Rethinking Atrous Convolution for Semantic Image Segmentation

LIANG-CHIEH CHEN, GEORGE PAPANDREOU, FLORIAN SCHROFF, HARTWIG ADAM

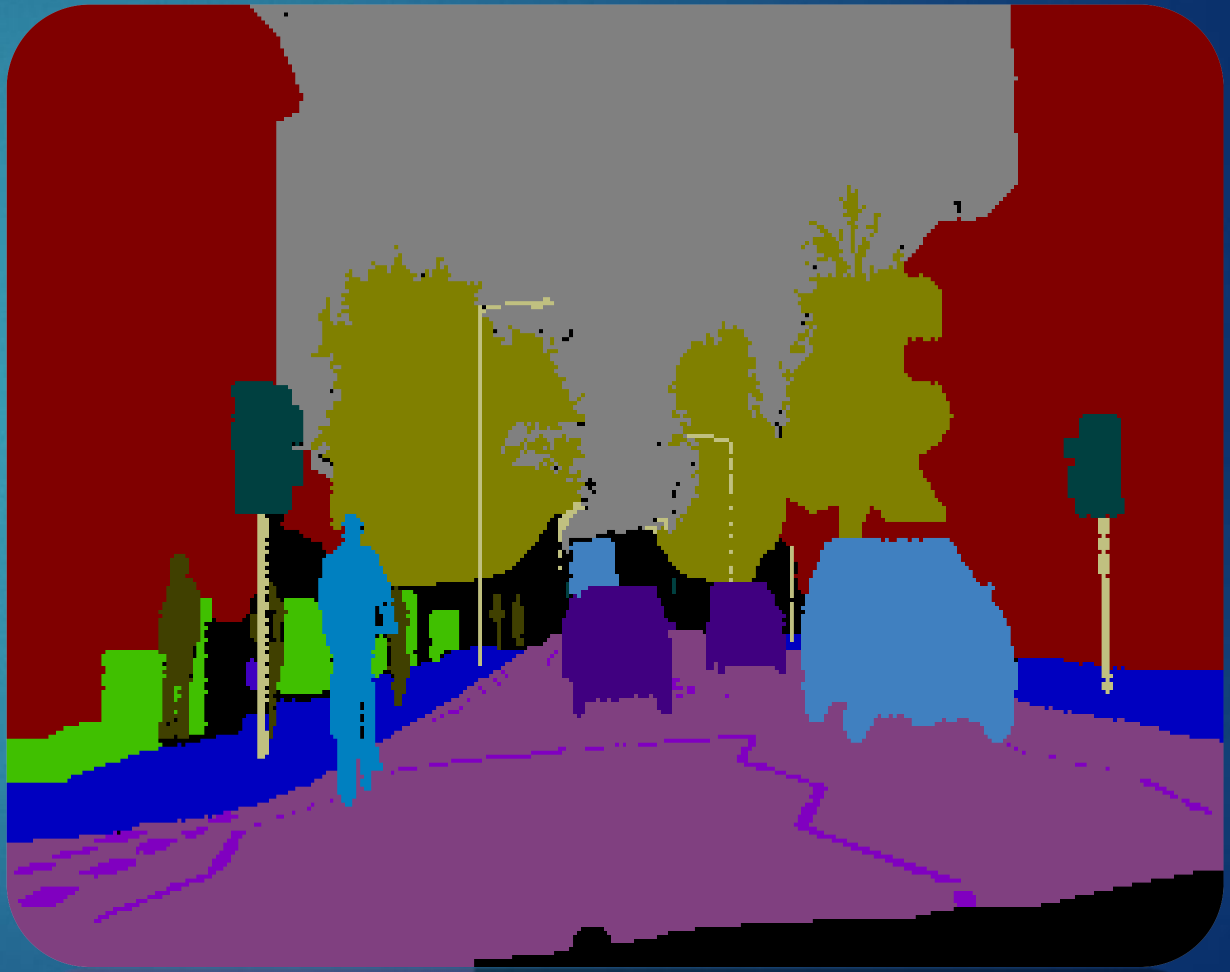
Sivan Doveh

Jenny Zukerman

Deep learning seminar 2017/18

Agenda

- Semantic segmentation
- Main idea of DeepLab
- DeepLabV1
- DeepLabV2
- DeepLabV3

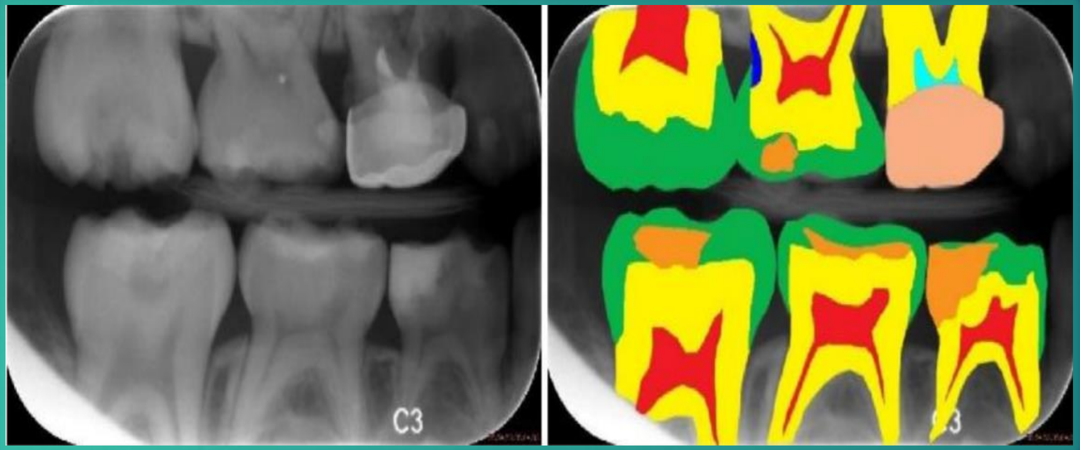


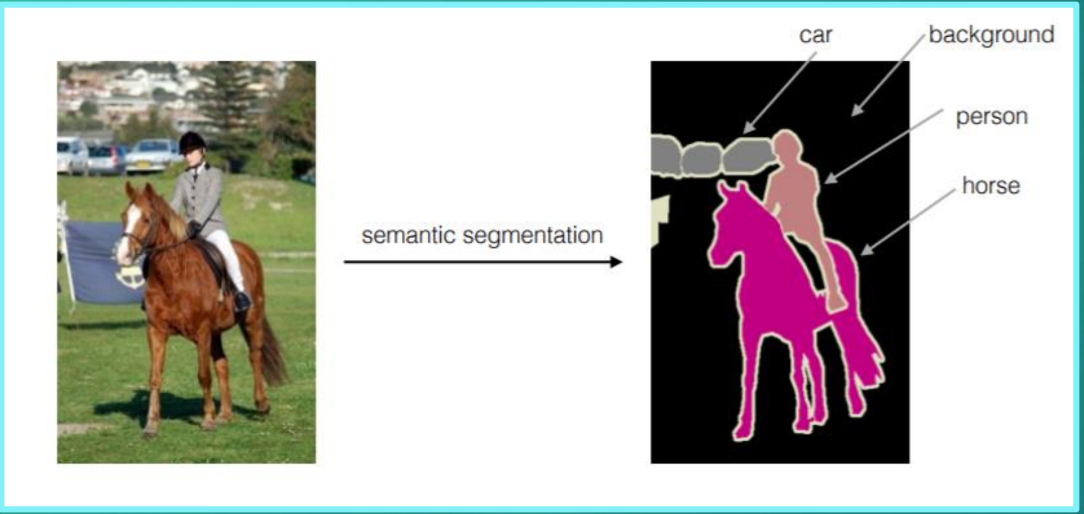
Semantic Segmentation

- Semantic Segmentation
  - **Semantic segmentation** is understanding an image at pixel level i.e, we want to assign each pixel in the image an object class
  - **Partitioning** an image into **regions** of **meaningful** objects.
  - Assign an object category label.

