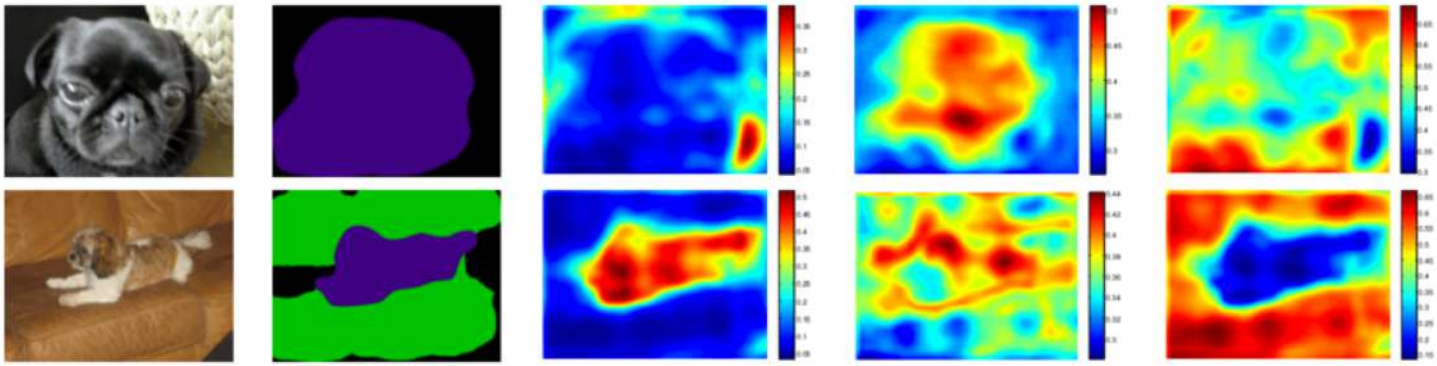
CVPR 2016

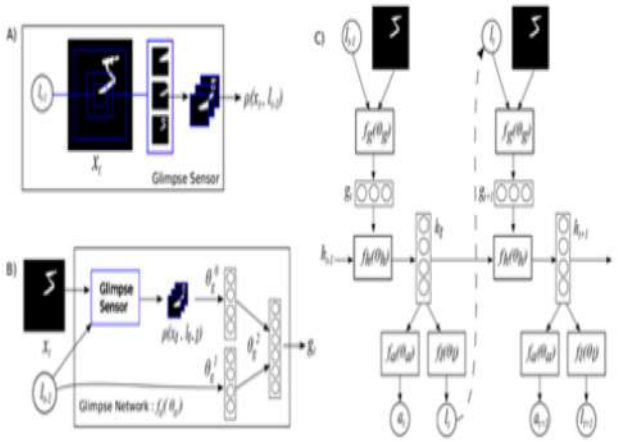
Attention to Scale: Scale-Aware Semantic Image Segmentation

Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, Alan L. Yuille
Baidu Research, Institute of Deep Learning (IDL)

