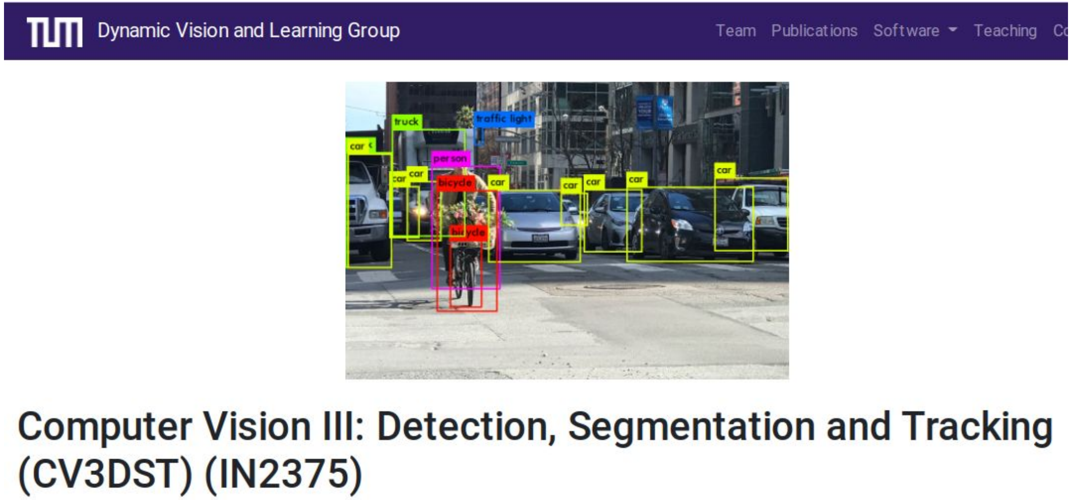
Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

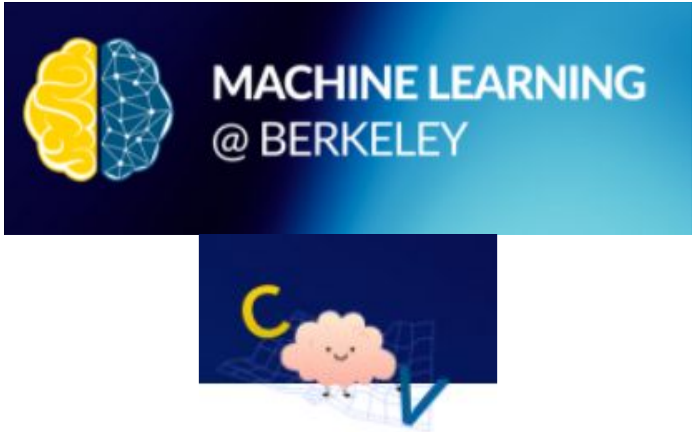
Semantic Segmentation

03.02.2023

PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

Contact: villar@ais.uni-bonn.de





CV3DST 2021 Lectures from TUM (Prof. Leal-Taixe)

https://dvl.in.tum.de/teaching/cv3dst-ws19/

Modern Computer Vision and Deep Learning @ Berkeley (Prof. Stuart Russell)

https://ml.berkeley.edu/decal/modern-cv

Motivation

Task Definition

Neural Network

'cat'

