Trinity College Dublin
The University of Dublin

V-SENSE

Semantic Segmentation

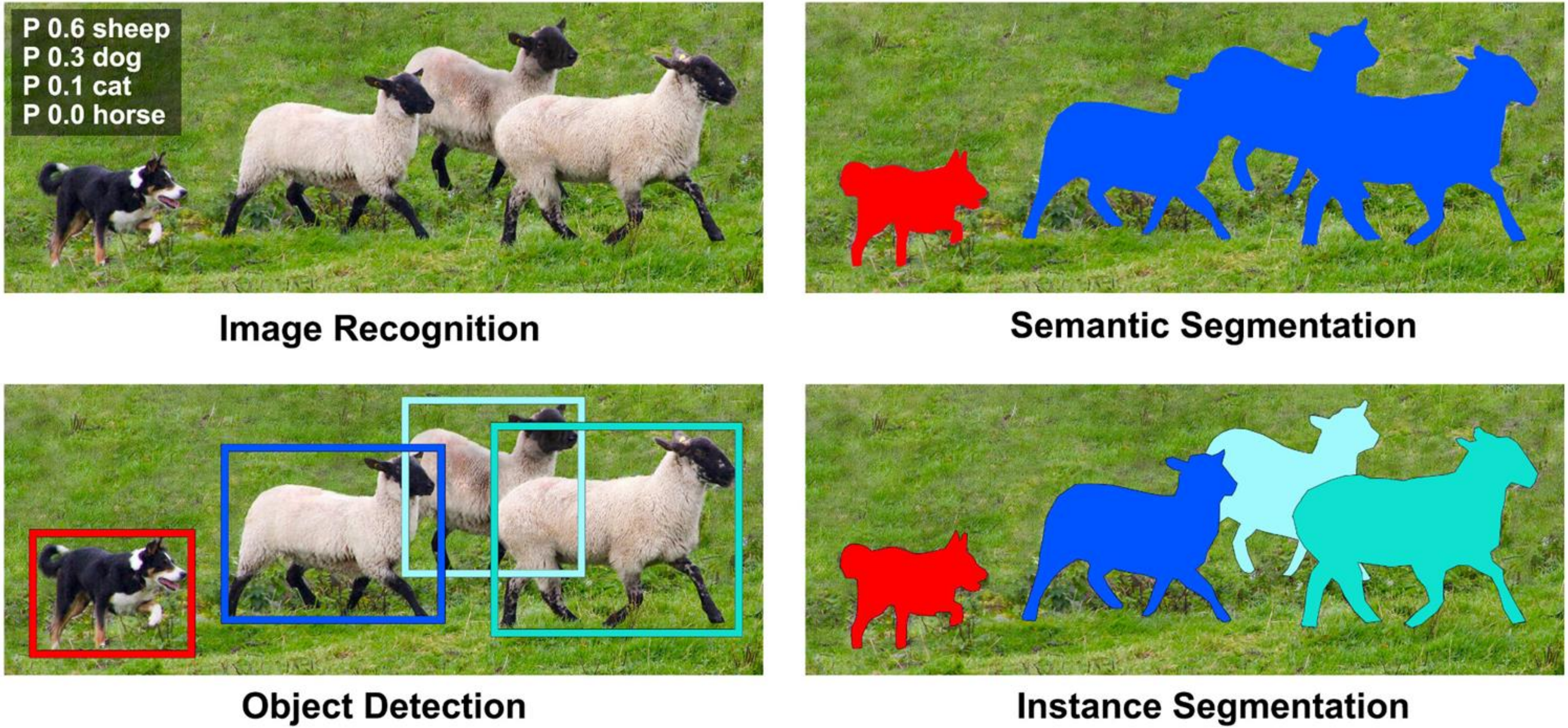
And other Image-to-Image translation tasks

Sebastian Lutz

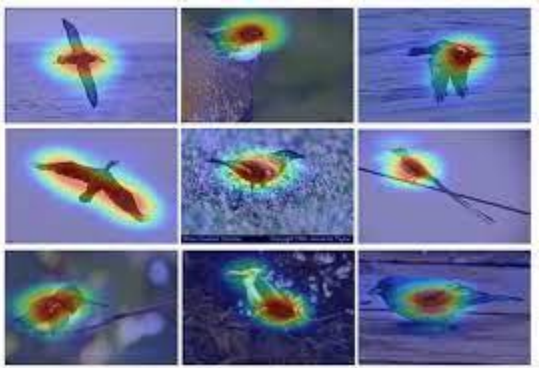
Content

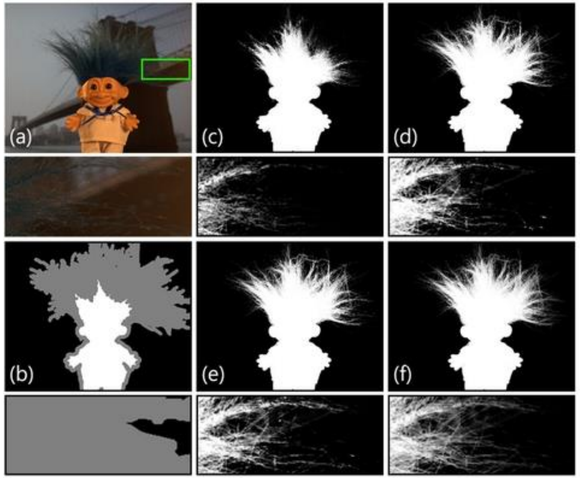
- **Defining semantic segmentation**
- **Seminal papers**
- **State of the art**
- **Datasets**
- **Training**
- **Examples**