Semantic Segmentation

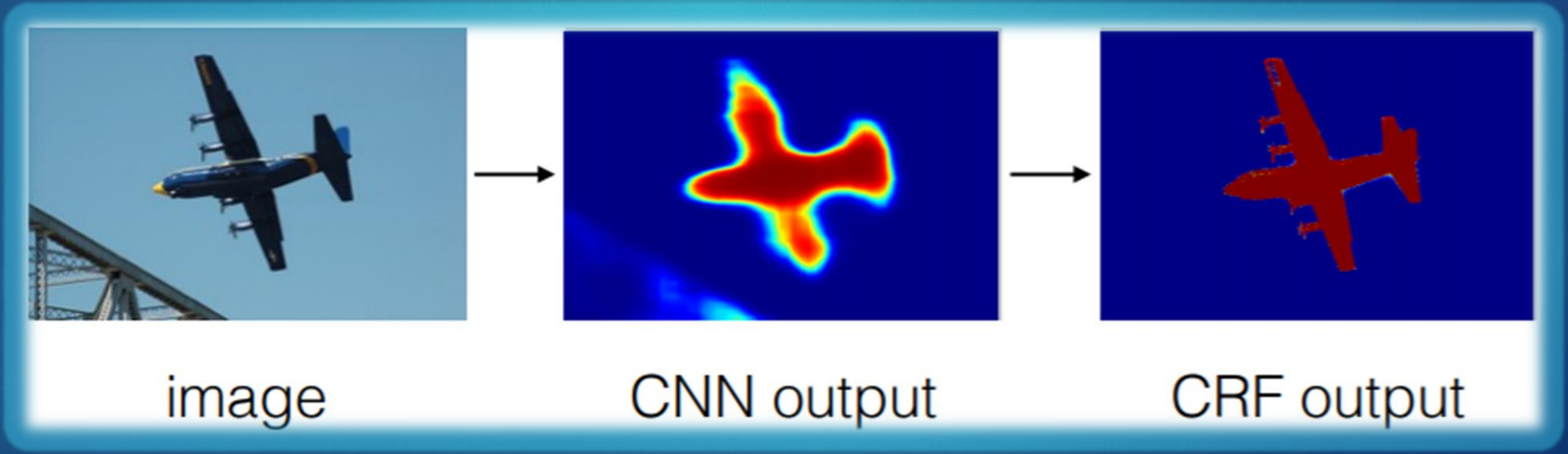
- Why semantic segmentation?
  - Autonomous driving
  - Medical purposes
- We will focus on 3 papers:
  - DeepLabV1
  - DeepLabV2
  - DeepLabV3



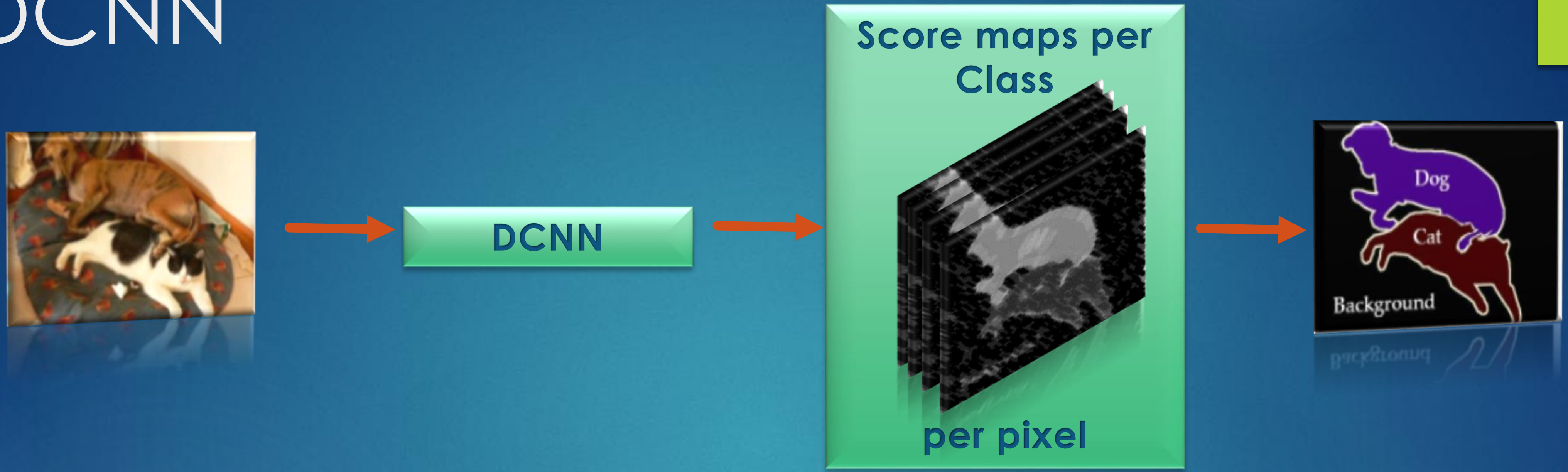


DeepLabV1 & DeepLabV2

- Use DCNN for classification to generate a rough prediction of segmentation (smooth, blurry heat map)
- Refine prediction with conditional random field (CRF)

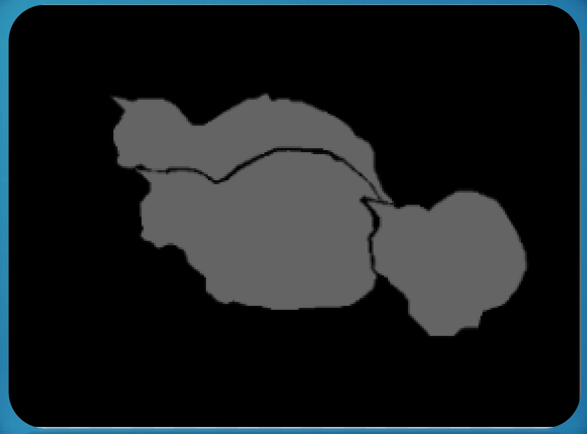


DCNN



- <u>What happens in standard DCNN?</u>
  - <u>Striding-</u>Smaller output size
  - <u>Pooling-</u>Invariance to small translations of the input