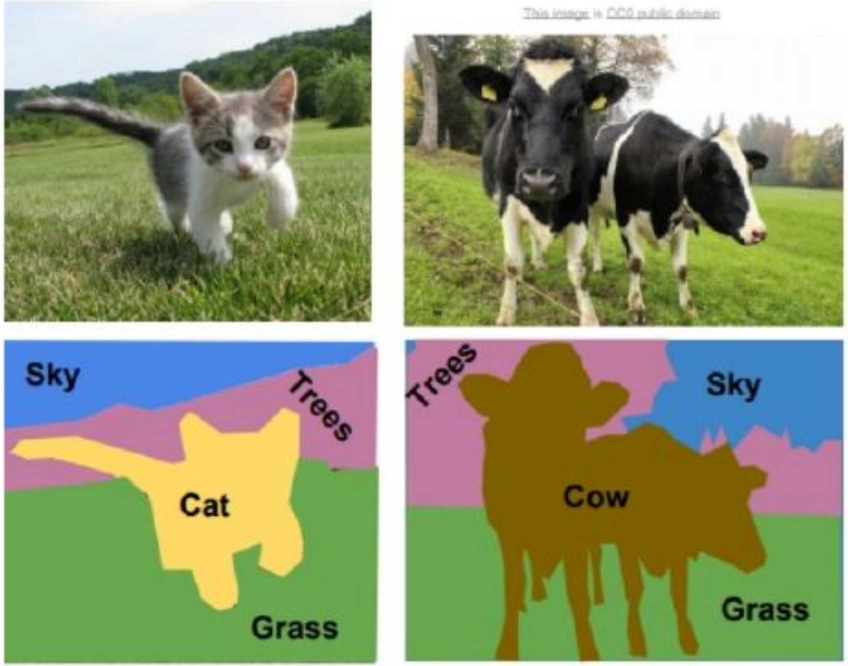
Classify the main object in the image

Neural Network

Cat, grass, sky, tree

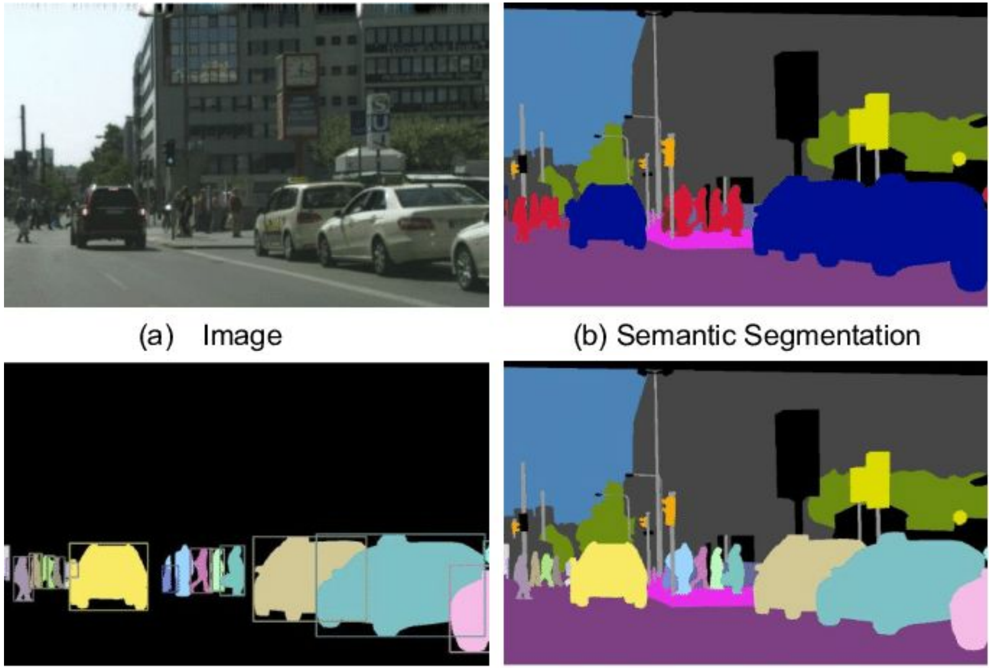
No objects, we classify each pixel

Semantic Segmentation



- Every pixel in the image is labelled with a category, e.g., sky, grass, person, …
- We do not differentiate between different instances of the same class (see cows in image 2)

Do not confuse…



(a) Image

(b) Semantic Segmentation

(c) Instance Segmentation

(d) Panoptic Segmentation

Label objects (car, person…) and stuff (sky, road, …), but no instance annotations

Segment different instances, but ignoring stuff (sky, road, grass …)

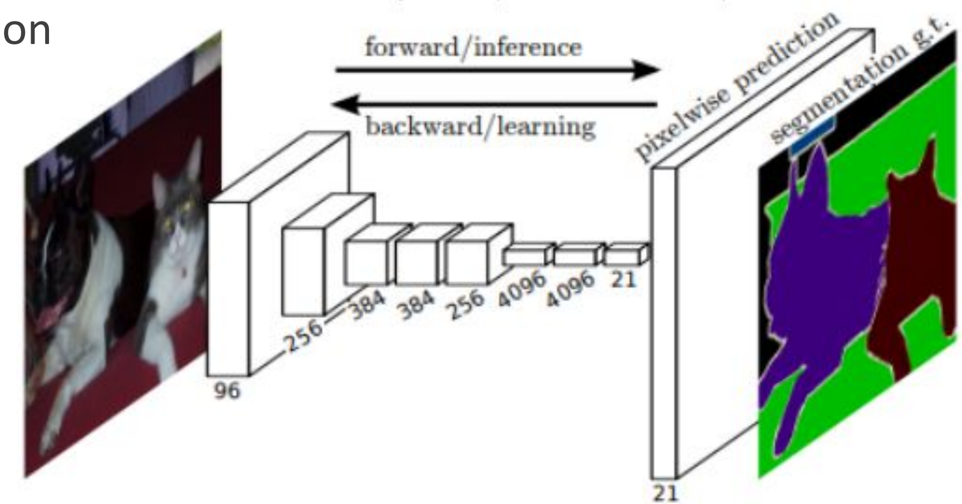
Combines semantic and instance segmentation