Defining semantic segmentation



Similar tasks





Classical segmentation

- Early methods used hand-crafted features to locate object boundaries
- They modelled dependencies of neighbouring pixels
- Slow
- Problems with occlusion

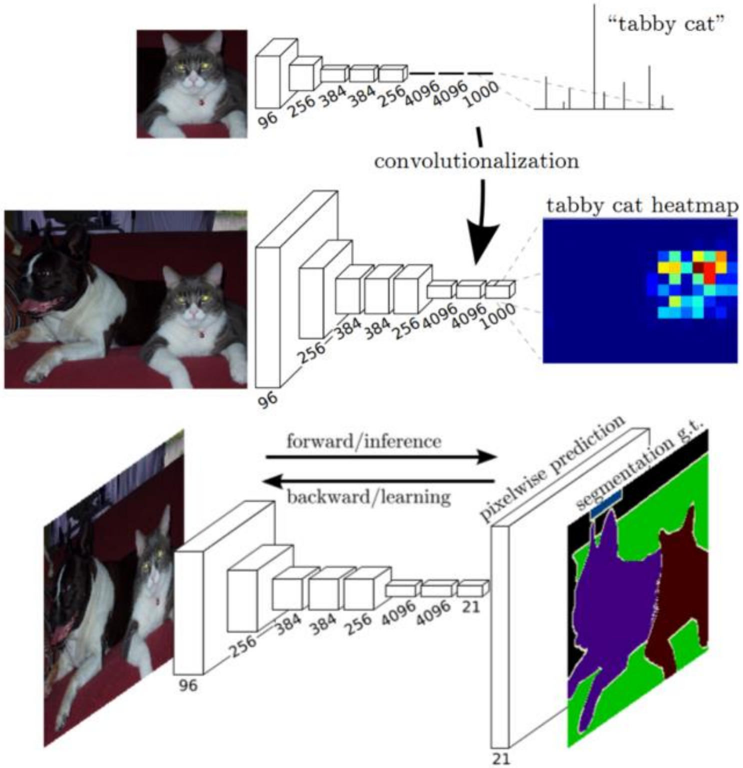
Early deep learning models

- Instead of hand-crafted features, features from classification networks were used
- Often, the fully-connected layers at the end were fine-tuned to the segmentation task
- These methods were constrained to patch-wise classification

Fully convolutional networks

- Transformation of fully-connected to convolutional layer
- Upsampling of smaller feature map through transposed convolution
- Prediction of any resolution
- Much faster than previous methods