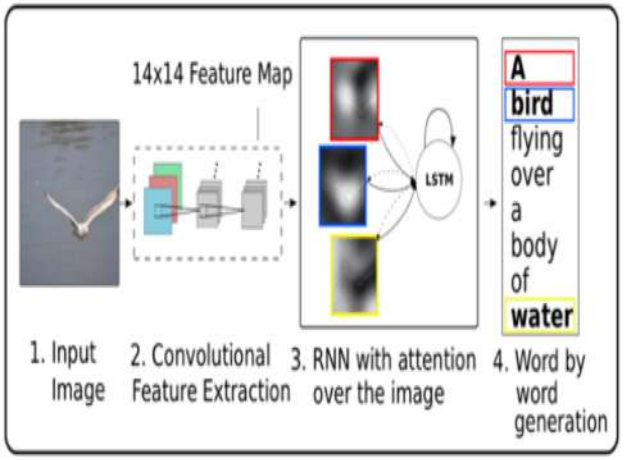


Background – Visual Attention Models

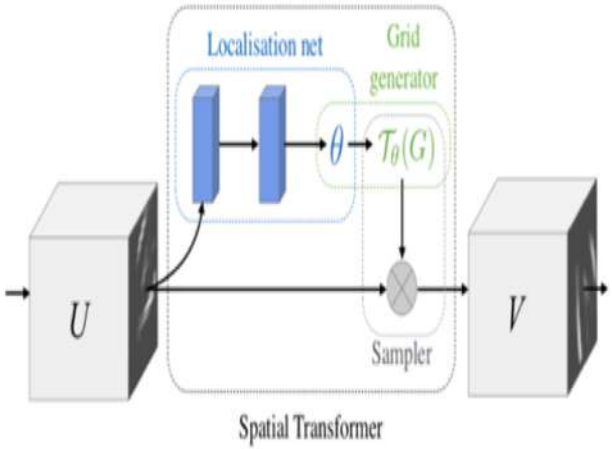
- Focus on discriminative locations / parts, reduce computational burden



[1]



[2]



[3]

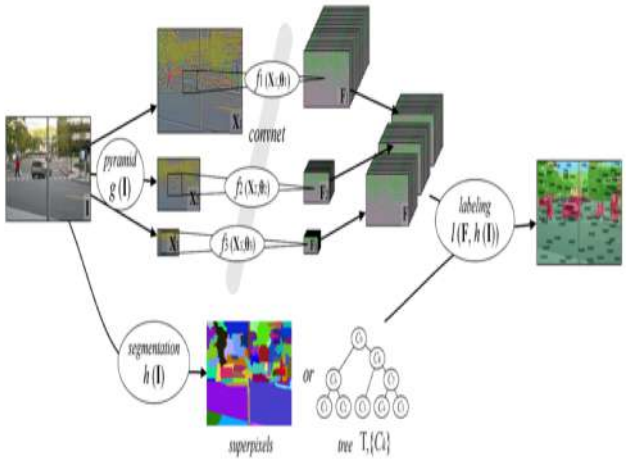
[1] Mnih et al, Recurrent Models of Visual Attention, NIPS 2014

[2] Xu et al, Show, attend and tell: Neural Image Caption Generation with Visual Attention, ICML 2015

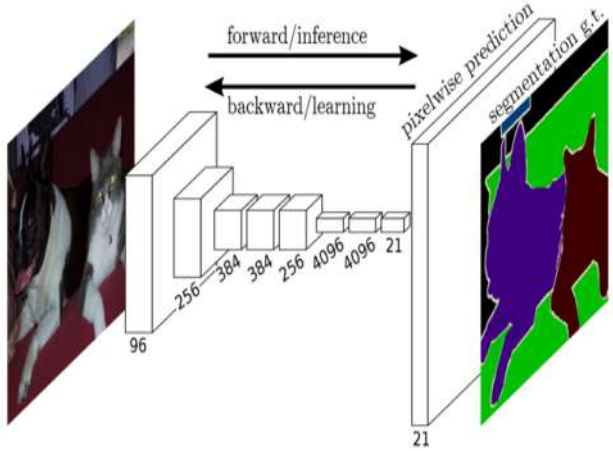
[3] Jaderberg et al, Spatial transformer networks, NIPS 2015

Background – Semantic Image Segmentation

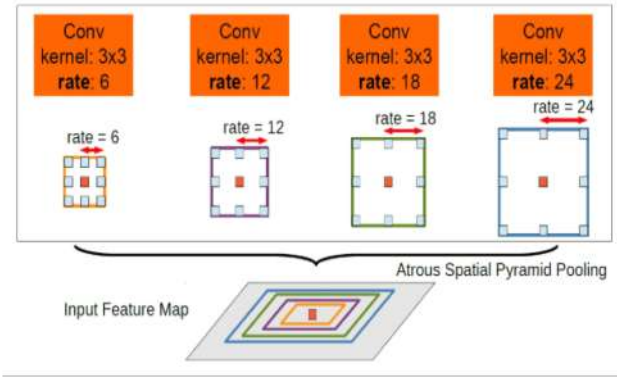
- Pixelwise prediction of object class labels



[1]



[2]



[3]