Fully-Convolutional Networks

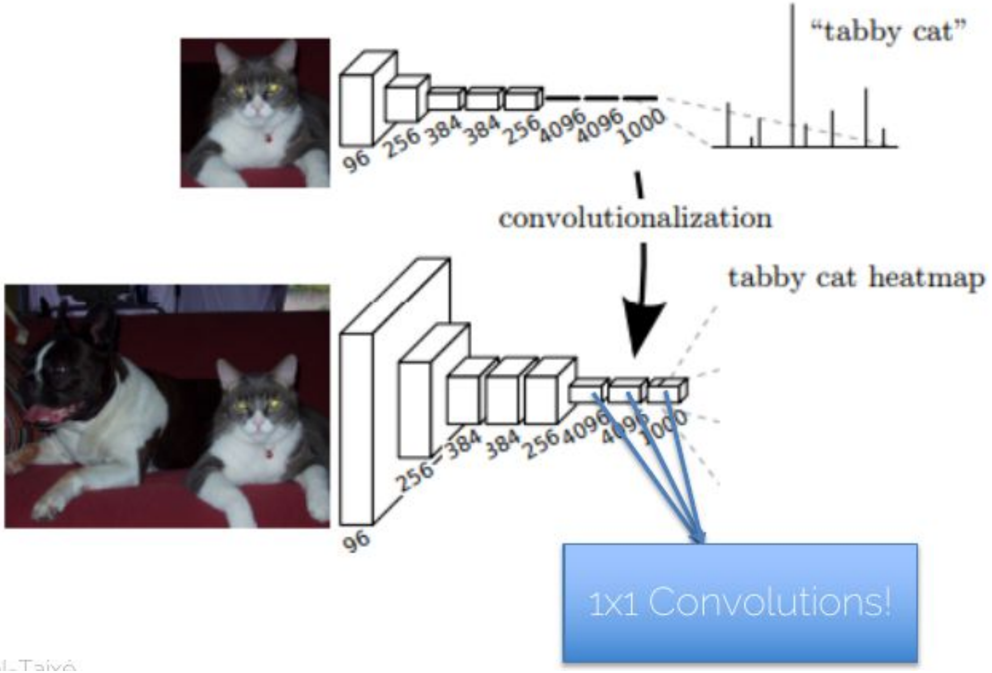
Fully Convolutional Networks (FCN)

- FCNs can handle any input/output size
- AlexNet moment of semantic segmentation
- Adapt a ConvNet from Classification:

1. Replace FC with convolutional layers
2. Convert to the original resolution in the last layer
3. Perform softmax-cross entropy between pixel-wise predictions and ground truth
4. Backprop and SGD



Convolutionalization

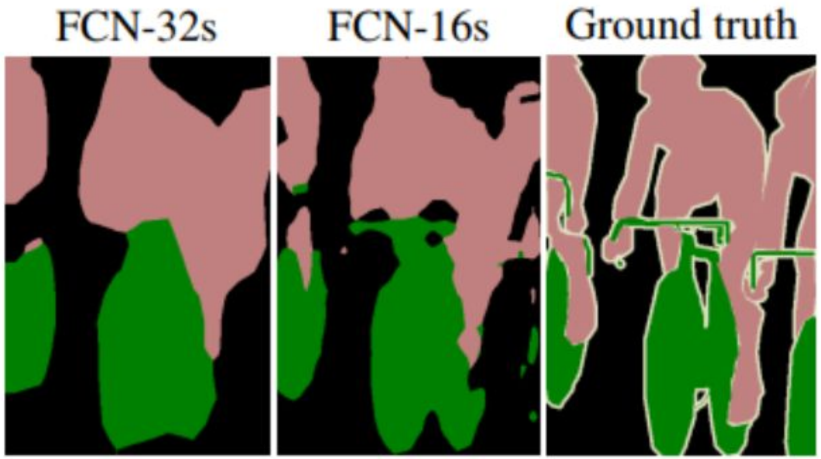


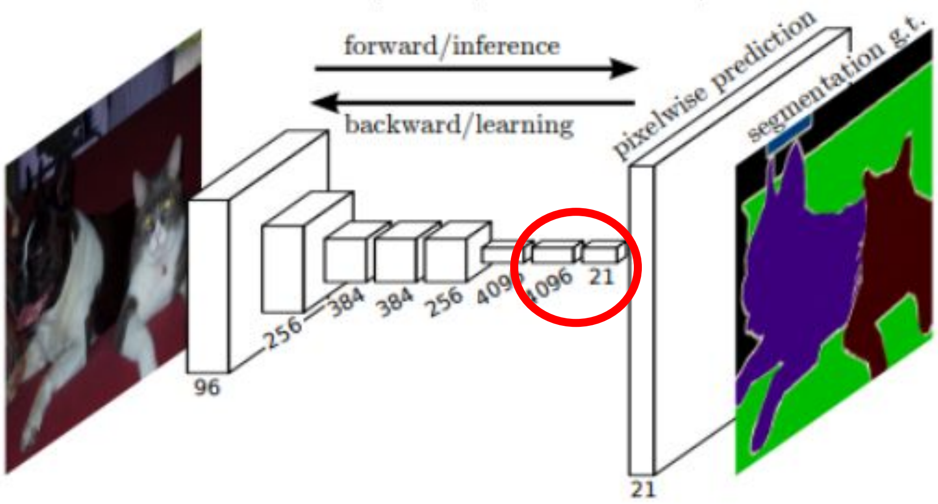
Q: What are 1x1 Convolutions for?

