DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Convolutional Neural Networks (CNN)
Part 3

Søren Olsen

Popular CNN-modules

Image classification, Object detection and Segmentation

- Loss functions
- AlexNet
- VGG11-16-19
- ResNet
- Feture Pyramid Networks (FPN)
- R-CNN-family, Unet

- Course Evaluation

Loss functions

- The loss function determines the performance!

- For regression use the sum of squared errors. This is a maximum likelihood estimate if the error is normally distributed.

- For classification use cross-entropy error and one hot encoding if multiple classes. Often, a soft-max activation function is applied to make the output integrate to 1.0.

- When bounding boxes and/or masks are to be estimated simultaneously, the loss function will be a composition.

Loss function for binary classification

- Assume the output $f(x,w)$ of a net is a real number in [0,1] and that the true value (the ground truth) y is binary, i.e. in {0,1}. Also assume that the probability follows a Bernulli distribution:

$$p(y|x,w) = f(x,w)^{y}\,[1\ - f(x,w)]^{1-y}$$

- Then the negative logarithm of a batch (y_n) of input values (x_n) leads to the **cross-entropy** measure:

$$-\sum^{N}[y_n \ln(f) + (1 - y_n)\ln(1\ - f)]$$

where f is s shorthand for $f(x,w)$.

from C. Igel

Soft-max

- To ensure that the net outputs values $f(x,w)$ in [0,1], a soft-max activation function usually is applied, i.e.:

$$f(x,w) \rightarrow \frac{e^{f(x,w)}}{\sum_{i=1}^{K} e^{f(x,w)}}$$

where K is the number of classes. Thus, after normalization $\sum_{i=1}^{K} f(x,w) = 1$.

from C. Igel

Multi class classification

- For K classes, One-hot encoding is used, i.e. each sample is coded with a single integer in [1:K]. Thus, y_i is a *K*-dimensional vector with exactly one 1, and zeros for the rest.

- Assuming that soft-max normalization has been done we have:

$$p(y|x,w) = \prod_{k=1}^{K} [f(x,w)]^{y_k}$$

from C. Igel

and further

- Taking the negative logarithm gives the cross-entropy for multiple classes:

$$-\ln p = -\sum_{n=1}^{N}\sum_{k=1}^{K}[y_n]_k\, ln[f[x_n, w)]_k$$

- Cross-entropy is the most used minimization measure used. However, it is often modified in several ways.

- from C. Igel

Class imbalance

If the number of observations N_i in different classes deviate a lot, the estimation will be biased and potentially wrong.

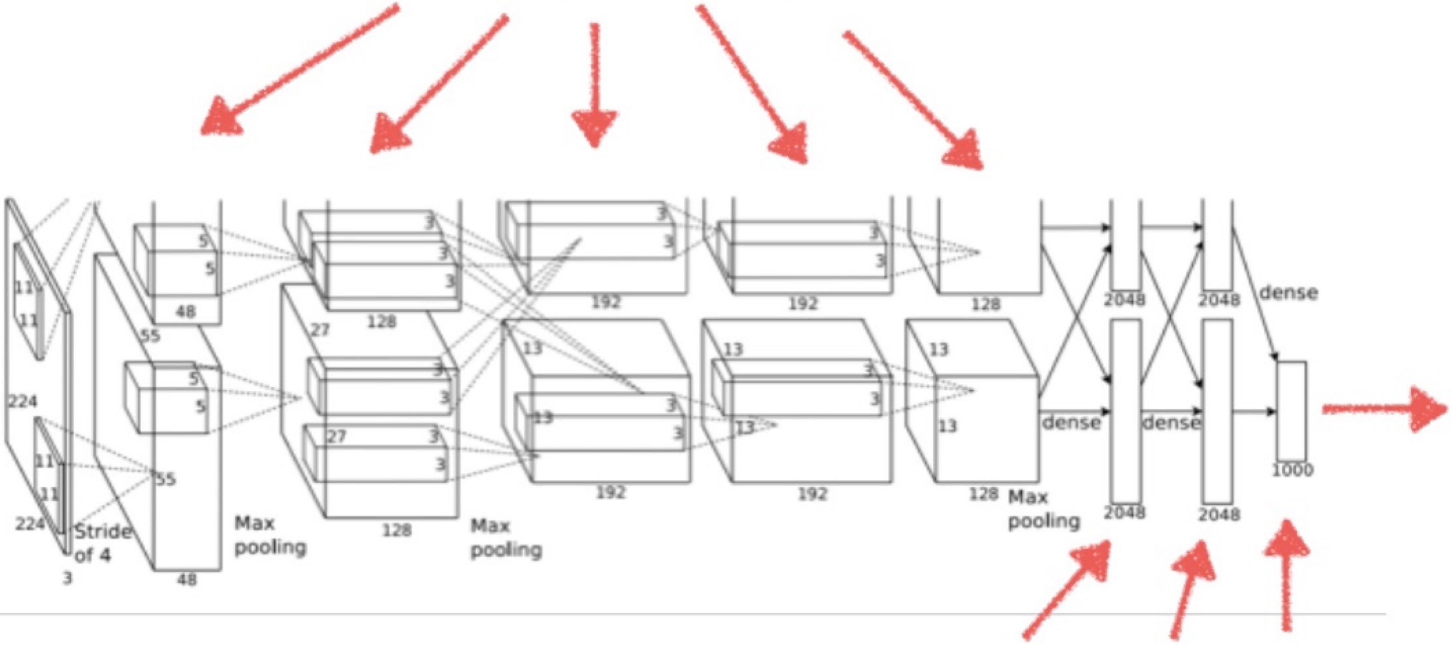
To remedy class imbalance one approach is to scale the error contribution in the loss function with the inverse number of class representatives.

$$\alpha_i = \frac{1}{N_i}$$

Modifying the loss to prevent class imbalance usually is mandatory and improves performance significantly.