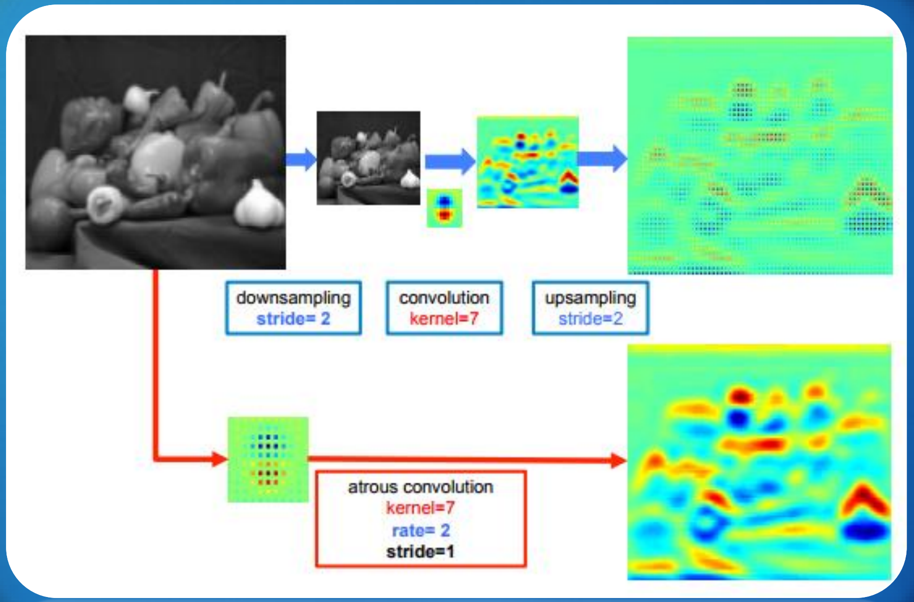
- <u>DeepLab solution</u>
  - Atrous convolution
  - CRFs (Conditional Random Fields)



DCNN – Atrous (Holes)

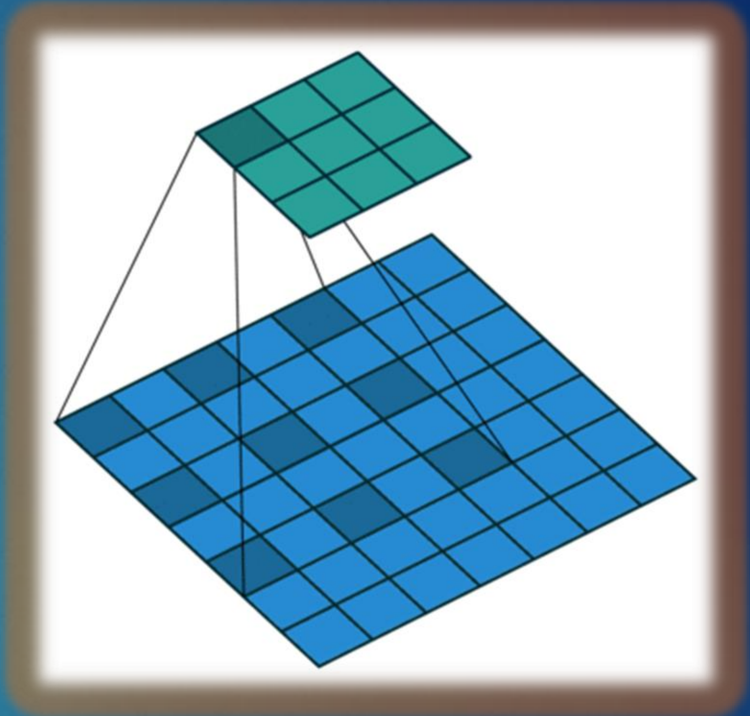
- Remove the last 2 pooling layers.
- Up-sample the original filter by a factor of the strides (rate = 2)

- Standard convolution → **responses at only 1/4 of the image positions.**
- Convolve image with a filter 'with holes' → **responses at all image positions**



DCNN – Atrous (Holes)

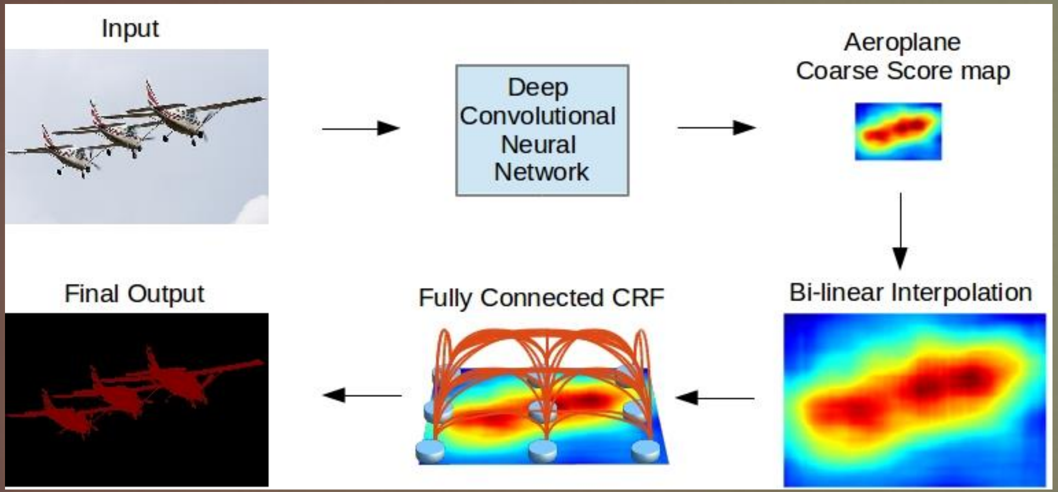
- Small field-of-view → accurate localization
- Large field-of-view → context assimilation
- Effective filter size increases

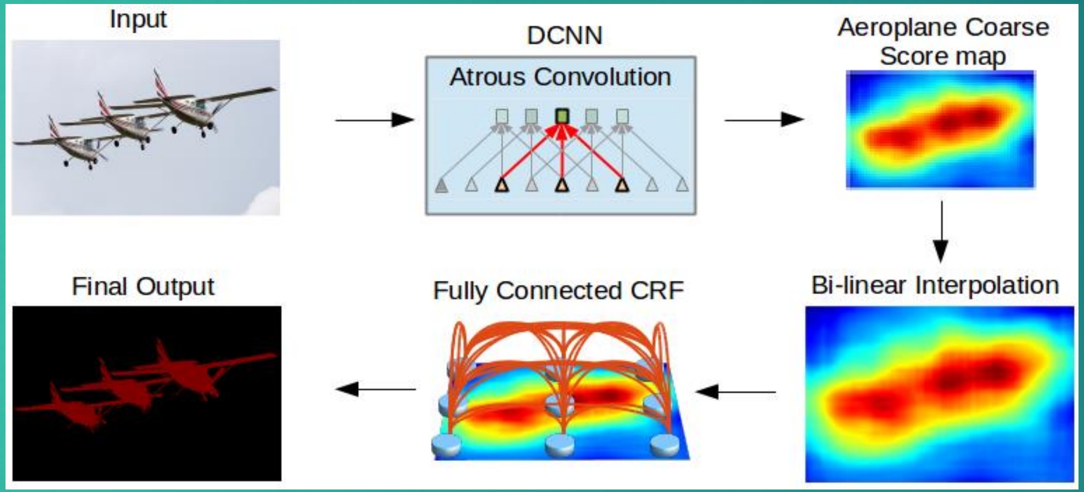


- Both the number of filter **parameters** and the number of **operations** per position stay **constant**

DCNN – Atrous (Holes)

- The authors found a good **efficiency/accuracy trade-off**, using atrous convolution to increase by a factor of 4 the density of computed feature maps, followed **by bilinear** interpolation (factor 8)





- the proposed approach **converts image classification** networks into **dense feature extractors** without requiring learning any extra parameters