A: Change the number of channels

Problem with FCN

- Lost resolution via downsampling
- Cannot really be recovered by interpolating



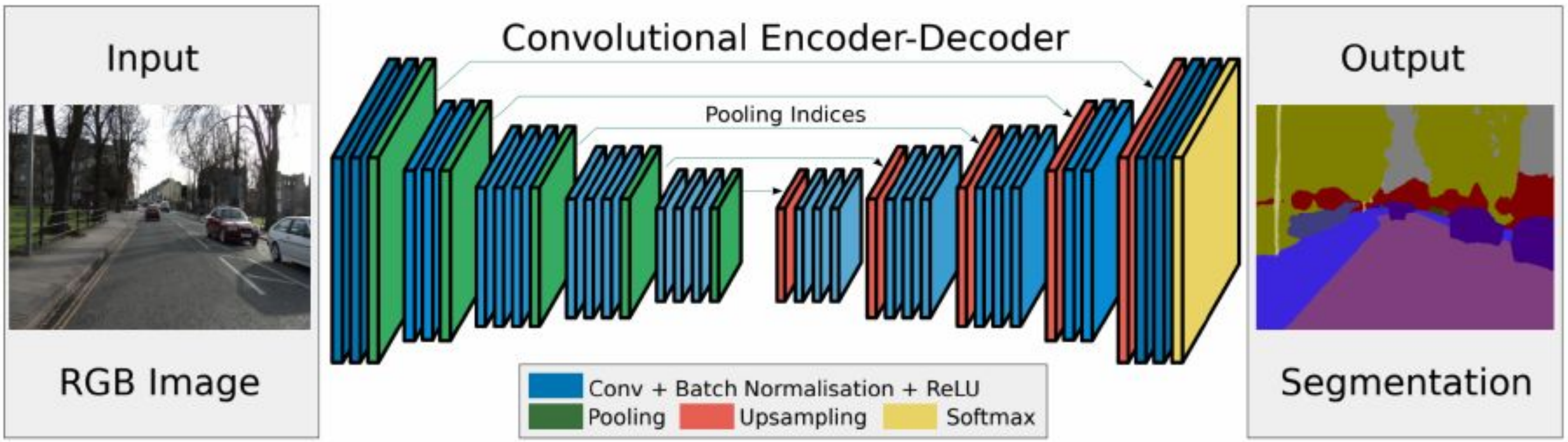


Autoencoder-Style Models

SegNet

- **Encoder:** convolutions + pooling
- **Decoder:** upsampling + convolutions

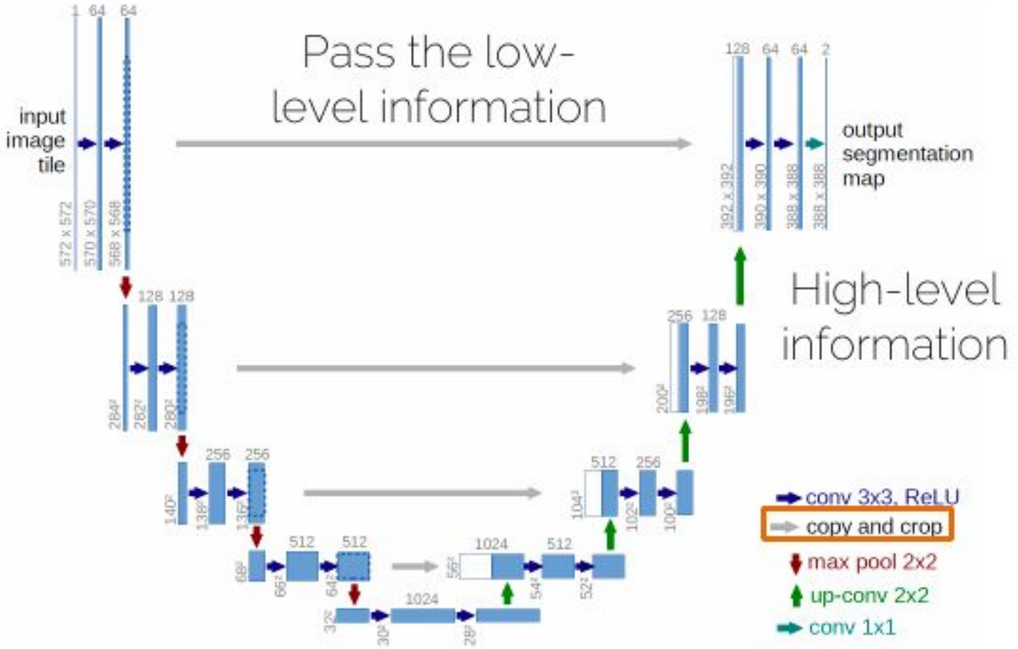