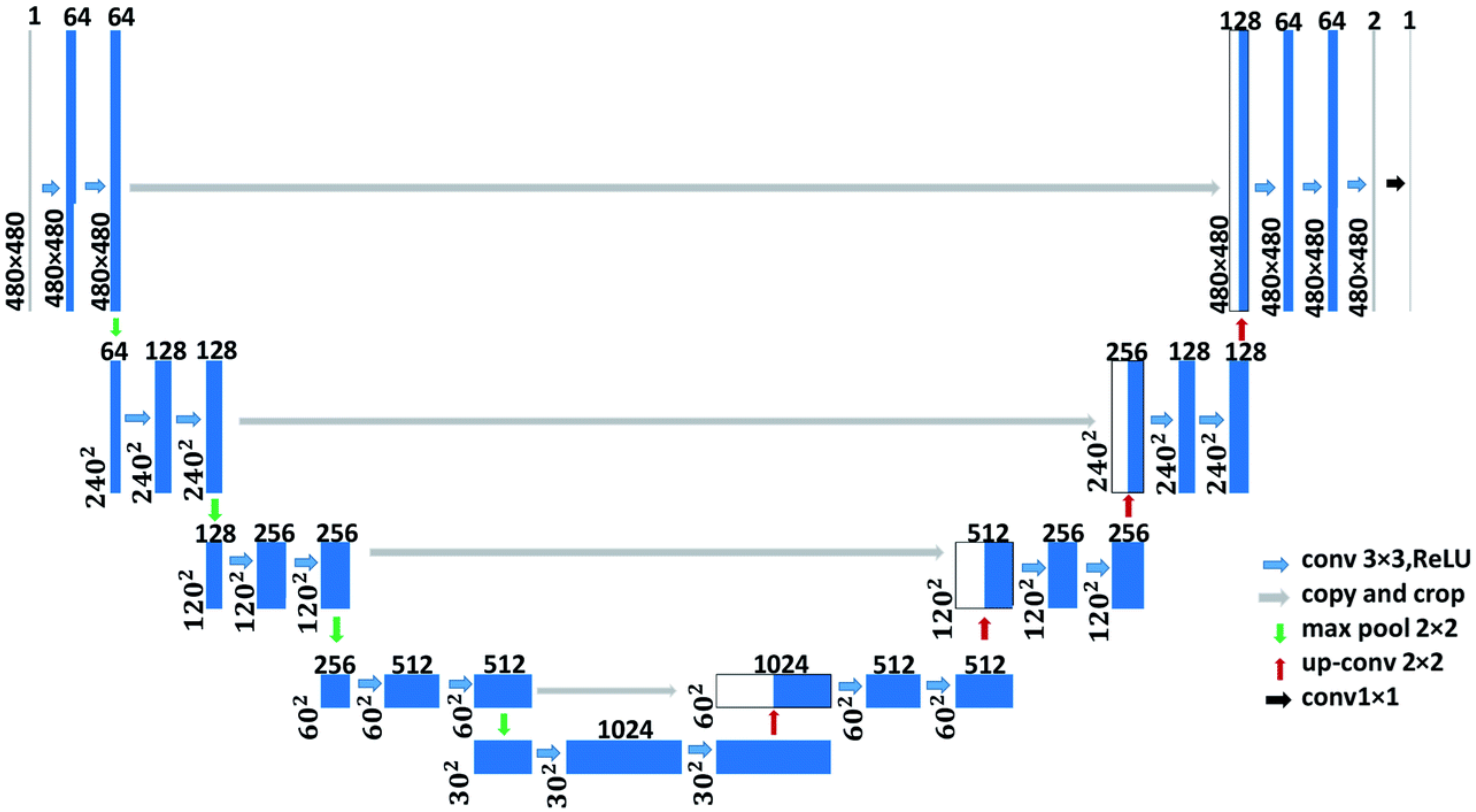
Fully convolutional networks



U-Net

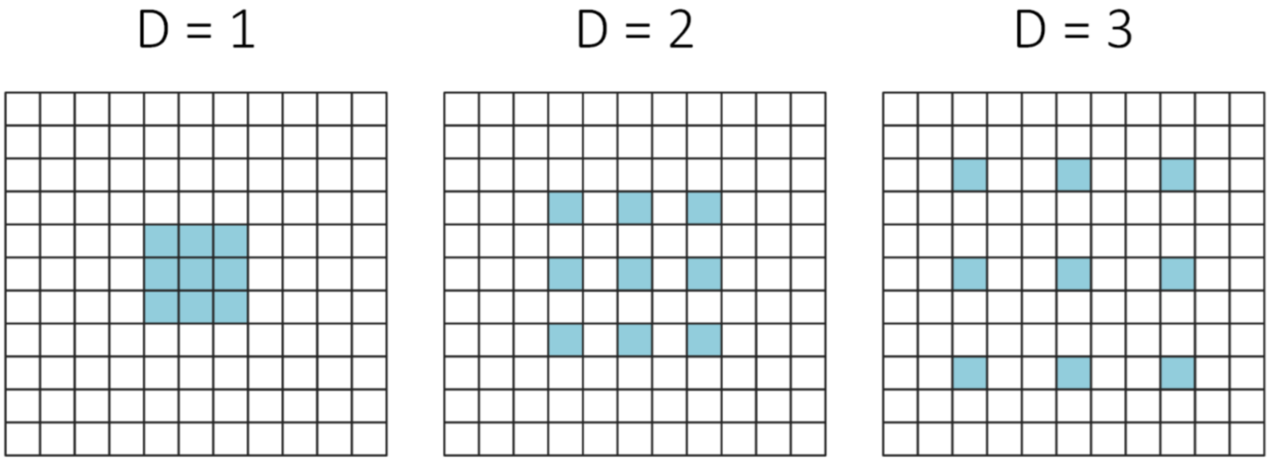
- Standard encoder in 5 blocks
- Symmetric decoder to gradually upscale feature maps
- Skip-connections to add local features from the encoder
    - Low-level features are helpful in recovering spatial information
    - Better object boundaries

U-Net



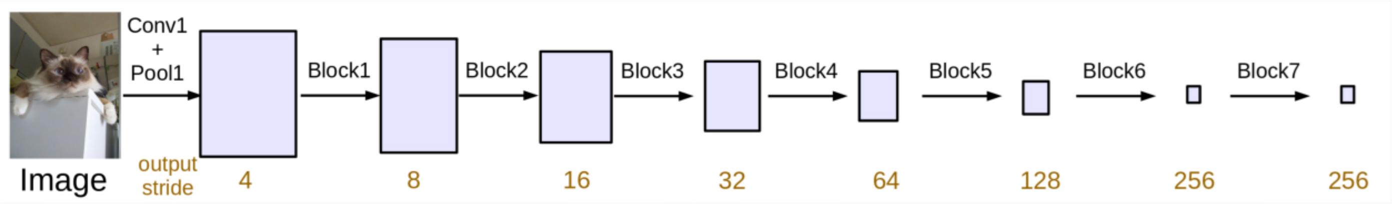
1
64
64
480×480
480×480
480×480
64
128
128
240^2
240^2
240^2
128
256
256
120^2
120^2
120^2
256
512
512
60^2
60^2
60^2
1024
30^2
30^2
30^2
1024
60^2
512
512
60^2
60^2
512
256
256
120^2
120^2
120^2
256
128
128
240^2
240^2
240^2
128
64
64
2
1
480×480
480×480
480×480
480×480
conv 3×3,ReLU
copy and crop
max pool 2×2
up-conv 2×2
conv1×1

Dilated convolutions

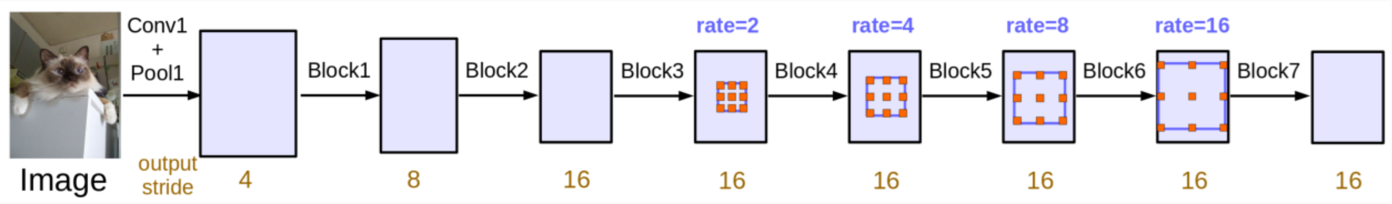


Same number of weights, larger receptive field
Also called “atrous convolutions”

Cascaded ResNet blocks



(a) Going deeper without atrous convolution.