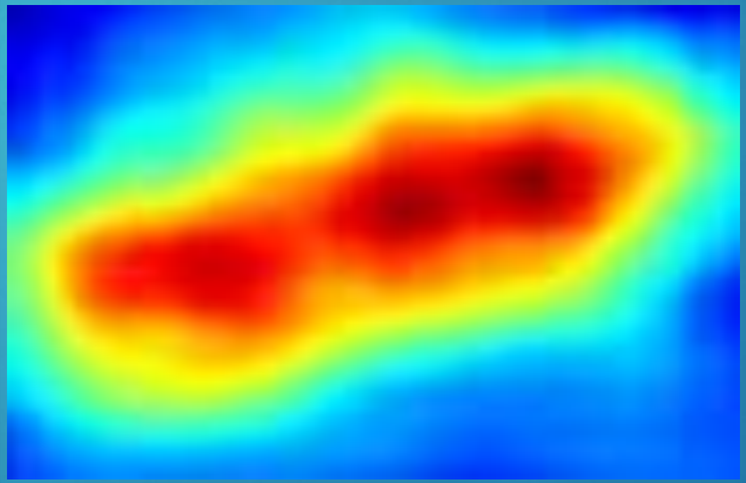
Atrous (Holes)-Some Results

Kernel	Rate	FOV	Params	Speed	bef/aft CRF
7×7	4	224	134.3M	1.44	64.38 / 67.64
4×4	4	128	65.1M	2.90	59.80 / 63.74
4×4	8	224	65.1M	2.90	63.41 / 67.14
3×3	12	224	20.5M	4.84	62.25 / 67.64

Conditional Random Field (CRF)

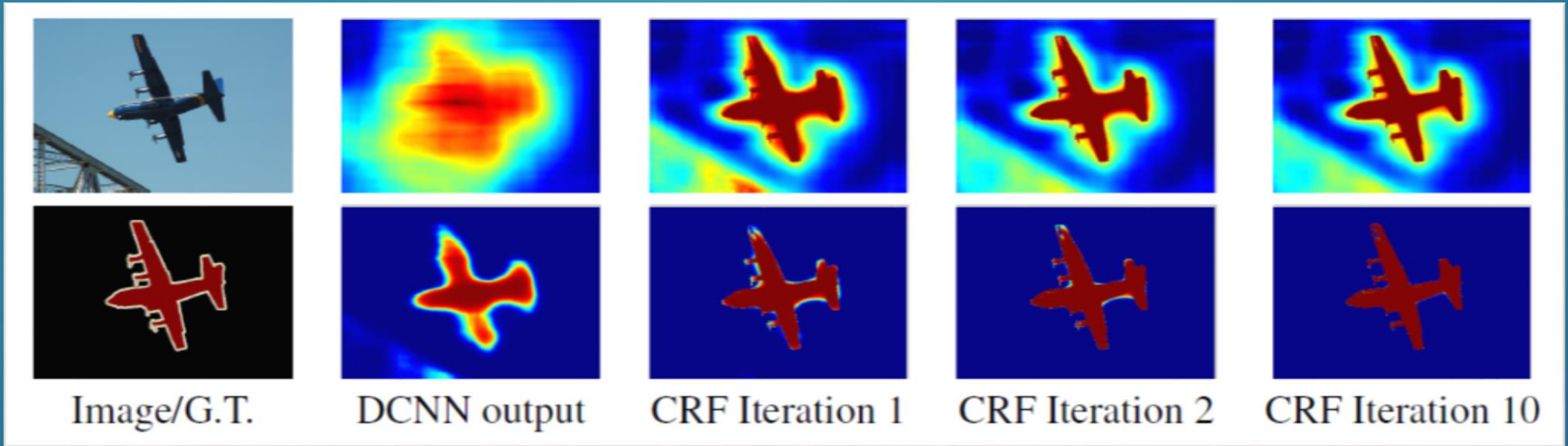
- DCNN trade-off:
  Classification accuracy $\leftrightarrow$ Localization accuracy

- DCNN score maps successfully predict classification and rough position.
- Less effective for exact outline.



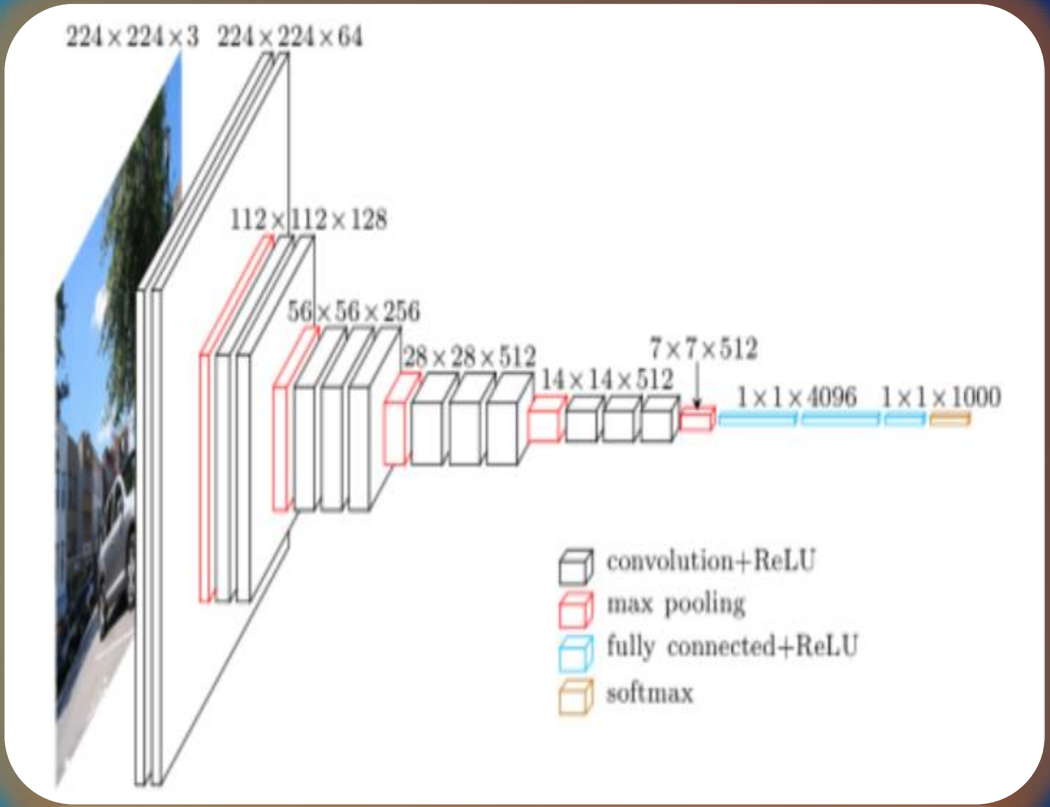
Conditional Random Field (CRF)

- CRF tries to model the relationship between pixels:
  - Nearby pixels more likely to have same label
  - CRF takes into account the label assignment probability at a pixel
  - Refine results by iterations