Rough upsampling +
refining the outcome



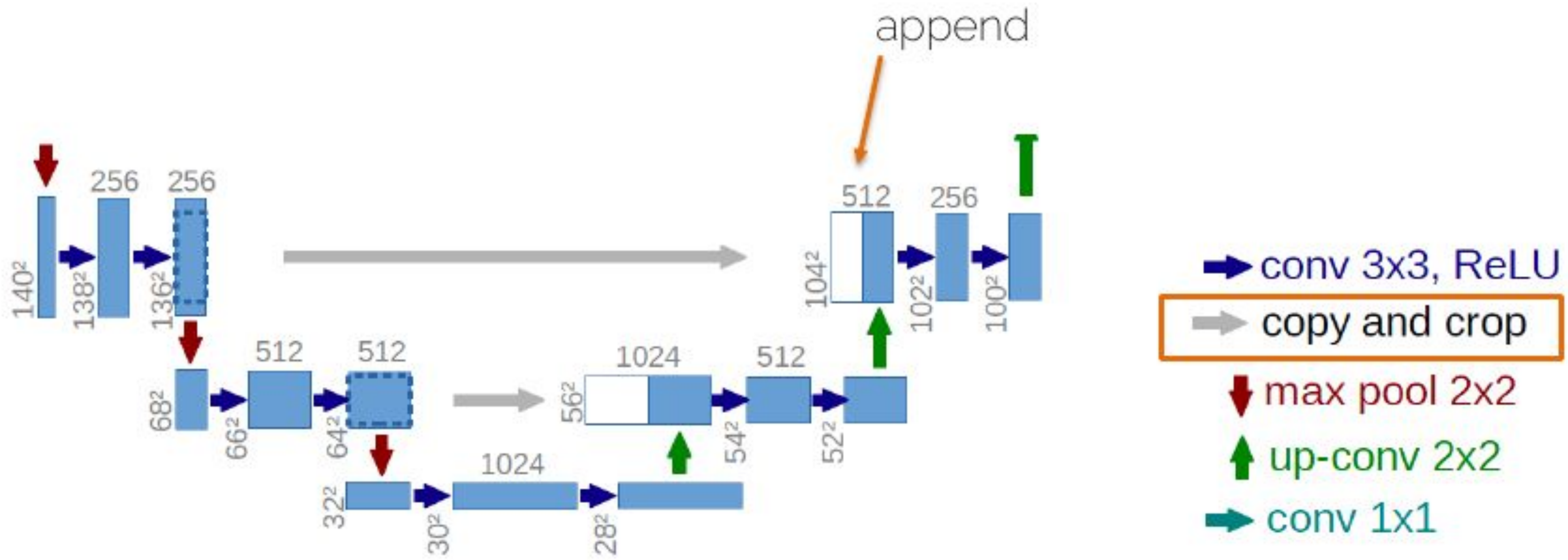
U-Net and Skip Connections

- Use information from multiple levels
  - Low-level information of high spatial resolution
  - High-level information (bottleneck) with low spatial resolution
  - Everything in between

Similar intuition as in ResNet!