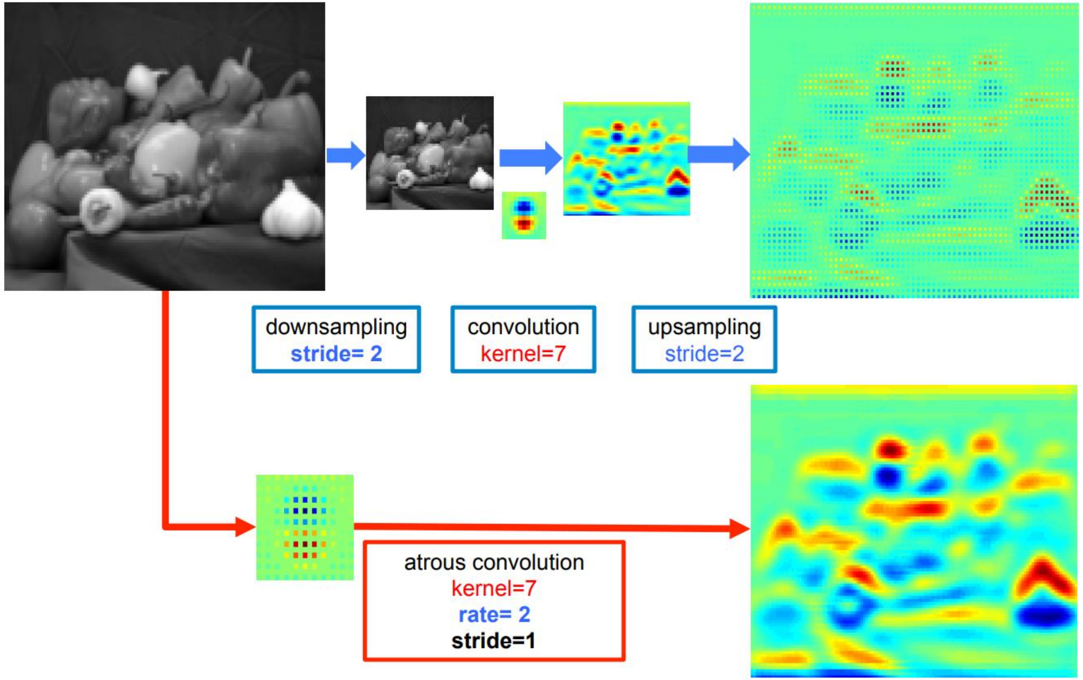


(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.
Figure 3. Cascaded modules without and with atrous convolution.

Keep the size of the receptive field without downscaling the feature maps → Less loss of spatial information