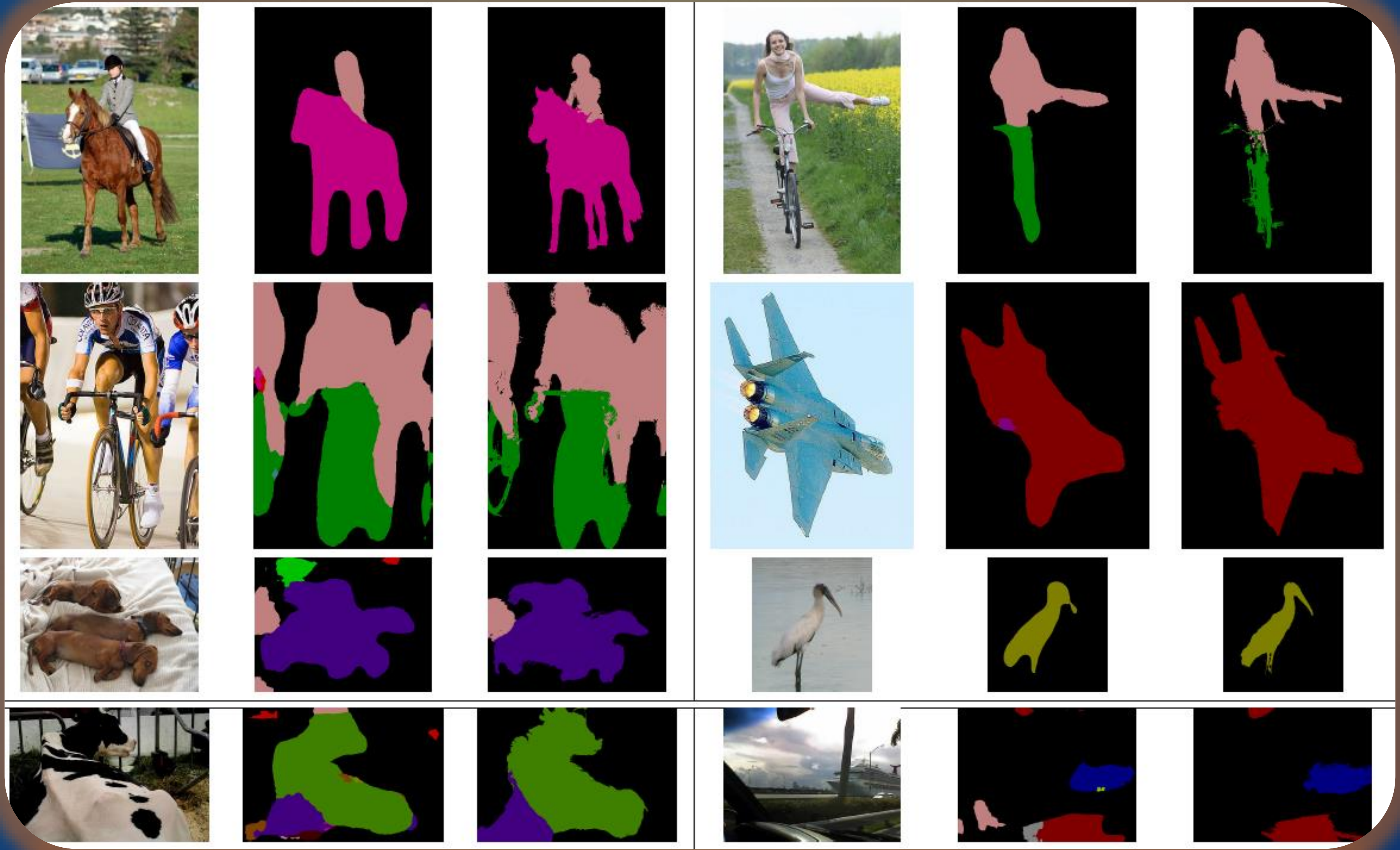


DeepLabV1

- DeepLab v1 is constructed by modifying VGG-16
- Fully connected layers of VGG-16 are converted to convolutional layers
- Subsampling is skipped after last two max-pooling layers
- Convolutional filters in the layers that follow pooling are modified to atrous
- Model weights of Imagenet-pretrained VGG-16 network are finetuned