AlexNet

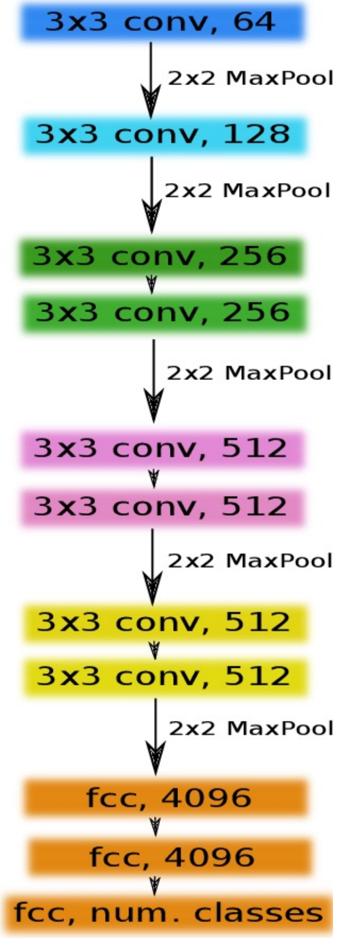
5 Convolutional Layers



1000-way softmax

3 Fully Connected Layers

VGG family – VGG11



3x3 conv, 64

2x2 MaxPool

3x3 conv, 128

2x2 MaxPool

3x3 conv, 256

3x3 conv, 256

2x2 MaxPool

3x3 conv, 512

3x3 conv, 512

2x2 MaxPool

3x3 conv, 512

3x3 conv, 512

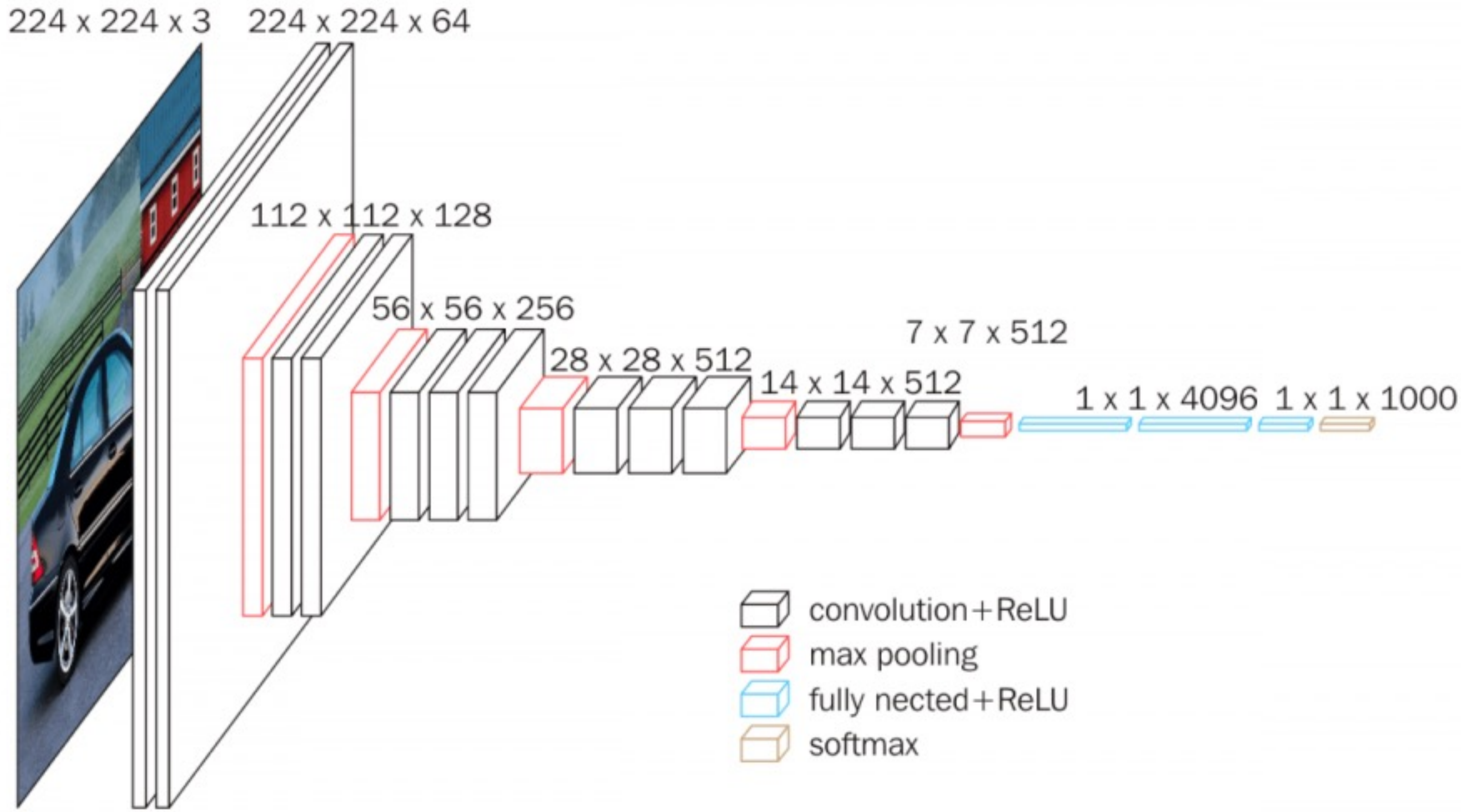
2x2 MaxPool

fcc, 4096

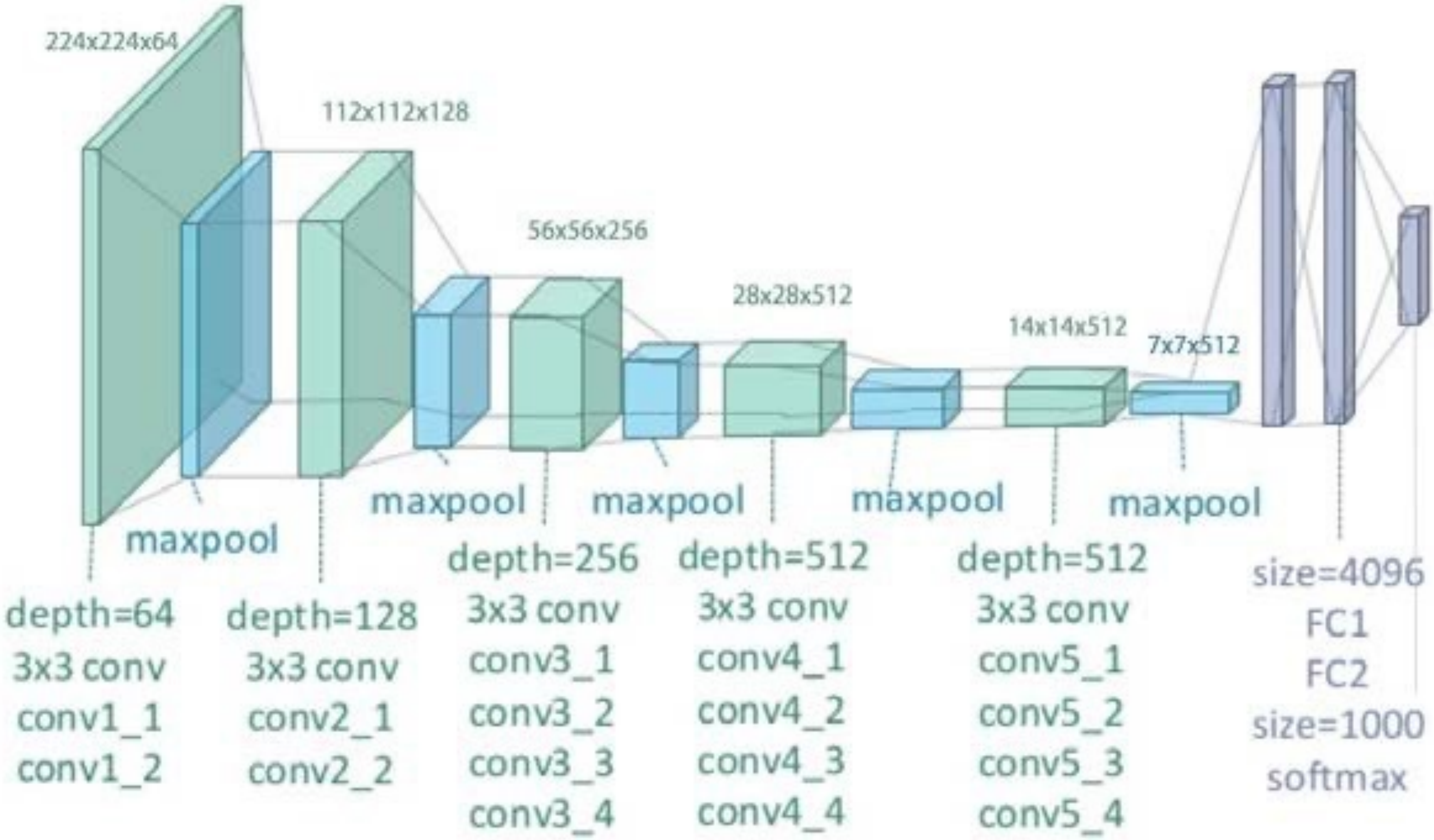
fcc, 4096

fcc, num. classes

VGG-family VGG-16



VGG-family VGG19



Vanishing gradient problem

- Researchers found that for shallow nets, increasing the number of layers improved performance. For deeper nets, the performance decreased. The feature abstraction change will be slower.

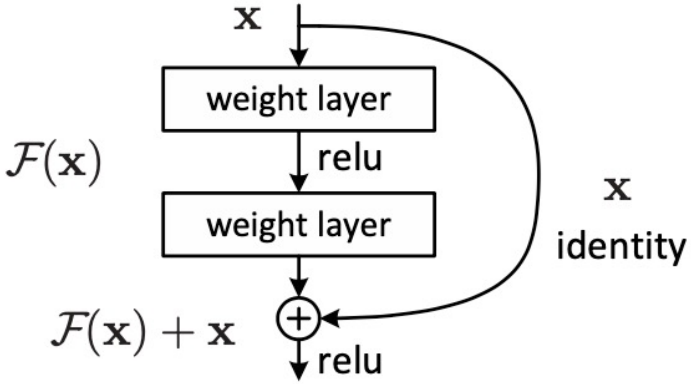
- The immediate consequence is that the training gradient becomes smaller between neighboring layers and tend to vanish.

- In He, Zhang, Ren, Sun: Deep Residual Learning for Image Recognition a new Residual learning building block is proposed and shown to solve the vanishing gradient problem.