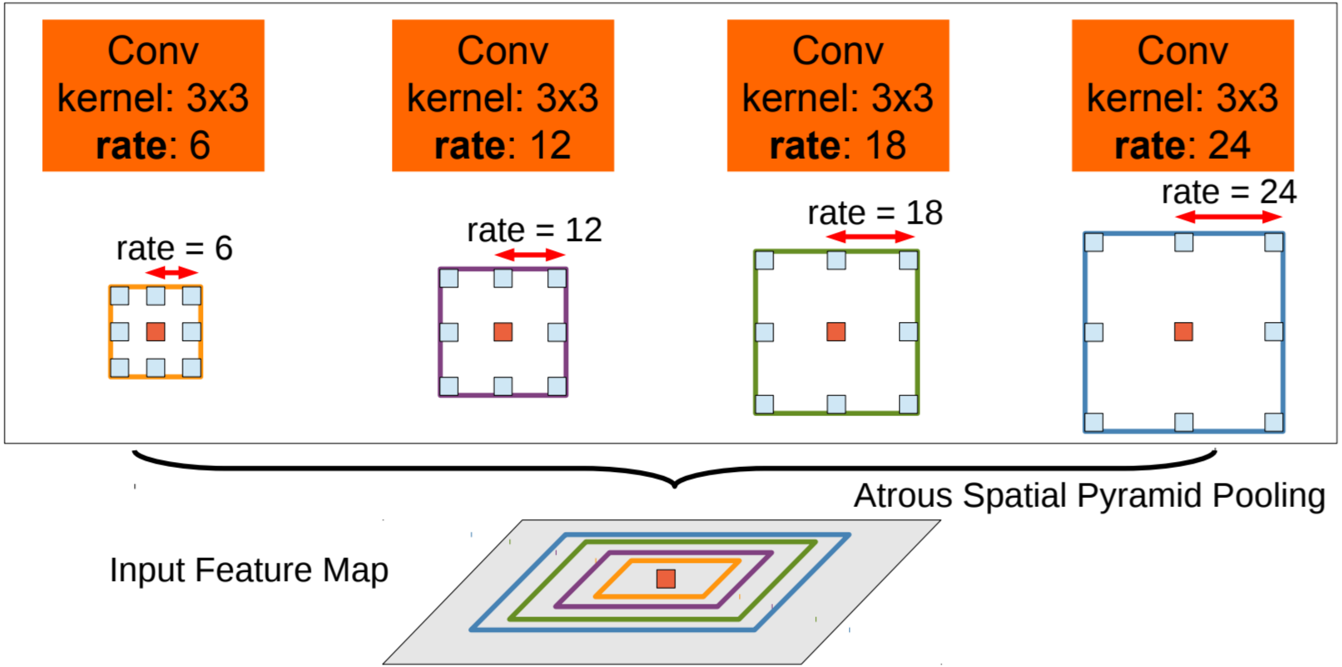
Dilated convolutions

- Recover local features using dilated convolutions



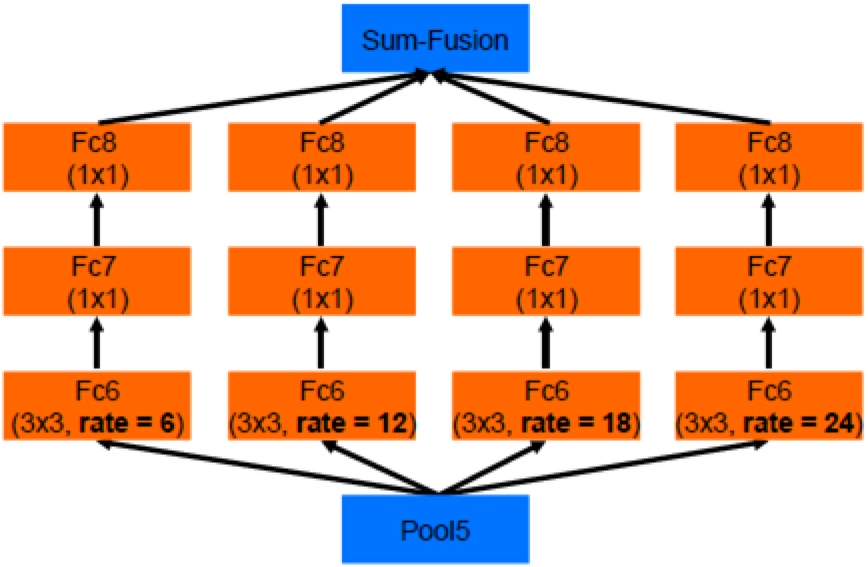
Atrous spatial pyramid pooling

- Exploiting multi-scale features through spatial pyramid pooling



Atrous spatial pyramid pooling

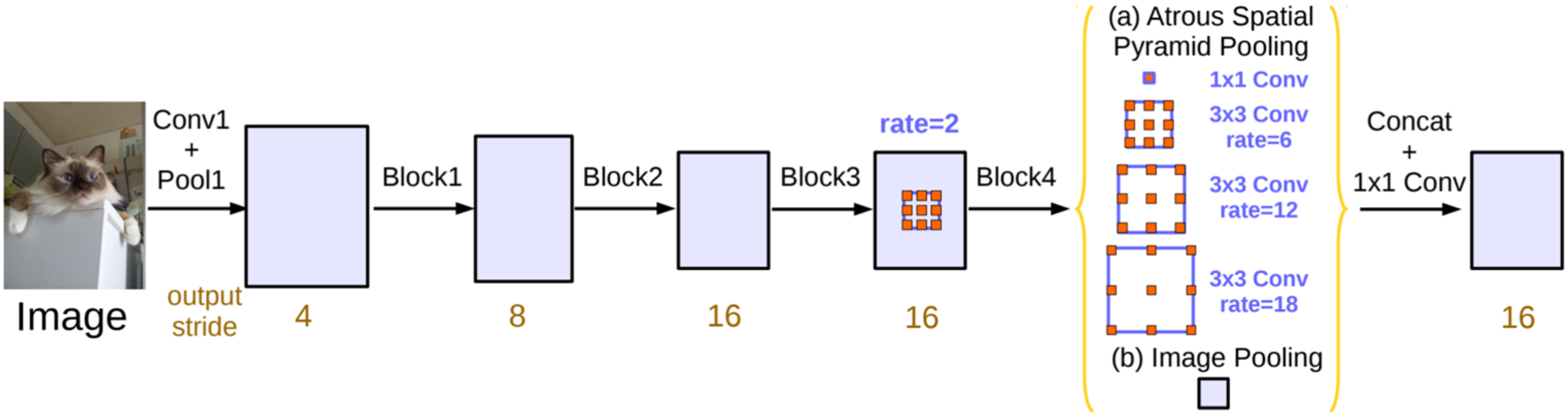
- Exploiting multi-scale features through spatial pyramid pooling



(b) DeepLab-ASPP

Deeplabv3

- Simple architecture with encoder, dilated convolutions and ASPP
- Another advantage of dilated convolutions: You can adjust the dilation rate during training and testing