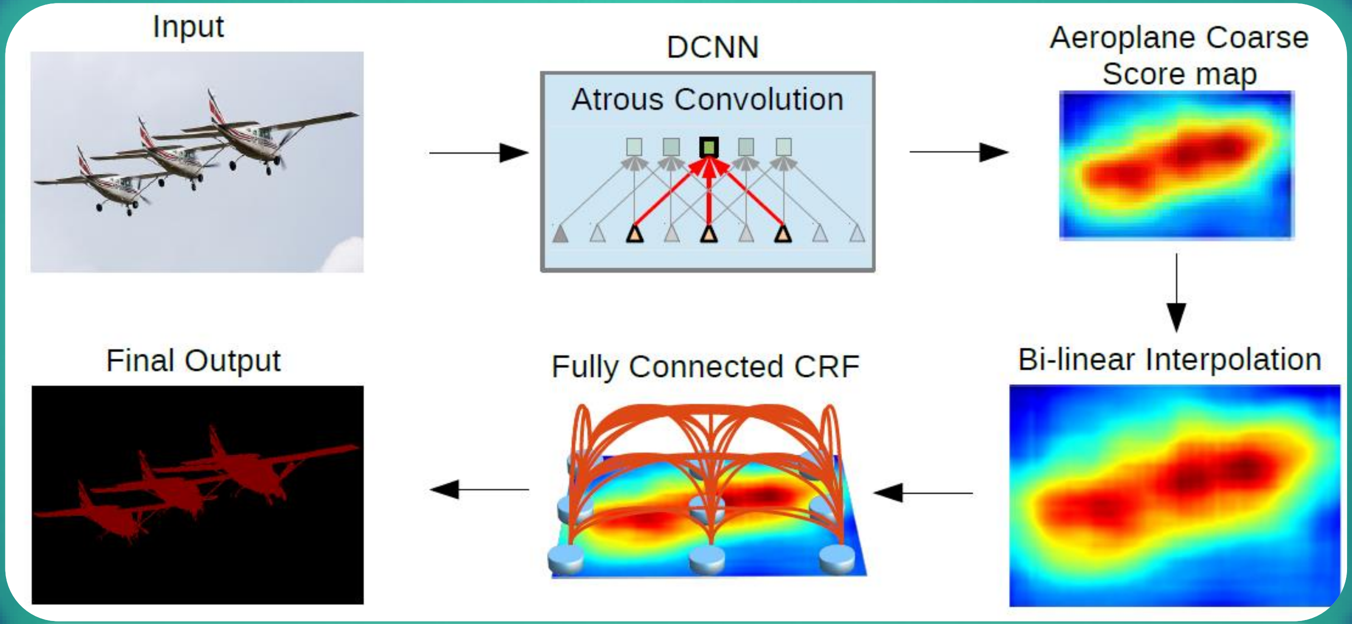


DeepLabV1 visualization



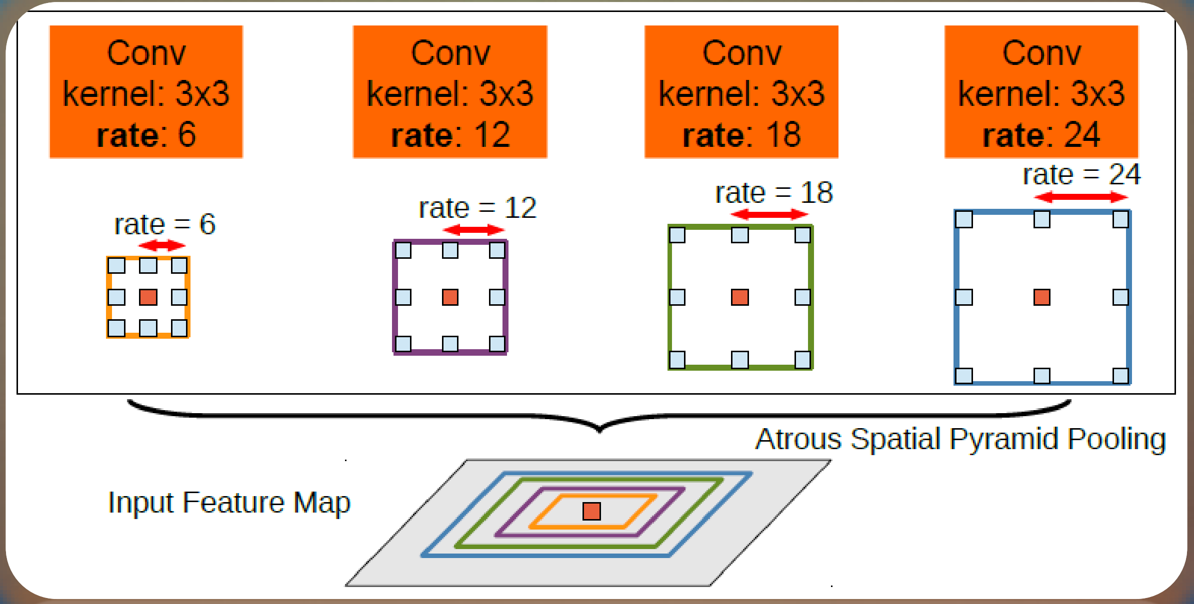
DeepLabV2

- Improvements compared to DeepLabV1:
  - Better segmentation of objects at multiple scales (using ASPP)
  - Adapting ResNet image classification DCNN
  - Learning rate policy



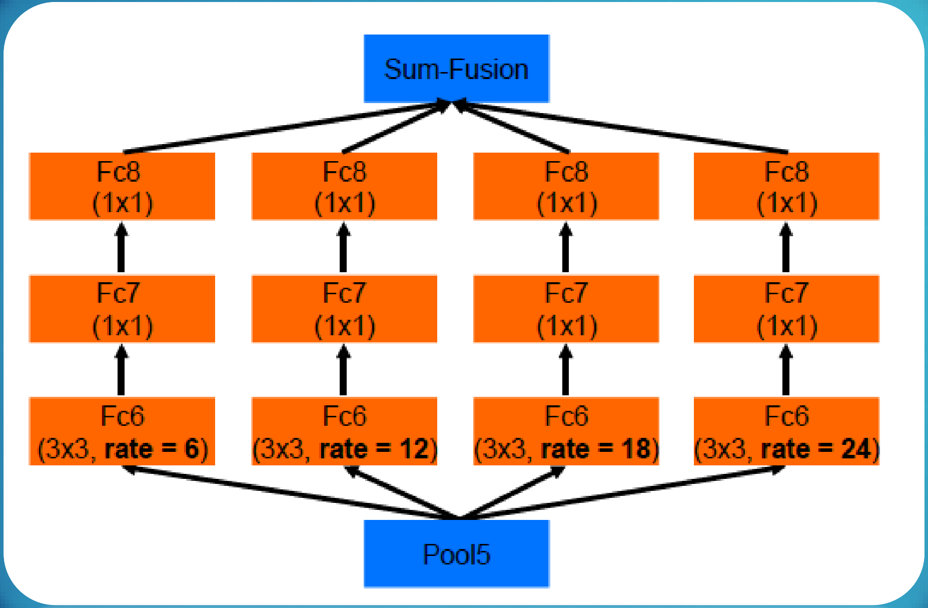
Atrous Spatial Pyramid Pooling (ASPP)

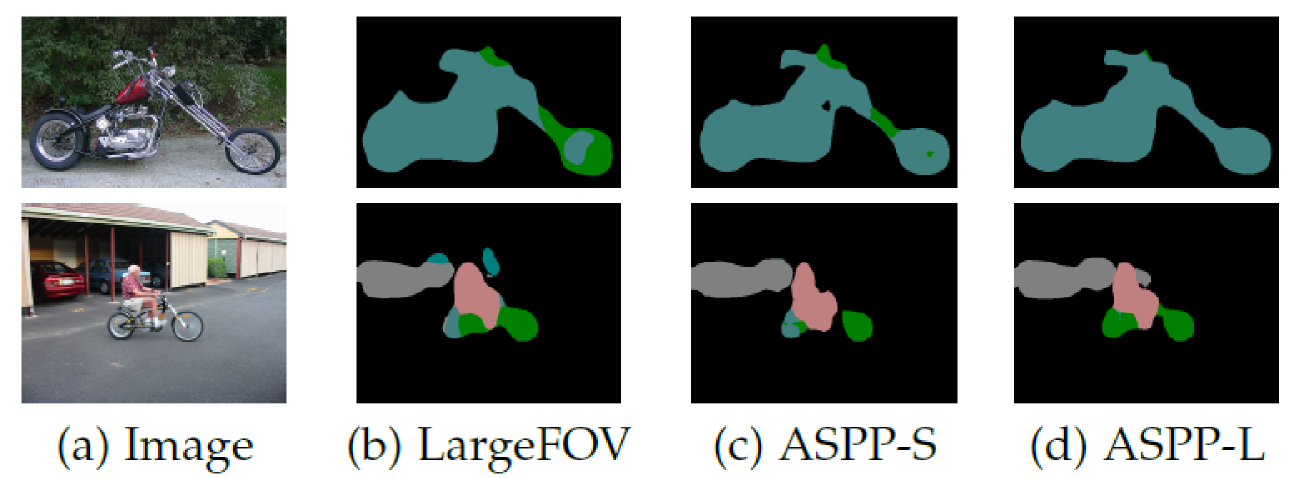
- Challenge: existence of objects at multiple scales
- Computationally efficient scheme of resampling a given feature layer at multiple rates prior to convolution
- Using multiple parallel atrous convolutional layers with different sampling rates



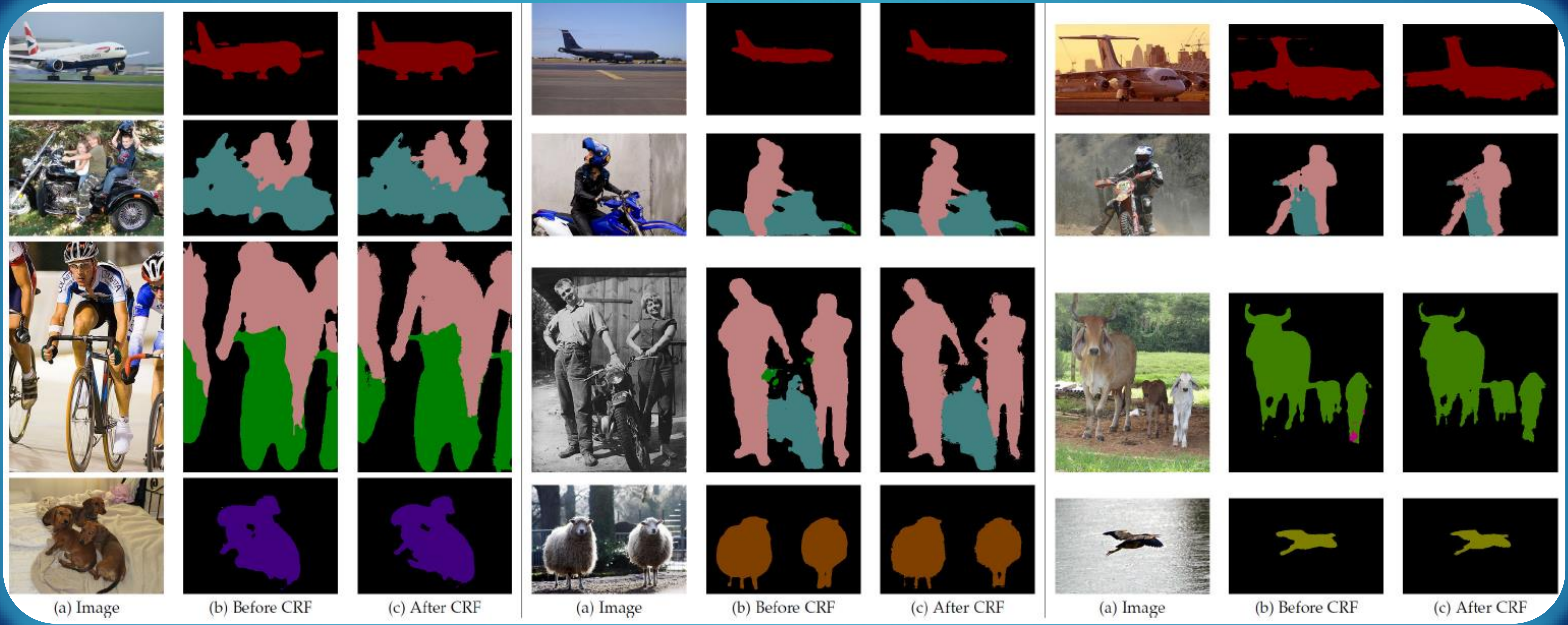
Atrous Spatial Pyramid Pooling (ASPP)