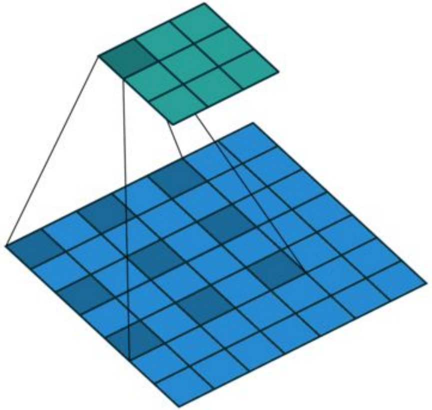
U-Net Zoomed In



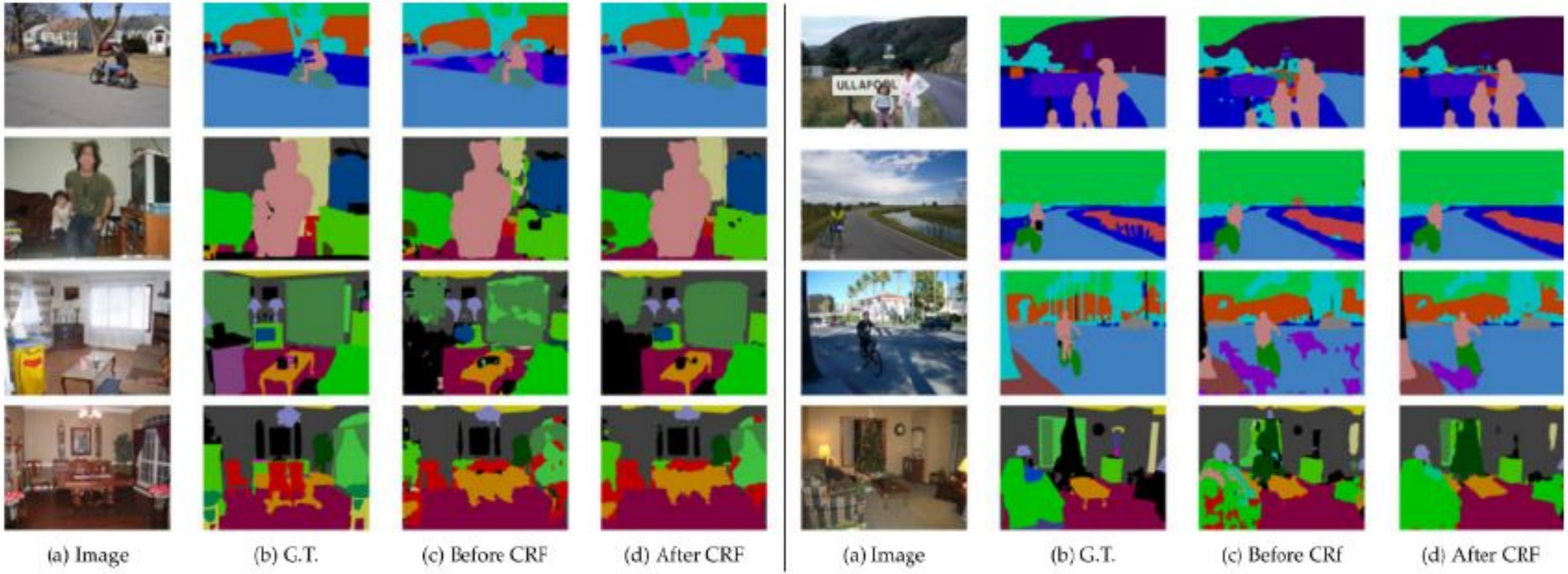
DeepLab

DeepLab Core Contributions

- Reduced feature resolution
  - ➢ Atrous (dilated) convolutions
- Objects of multiple scales
  - ➢ Pyramid Pooling
- Poor localization of edges
  - ➢ Refinement via Conditional Random Field (CRF)

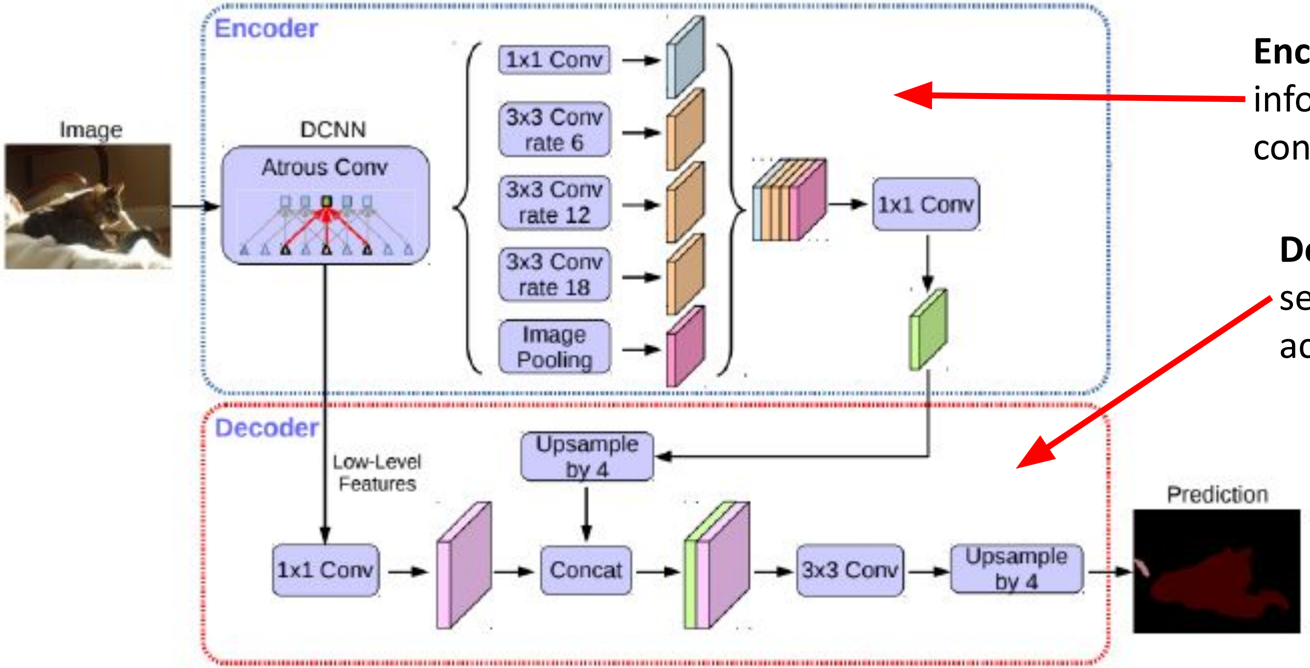


DeepLab



(a) Image
(b) G.T.
(c) Before CRF
(d) After CRF
(a) Image
(b) G.T.
(c) Before CRf
(d) After CRF