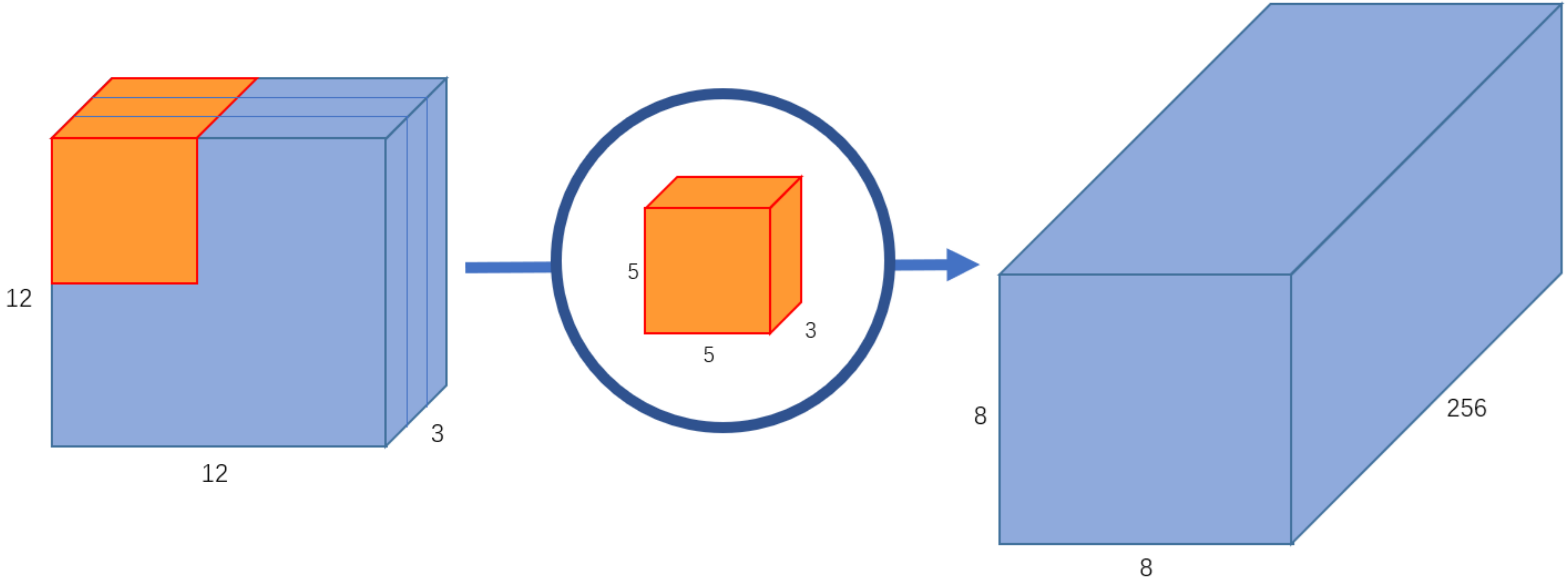


Depthwise separable convolutions

- Separate the depth and spatial dimensions of a convolution
- Instead apply a depthwise convolution first, followed by a pointwise convolution
- Less parameters → Less likely to overfit, saves GPU memory

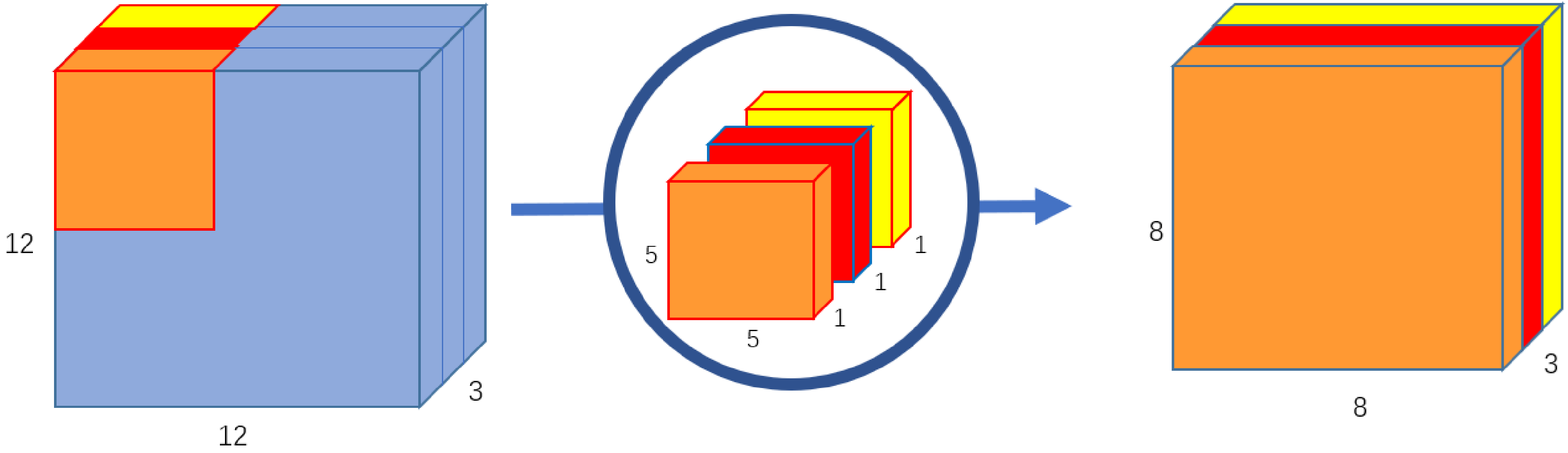
Depthwise separable convolutions

- Normal convolution (5 x 5 x 3 x 256 = 19200 parameters)