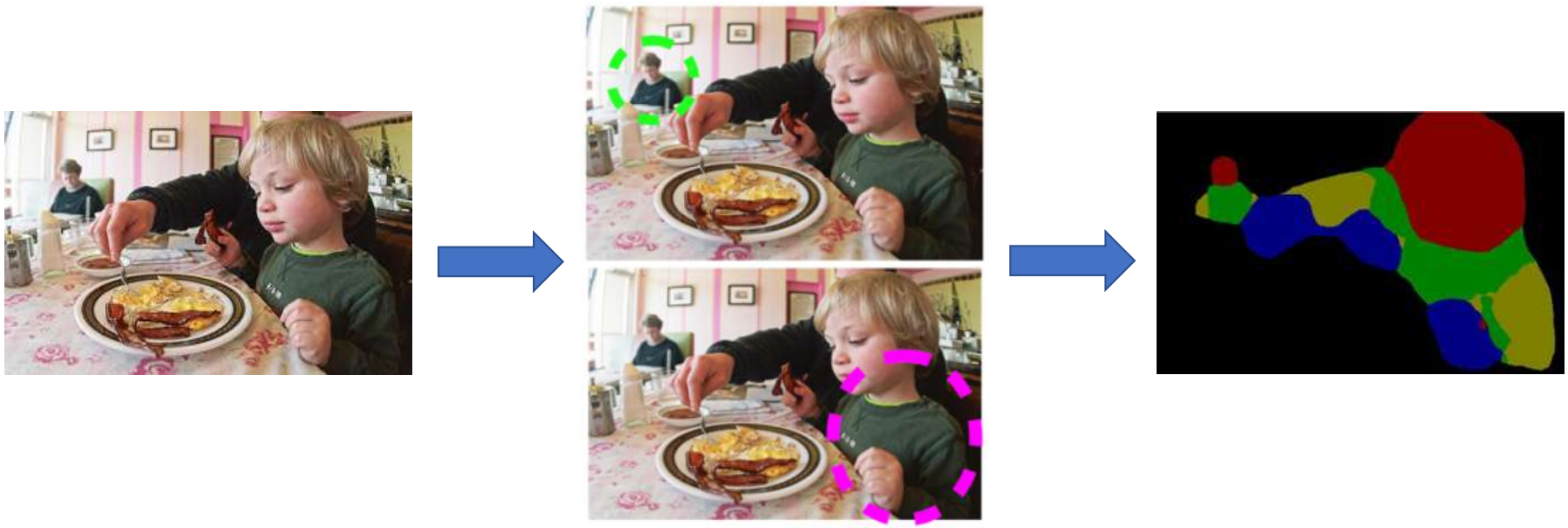
[1] Farabet et al, Learning Hierarchical Features for Scene Labeling, TPAMI 2013

[2] Long et al, Fully Convolutional Networks for Semantic Segmentation, CVPR 2015

[3] Chen et al, DeepLab: Semantic Image Segmentation with Atrous Convolution, PAMI 2017

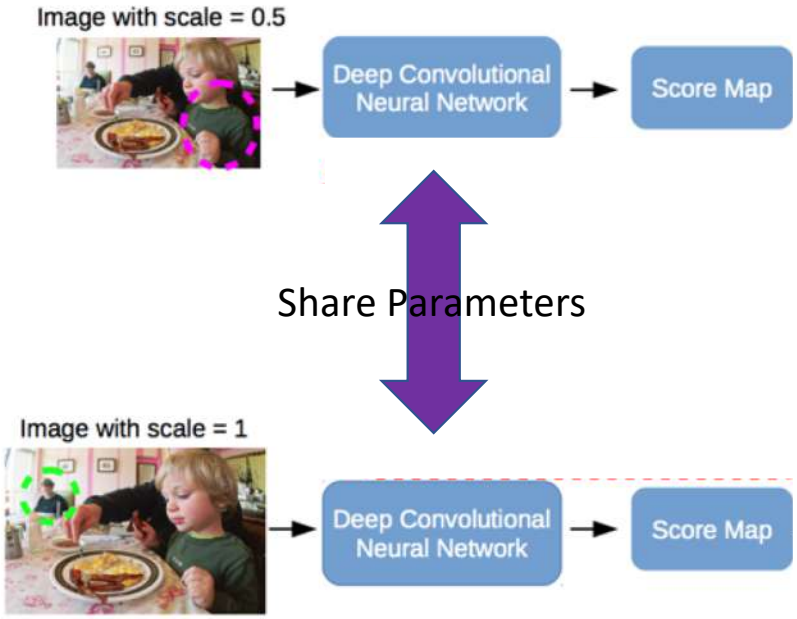
Motivation - Pixelwise Attention to Scale

- The scale is a factor of both object size and depth to the camera.
- When objects are large, the visual receptive field should also be large.