Residual building bloak

- Imagine a CNN performing a sequence of mappings F_i between different representations $x_i$of the image
- $(RGB) \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_i \longrightarrow x_{i+1} \longrightarrow \cdots \longrightarrow x_n$
- The mapping made by a unit F_i must then correspond to the residual/difference $x_{i+1} - x_i$.
- Conversely, $x_{i+1} = x_i + F_i$

Residual building block



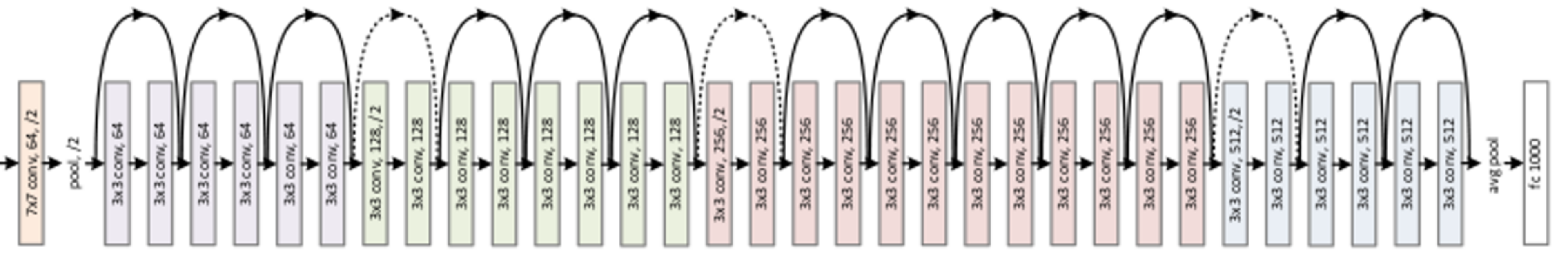
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning: a building block.

The residual building block is used in ResNet50, ResNet101 and ResNet152 (reference networks) and is shown to work for a net with 1202 layers.

For very deep structures also, the Bottleneck building block is popular.

ResNet: ResNet34 (skipping, 18, 50, 101, 152, 1202)



Today, ResNet50, ResNet101 and Resnet152 serve as common backbone-net in many applications.

The Bottleneck building block

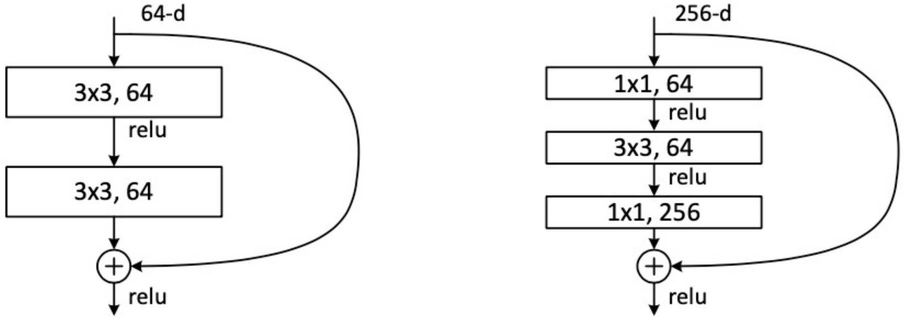


Figure 5. A deeper residual function $\mathcal{F}$ for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a "bottleneck" building block for ResNet-50/101/152.

The bottleneck idea is to downscale the channel number before convolution and upscale again after convolution. Here combined in a residual context.

Questions ?

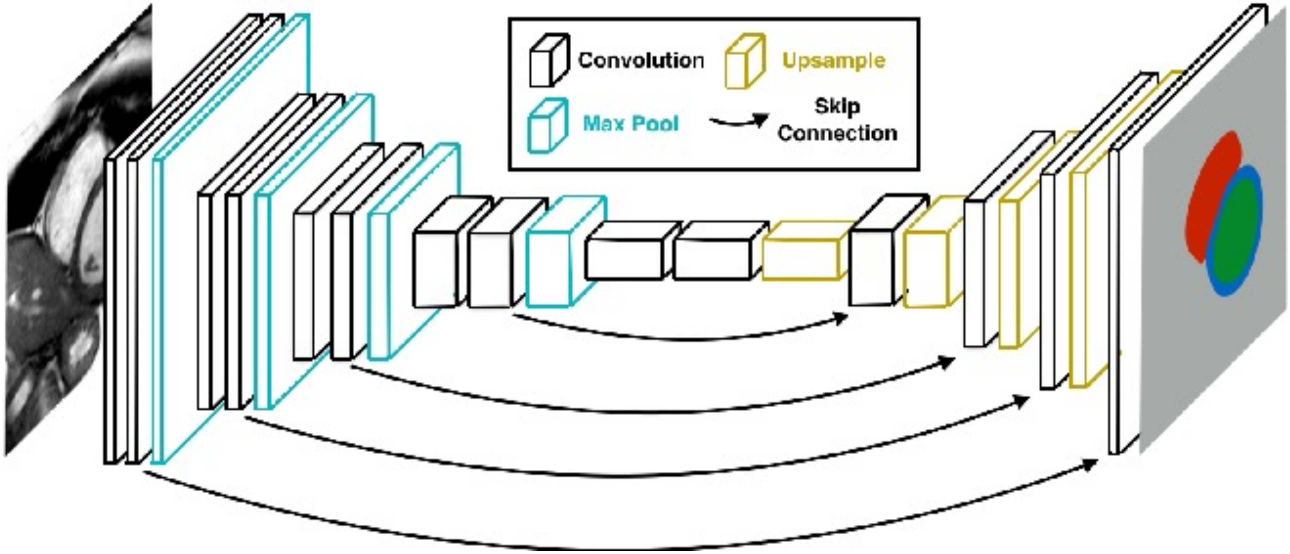
Many other CNN constructions

- We will skip a lot of contributions, including
  - Inception Net family
  - Squeeze and Exitation Net
  - DenseNet
  - R-FCN
  - DeepNet
  - Efficent Net
  - Yolo, SSD
  - R-CNN family
  - etc.

- Instead, I will sketch a single semantic segmentation method.

Skip connections

- Skip connections may be done as in residual building blocks but may also span a larger range. In e.g. UNet information from the first levels is carried through



The combination of skip connections and upsampling (e.g. for use in semantic segmentation) may be seen as an alternative to use of a feature pyramid.