- The features extracted for each sampling rate are further processed in separate branches and fused to generate the final result





DeepLabV2 Visualization

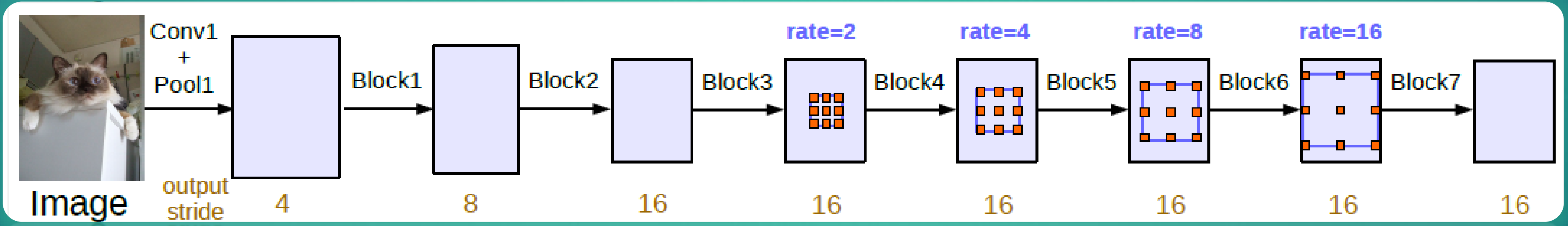


DeepLabV1 & DeepLabV2

- Advantages:
  - **Speed:** By virtue of the 'atrous' algorithm, dense DCNN operates at 8 fps, while fully-connected CRF requires 0.5 second
  - **Accuracy:** state-of-the-art results achieved on several state-of-art datasets
  - **Simplicity:** the system is composed of a cascade of two fairly well-established modules, DCNNs and CRFs

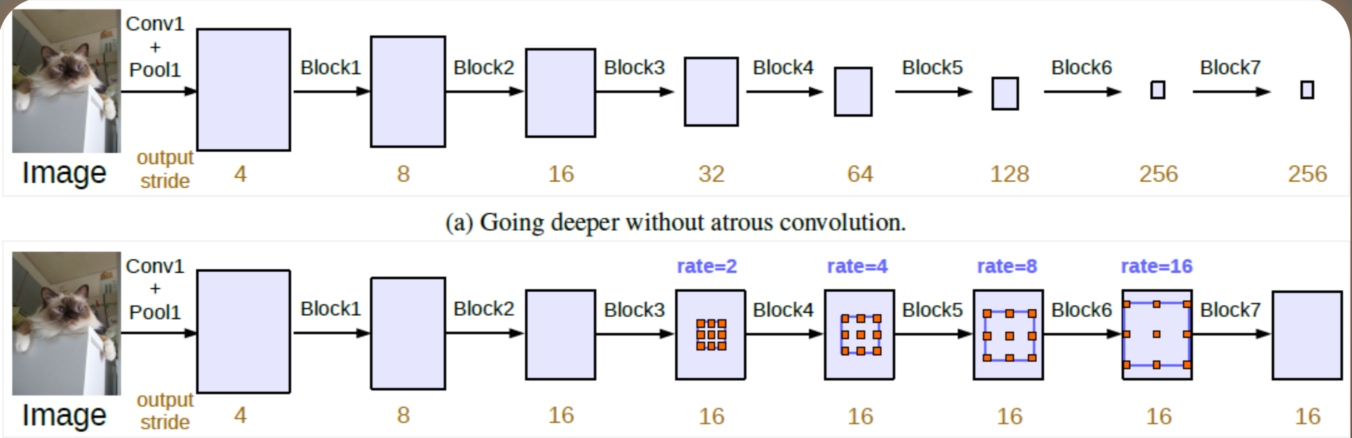
DeepLabV3

- Changes compared to DeepLabV1 & DeepabV2:
  - The proposed framework is general and could be applied to any network
  - Several copies of the last ResNet block are duplicated, and arranged in cascade
  - Batch normalization is included within ASPP
  - CRF is not used



Inside ResNet Block

- Duplicate several copies the last ResNet block (Block 4) and arrange in cascade
  - In the proposed model, blocks 5-7 are duplicates of block 4
- Three convolutions in each block
- Last convolution contains stride 2 except the one in last block
- In order to maintain original image size, convolutions are replaced with atrous convolutions with rates that differ from each other with factor 2

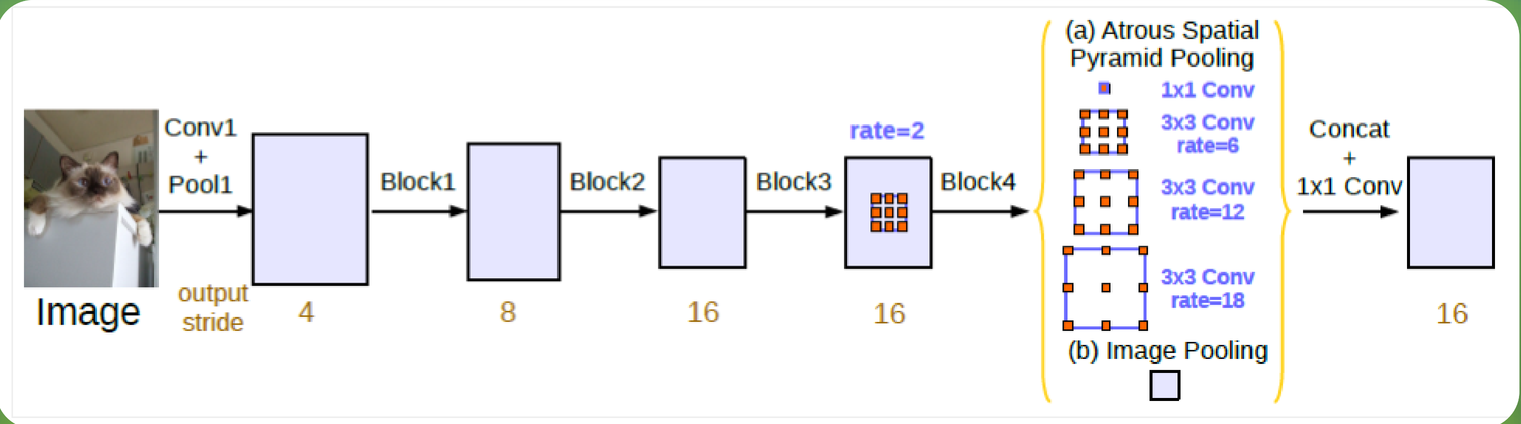


(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

Figure 3. Cascaded modules without and with atrous convolution.