DeepLab v3+

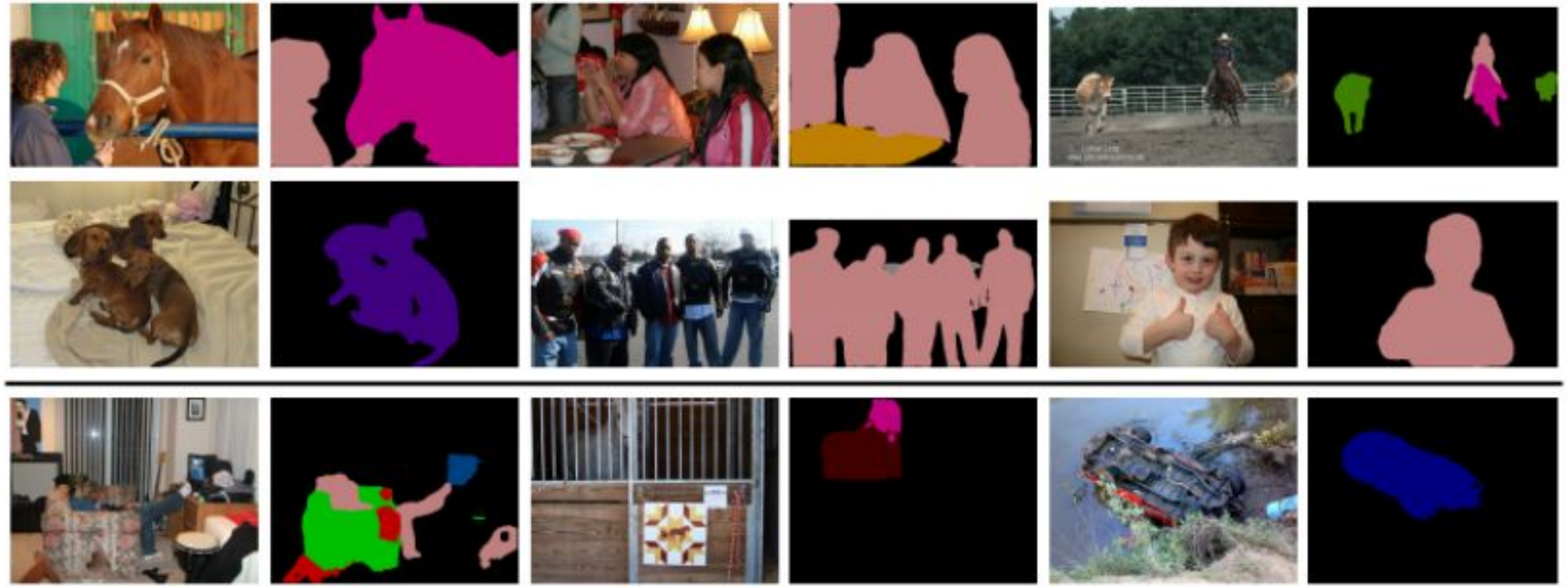


Encode multi-scale contextual information by applying atrous convolutions at multiple scales

Decode and **refine** segmentation results across boundaries

Extended use of depth-wise separable convolutions

DeepLab v3+

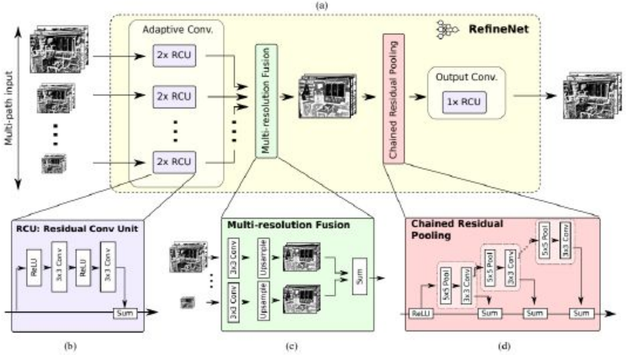


Still considered state of the art!

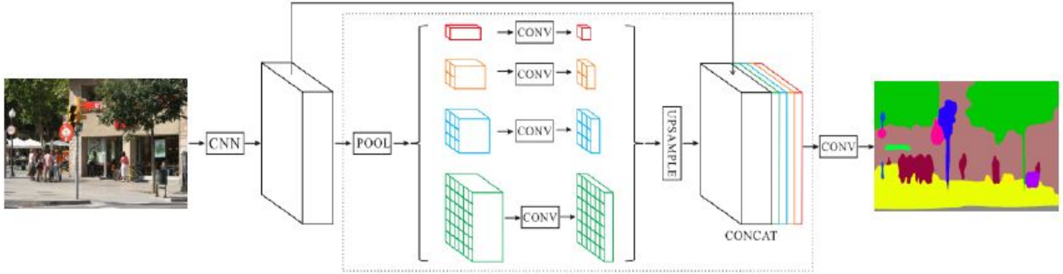
Many More Models

Many More Models

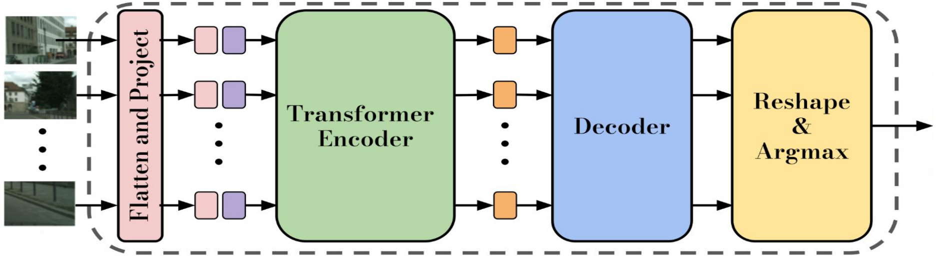
RefineNet



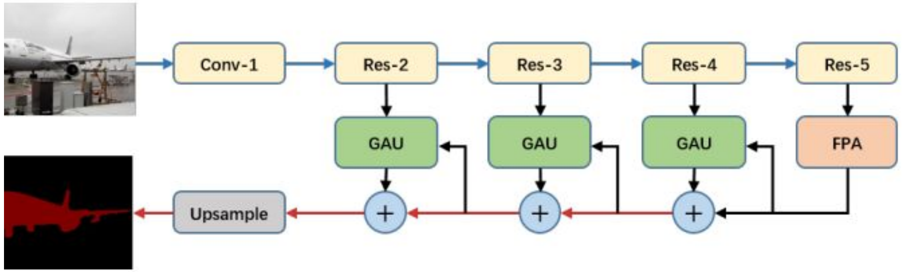
PSPNet



Segmenter



PAN-Net



Datasets & Evaluation

Datasets

Pascal VOC 2012: 9993 natural images divided into 20 classes.	Cityscapes: 25K urban-street images divided into 30 classes.	ADE20K: 25K (20 stands for 20K training) scene-parsing images divided into 150 classes.	Mapillary Vistas: 25K street level images, divided into 152 classes.