UNet

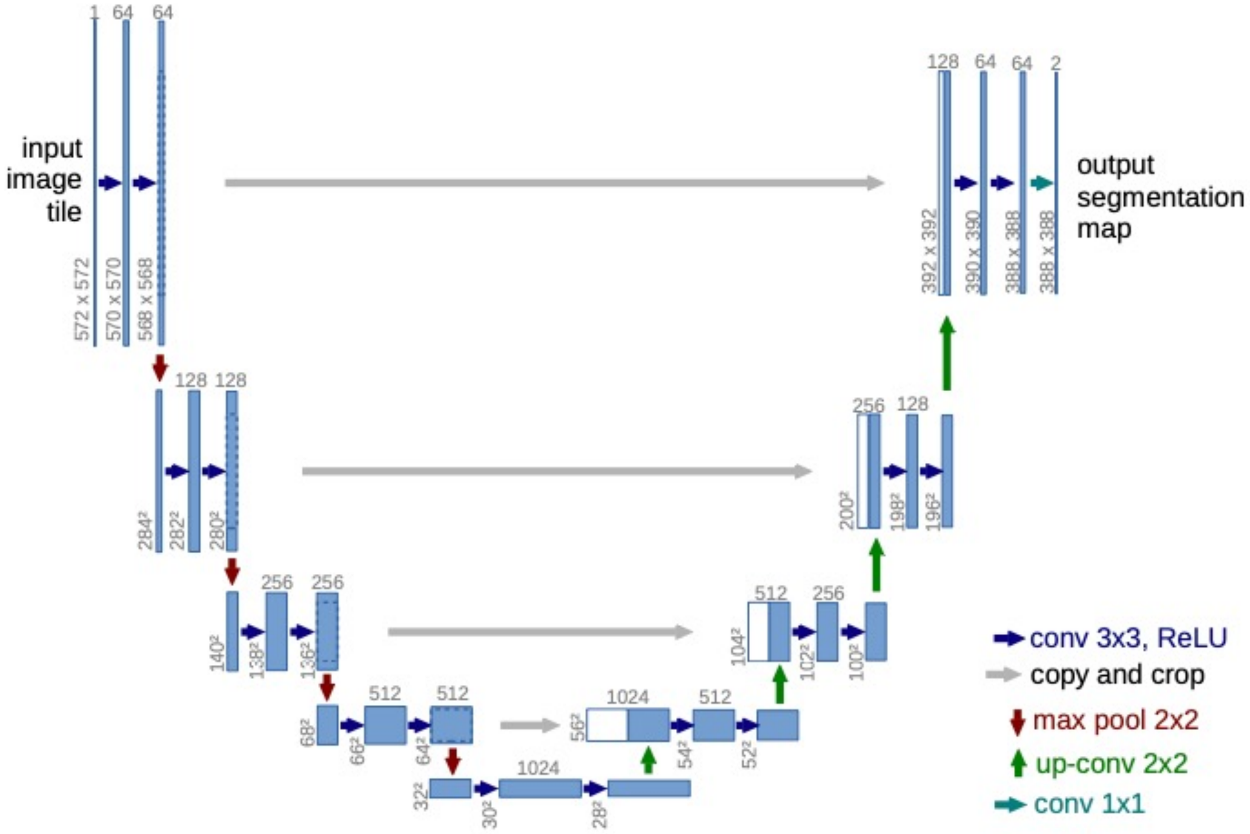
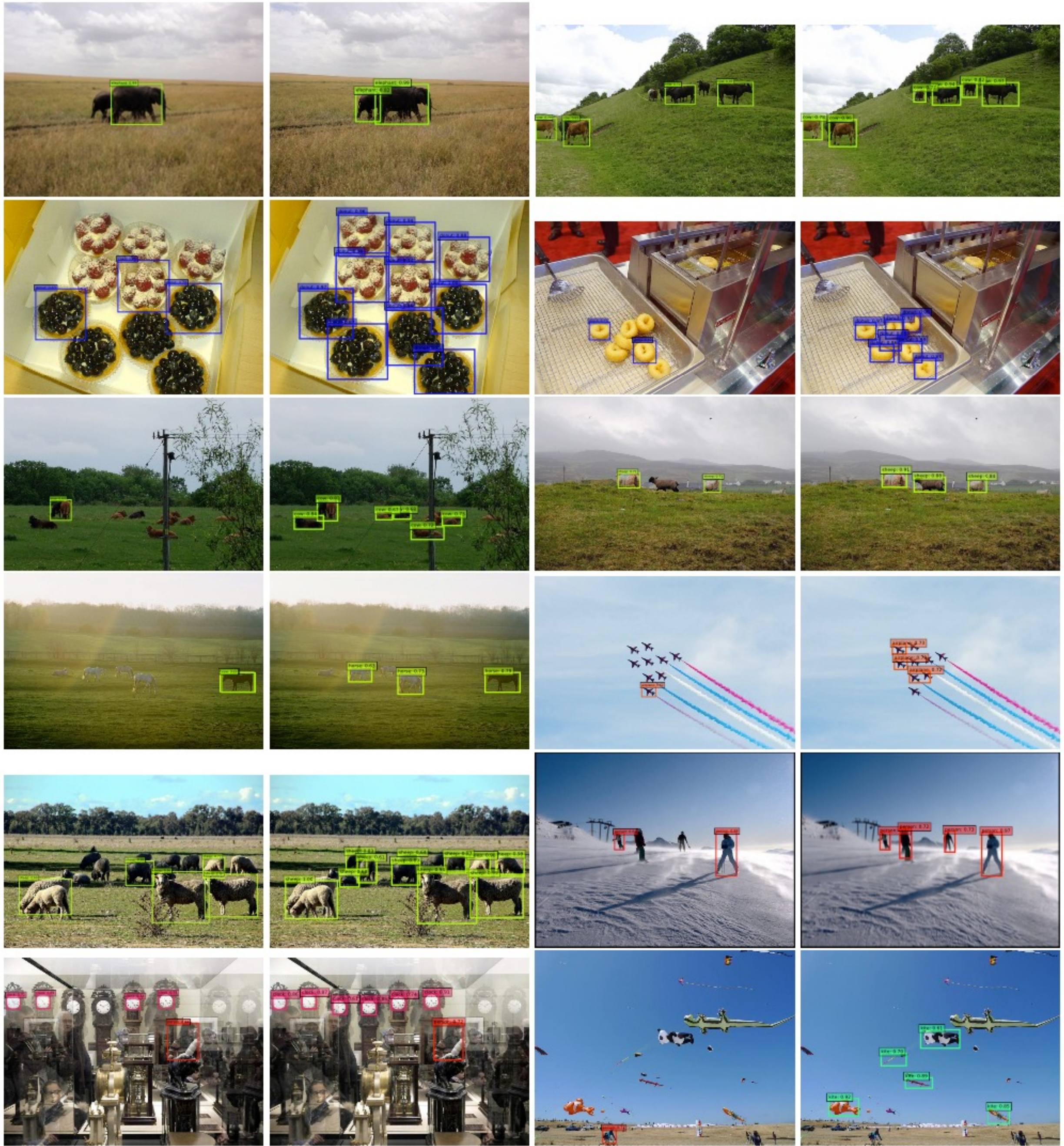