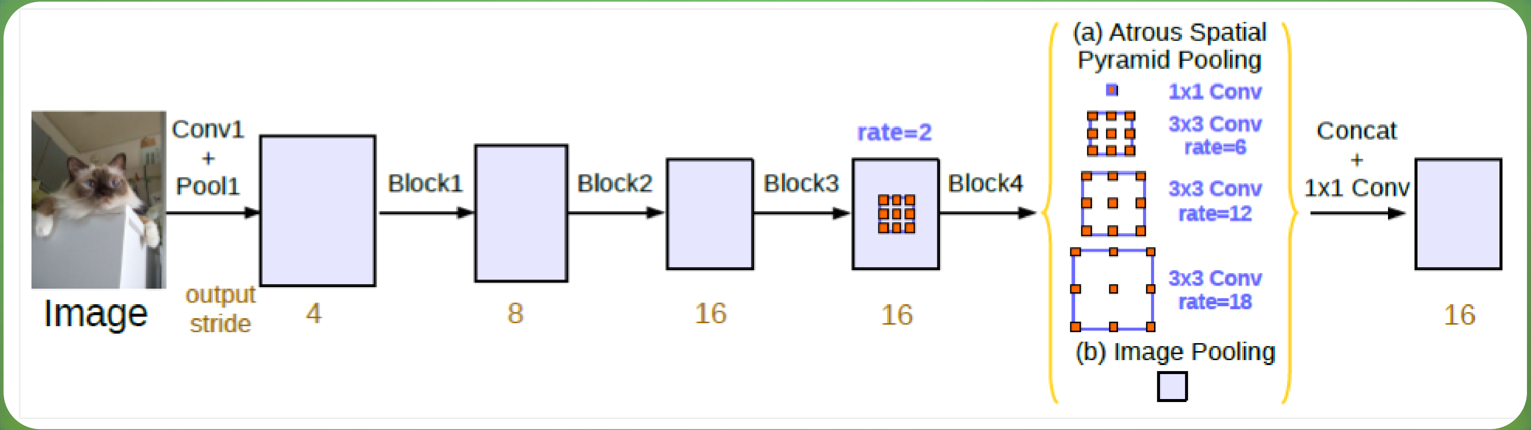
DeepLabV3 - ASPP

- Batch normalization is included within ASPP
- As the sampling rate becomes larger, number of valid filter weights becomes smaller
- Global average pooling on last feature map of the model



DeepLabV3 - ASPP

- Improved ASPP consists:
  - One 1x1 convolution and three 3x3 convolutions with (6,12,18) rates - all with 256 filters and batch normalization
  - image-level features (global average pooling)
- Resulting features from all branches are concatenated and pass through 1x1 convolution

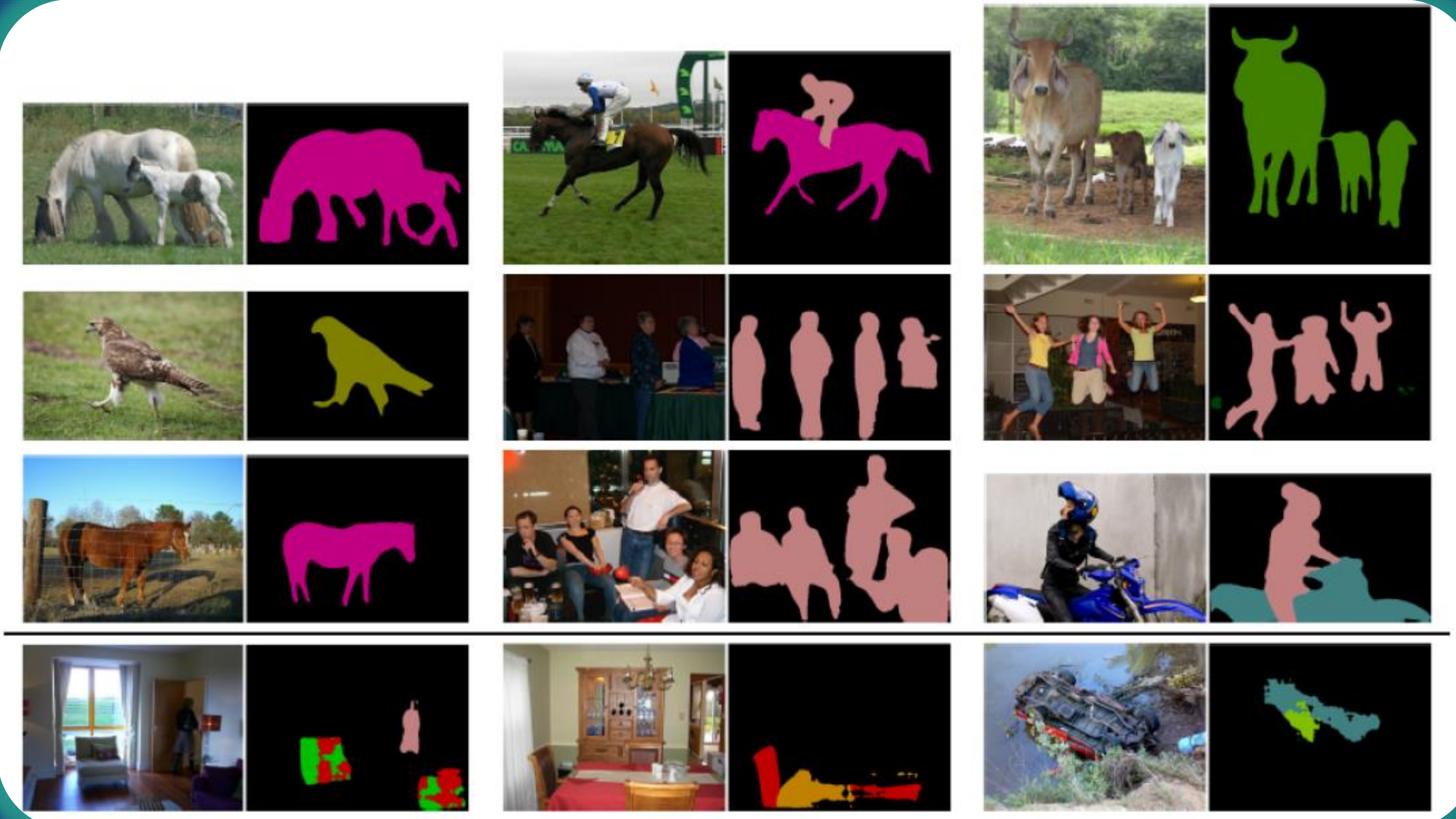


DeepLabV3 - Results

- Best result includes:
  - ASPP
  - Output stride of 8
  - Flip and rescale augmentation
- **Outperforms DeepLabV2 (77.69%)**

Method	OS=16	OS=8	MS	Flip	mIOU
MG(1, 2, 4) + ASPP(6, 12, 18) + Image Pooling	✓				77.21
		✓			78.51
		✓	✓		79.45
		✓	✓	✓	$\boxed{79.77}$

DeepLabV3 Visualization



Questions?

student

student

student

student

Student with question