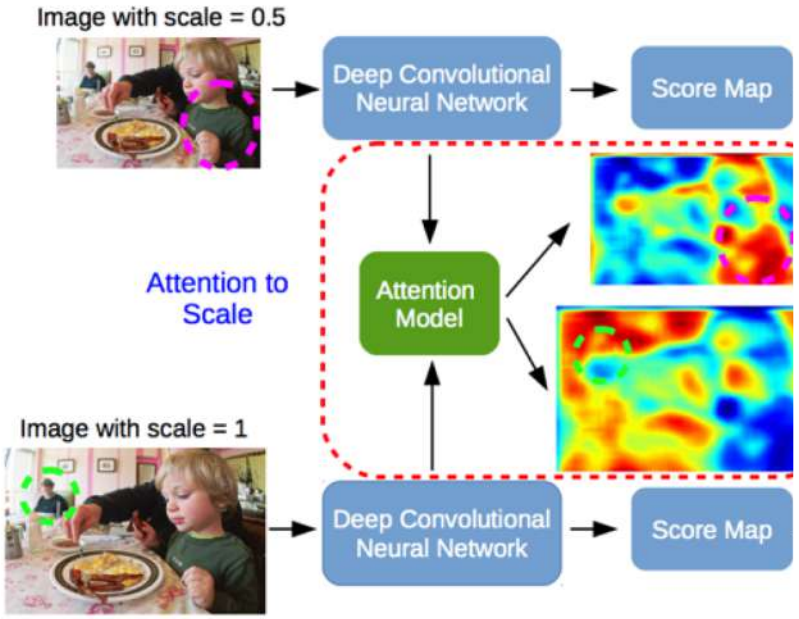
Model Architecture

- Allowing networks share parameters at different scales.



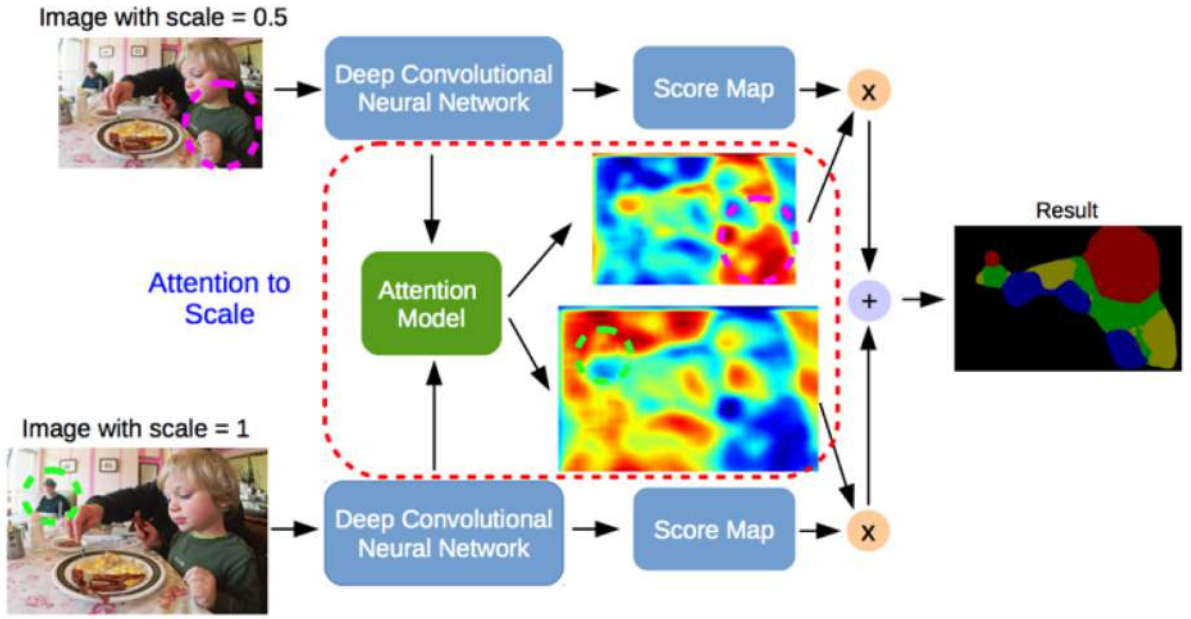
Model Architecture

- Allowing networks share parameters at different scales.