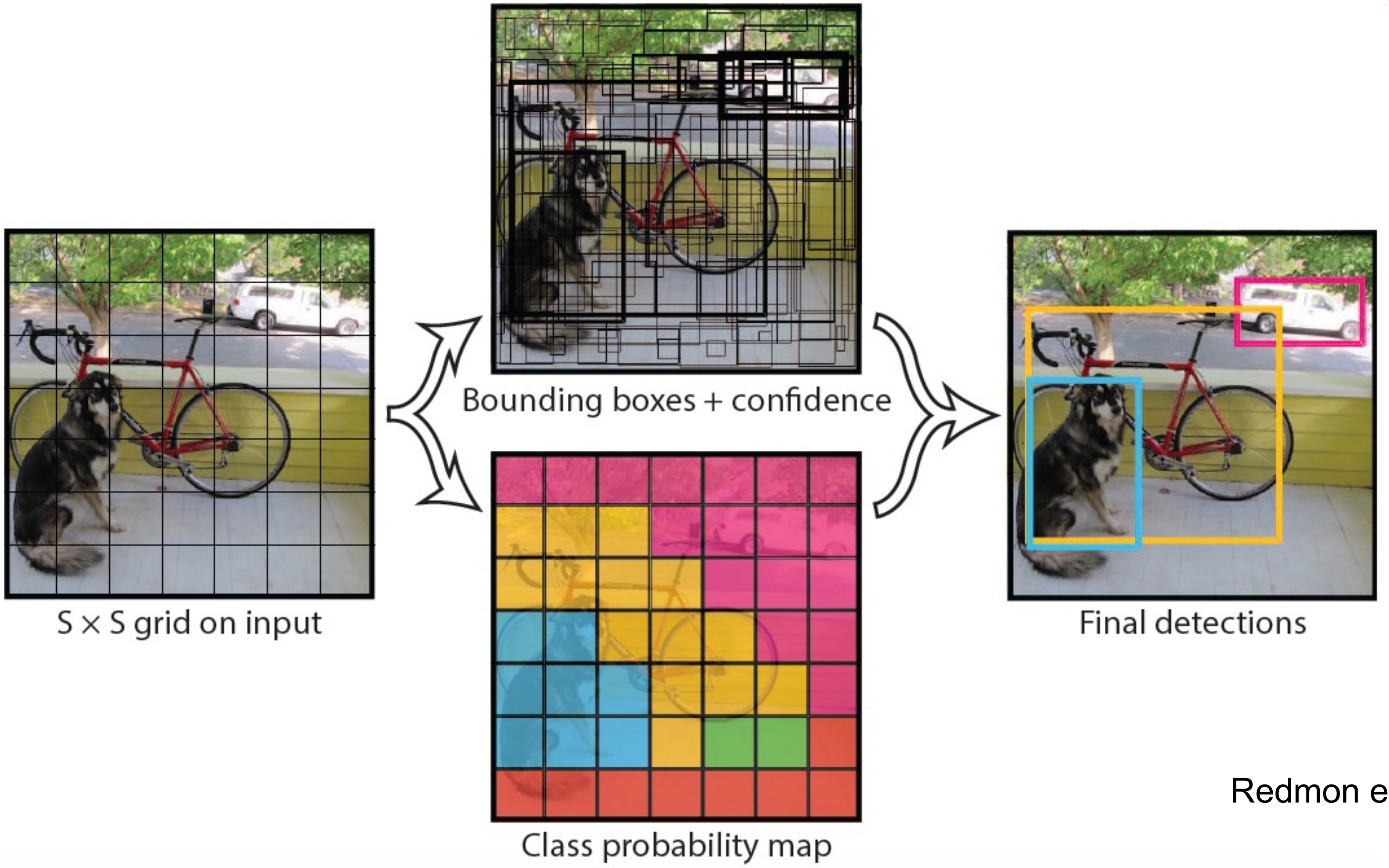
Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Examples

- The next slides how a few examples by YOLO, HTC and Mask R-CNN on object detection, classification and instance segmentation.