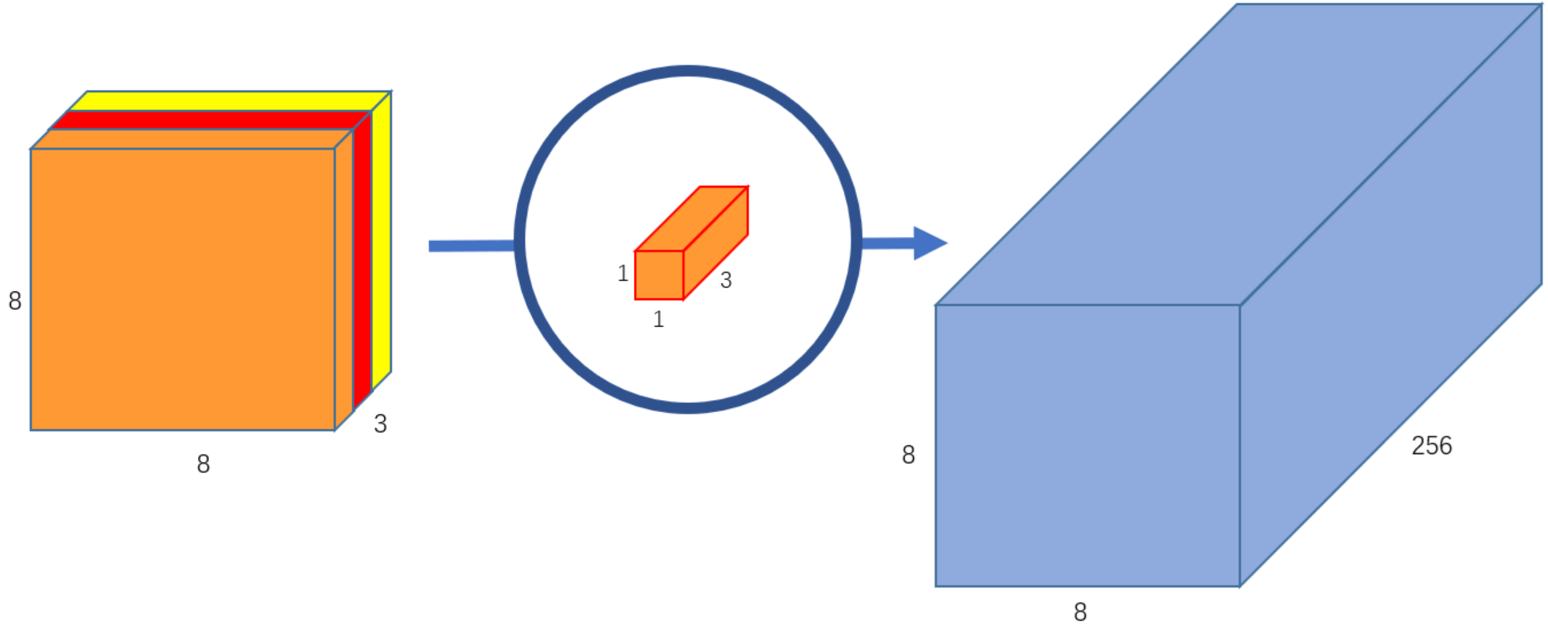
Depthwise separable convolutions

- Depthwise convolution (5 x 5 x 3 = 75 parameters)



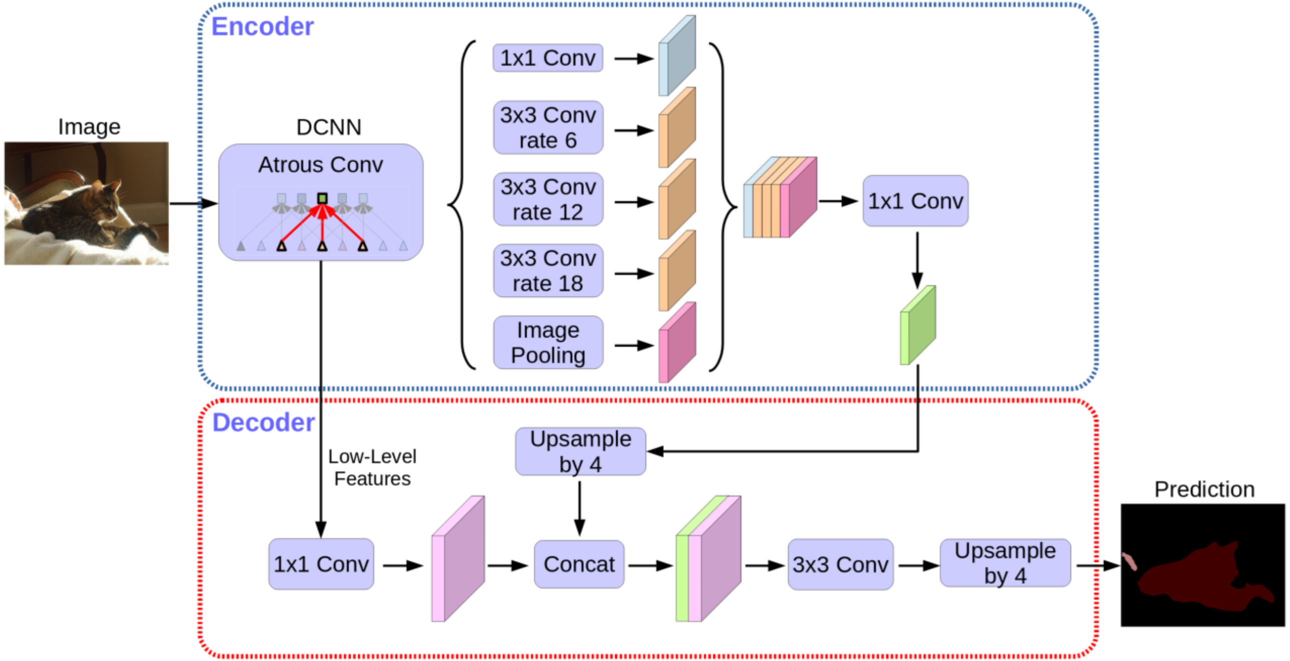
Depthwise separable convolutions

- Pointwise convolution (1 x 1 x 3 x 256 = 768 parameters)



Deeplabv3+

- Dilated convolutions, ASPP and decoder



Datasets

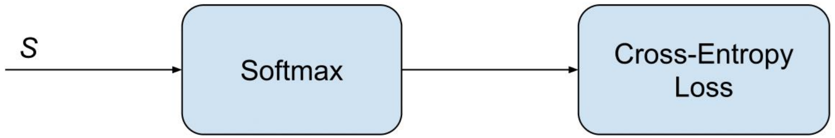
- **Pascal VOC**
  - 20 object categories
  - 9963 labelled images
- **MSCOCO**
  - 80 object categories
  - >200k labelled images
- **Cityscapes**
  - 30 object categories
  - 5000 finely labelled images
  - 20000 coarse labelled images

MSCOCO

Metrics

- **Pixel Accuracy**
  - Percentage of correctly classified pixels
  - Problem with class imbalance
- **Intersection over Union**
  - Overlapping area between prediction and ground truth divided by union of prediction and ground truth (per class)
- **Dice coefficient (F1-Score)**
  - Twice the area of overlap divided by the total number of pixels