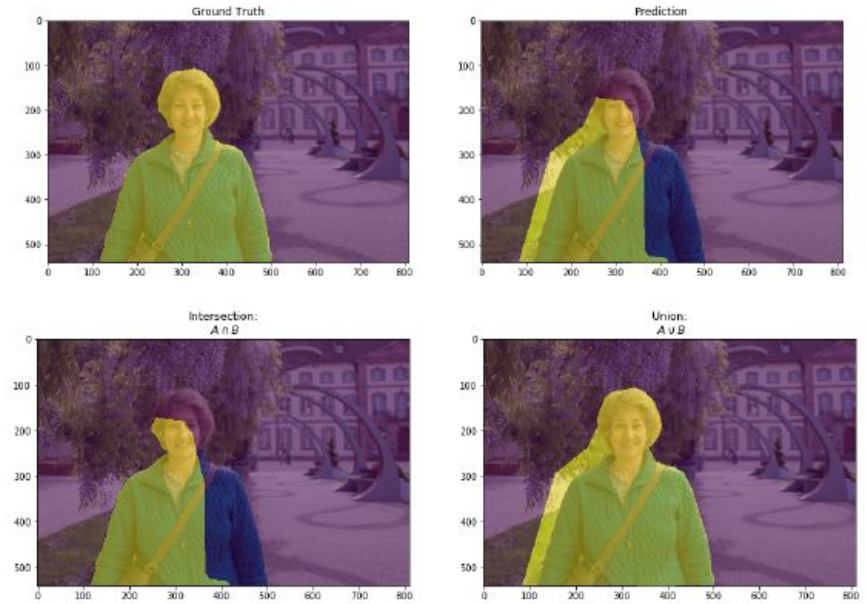
Models are often pre-trained in the large MS-COCO dataset, before finetuned to the specific dataset.

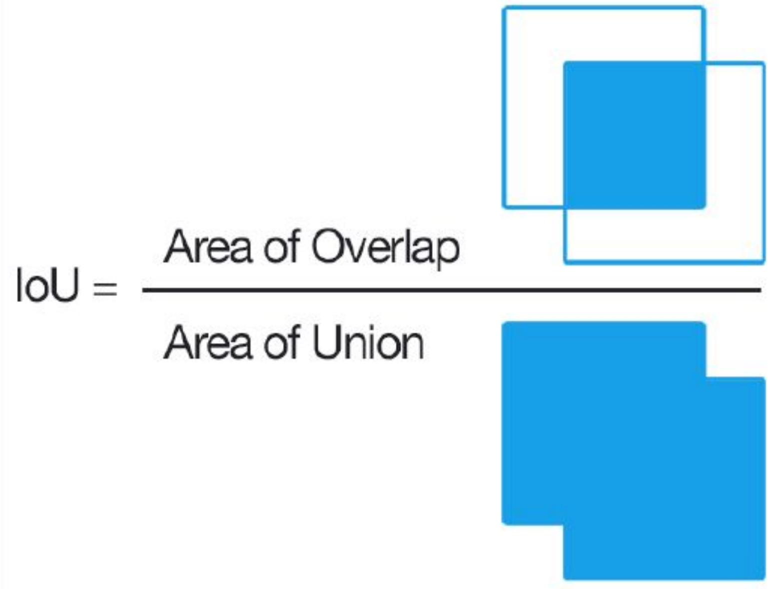
Segmentation Accuracy

- Percentage of correctly classified pixels
- Often given in a per-class basis, but you can compute it globally
- Can be misleading if some classes are underrepresented
  - Pedestrian vs road
  - Bird vs tree

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Intersection over Union





- **mIoU:** Compute IoU for each class, and average across classes

References

1. https://www.youtube.com/watch?v=XMSjOatyH0k&list=PLog3nOPCjKBkamdw8F6Hw_4YbRiDRb2rb&index=11
2. Long, Jonathan, et al. "Fully convolutional networks for semantic segmentation." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.
3. Badrinarayanan, Vijay, et al. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2017
4. Ronneberger, Olaf, et al. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention (MICCAI). 2015.
5. Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence (TPAMI) 2017
6. Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint, 2017.

References

7. Lin, Guosheng, et al. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." IEEE conference on computer vision and pattern recognition (CVPR). 2017.
8. Strudel, Robin, et al. "Segmenter: Transformer for semantic segmentation." IEEE/CVF International Conference on Computer Vision (ICCV). 2021.