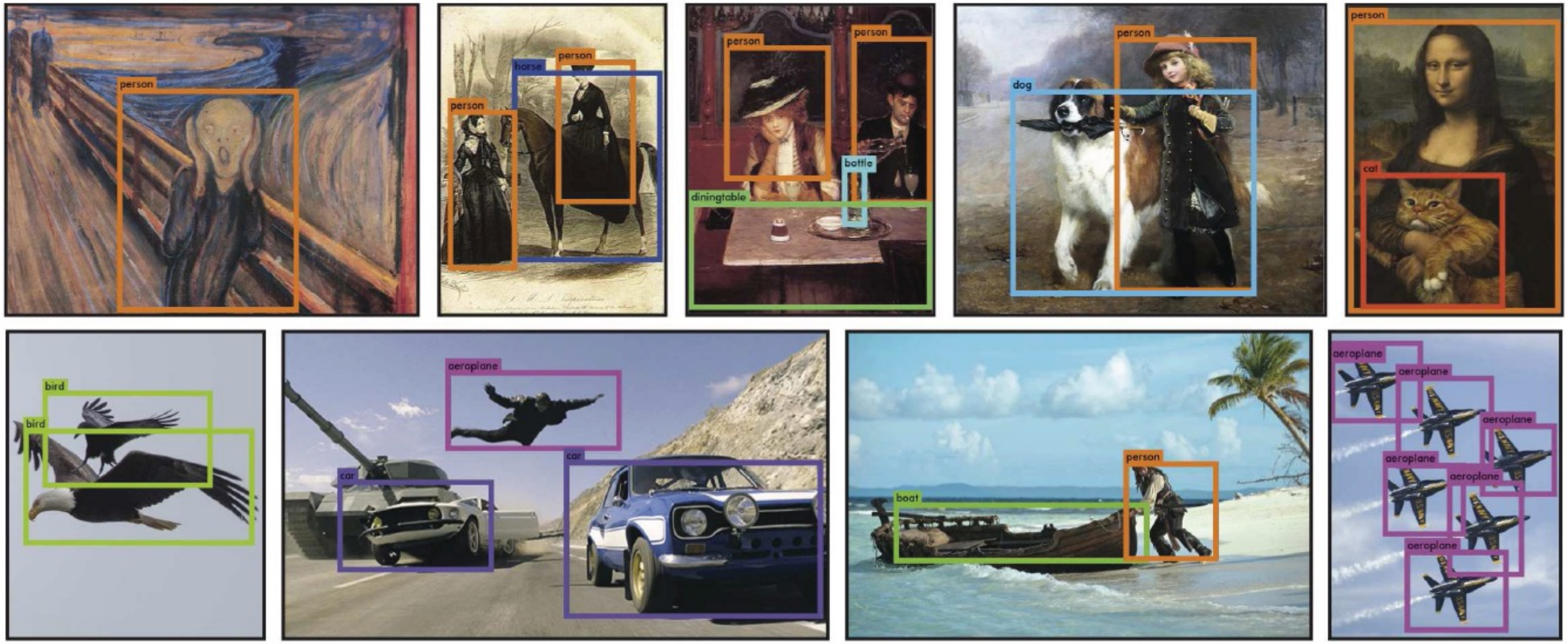
You Only Look Once (YOLO)



Redmon et al, 2016

In the paper, cell grid is 7 x 7 with B=2 bounding boxes per cell.
Therefore YOLO can only detect up to 98 objects in an image.