Losses



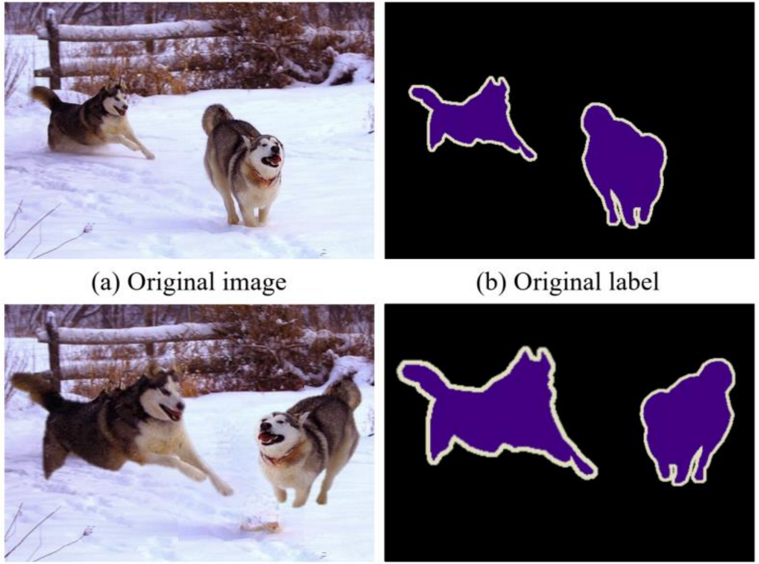
$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CE = -\sum_i^C t_i log(f(s)_i)$$

- **Categorical Cross-entropy**
  - Default classification loss
- **Focal loss**
  - Weighs the contribution of each sample to the loss
  - Already well classified samples have less contribution
  - Solution to class imbalances in the dataset
  - Binary loss, applied to each class
  - If $\boldsymbol{\gamma}$ = 0, equivalent to BCE

$$FL = -\sum_{i=1}^{C=2} (1 - s_i)^{\gamma} t_i log(s_i)$$

Data augmentation

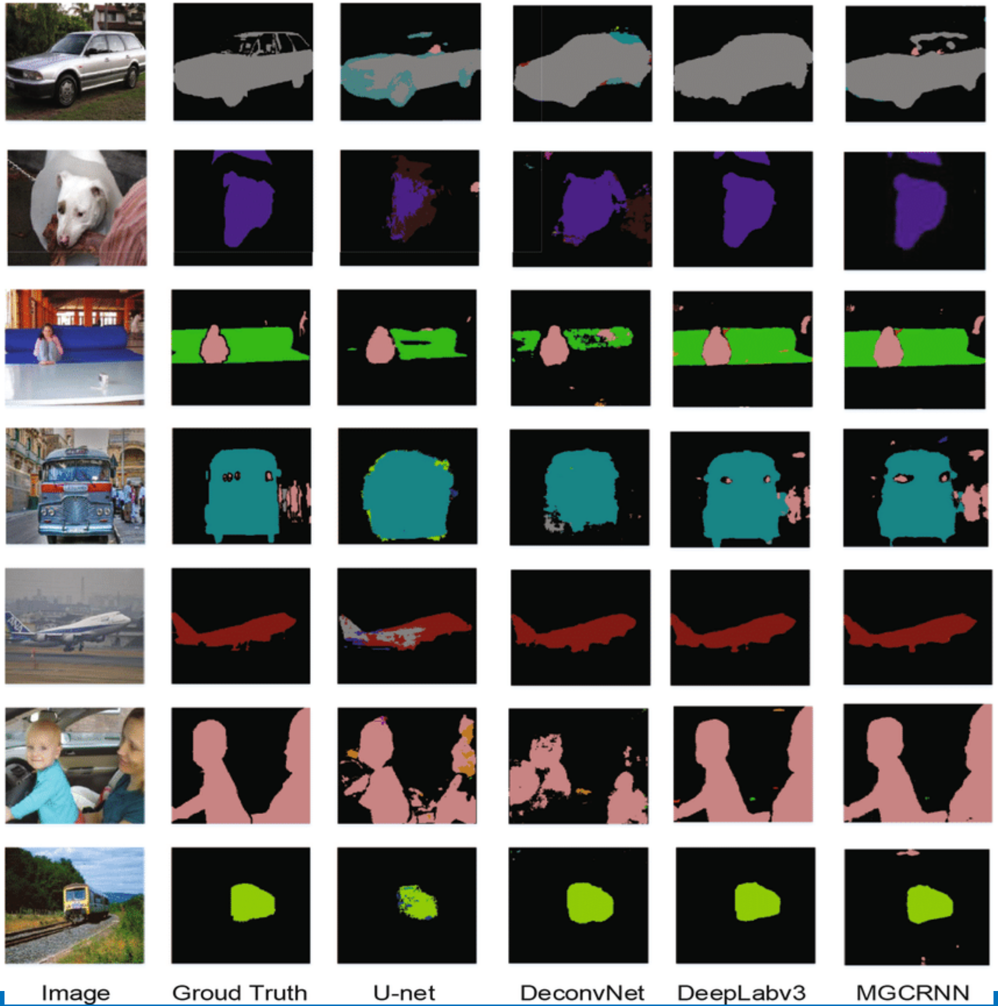
- Traditional data augmentation
    - Randomly crop, rotate, shift colors, etc.
- Specific data augmentation
    - Manipulate objects in the sample
- Test-time data augmentation
    - Test the same image multiple times
    - Each with slightly different rotation
    - Average results



(a) Original image (b) Original label

(c) Augmented image (d) Augmented label

Examples



Trinity College Dublin
The University of Dublin

V-SENSE

Many Thanks!