YOLO: Examples of results



Redmon et al, 2016

mask R-CNN results

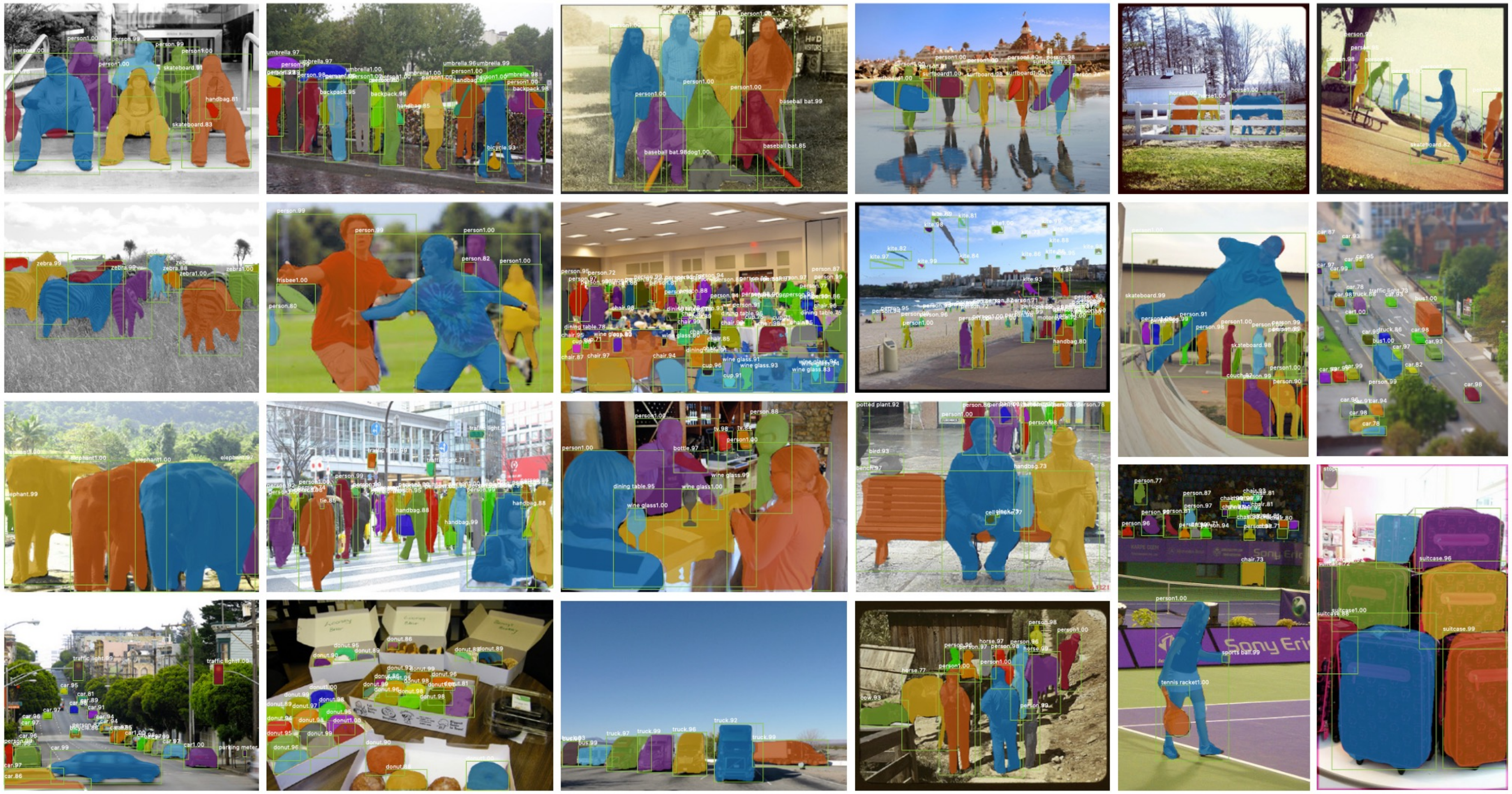


Figure 4. More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

HTC results on COCO

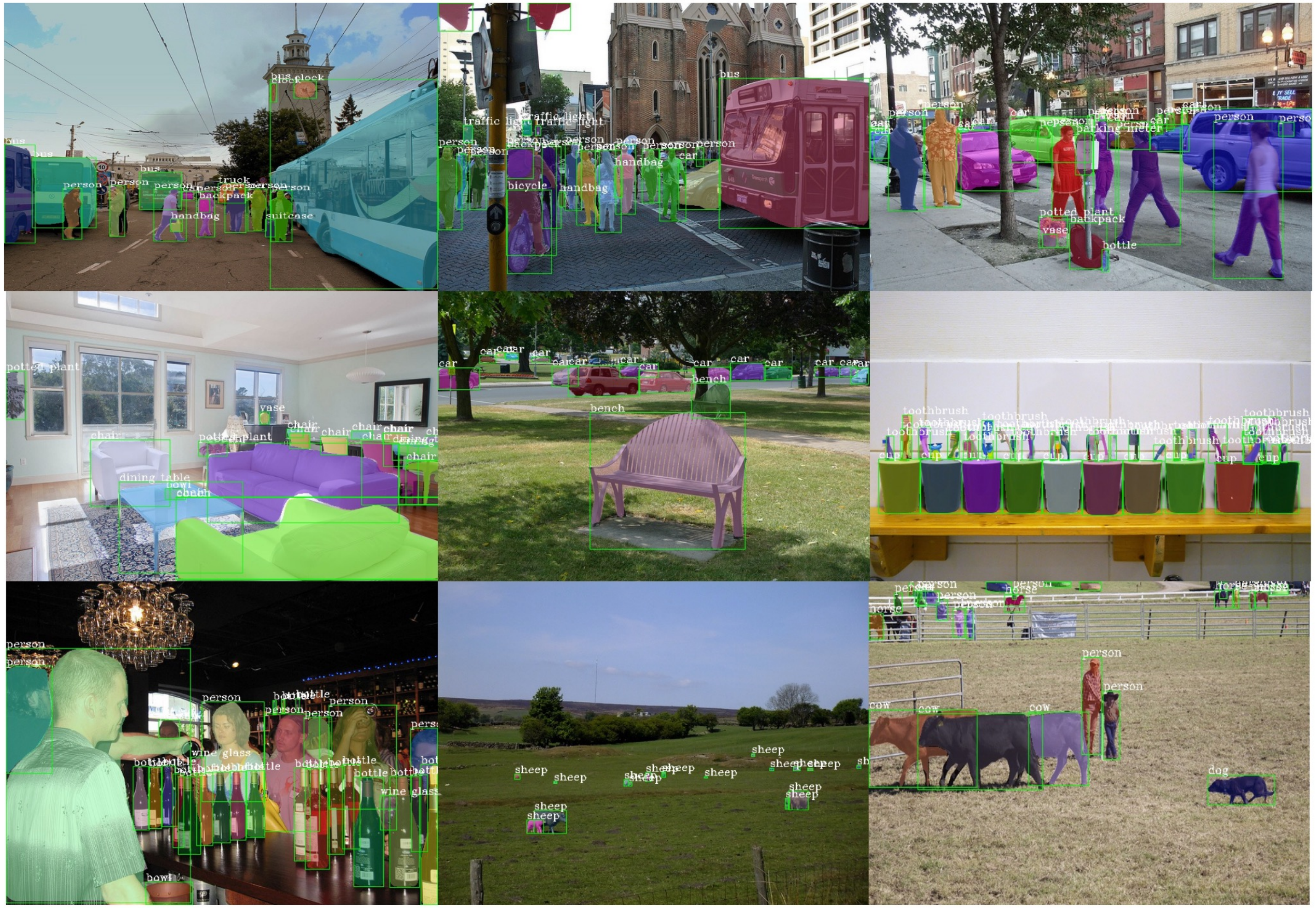


Figure 4: Examples of segmentation results on COCO dataset.

NO MORE CNN

- Questions (almost last chance)

Results from Evaluation

- You think you work too hard, but you have just used the allocated time.

- You are satisfied with the feedback and help from the TA’s.

- You would prefer that all lectures were recorded, but you are split wrt. the benefits of online teaching.

- You find the teaching material relevant, the academic level appropriate, and the course outcome useful.

- Do you have any comments, further suggestions etc. before we close ?

Farewell

We hope that you will be able to use the techniques that we have been through.

Best wishes for your future study and life

Francois and Søren,
Navid, Steffen, Peidi