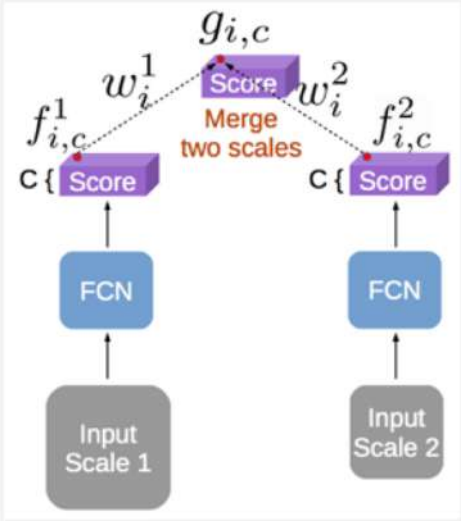
Model Architecture

- Allowing networks share parameters at different scales.



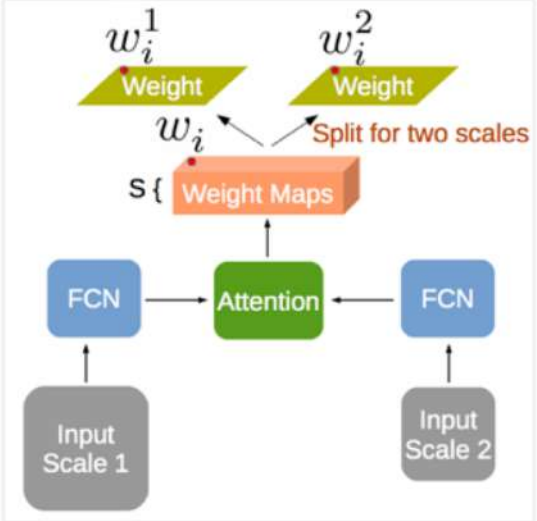
Attention Details

- The attention model takes as input the convolutional fc7 features from VGG-16, and pass through two layers (512 3 × 3 filters + 1 × 1 S filters).



$$g_{i,c} = \sum_{s=1}^{S} w_i^s \cdot f_{i,c}^s$$

$$w_i^s = \frac{\exp(h_i^s)}{\sum_{t=1}^{S} \exp(h_i^t)}$$

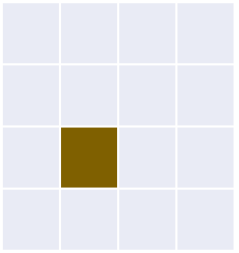


Attention Pooling – A Generalization

- Previous works usually apply average-pooling or max-pooling over multi-scale features.
- Attention pooling is a generalization over ave-pooling and max-pooling.

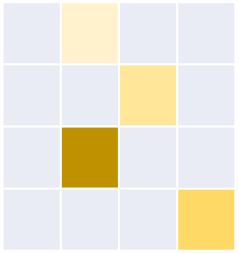
Ave-pooling

$$\alpha_i = \frac{1}{N} \forall i \in [1, N]$$



Max-pooling

$$\alpha_i = 1 \; \exists \; i$$
$$\alpha_j = 0 \; \forall \; j \neq i$$



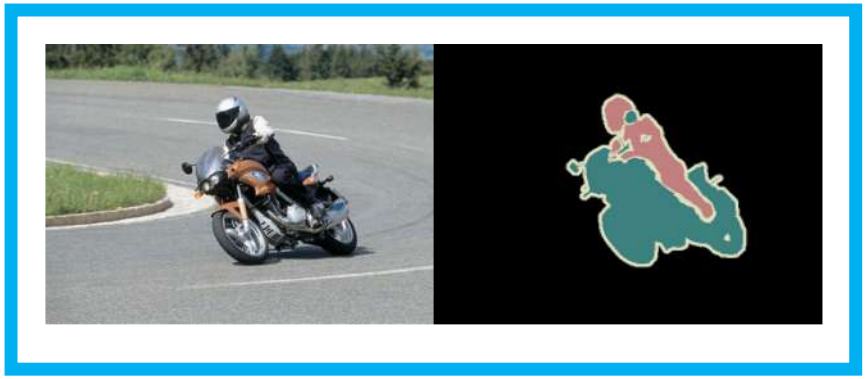
Attn-pooling

$$\sum_{i=1}^{N} \alpha_i = 1, \alpha_i \geq 0$$

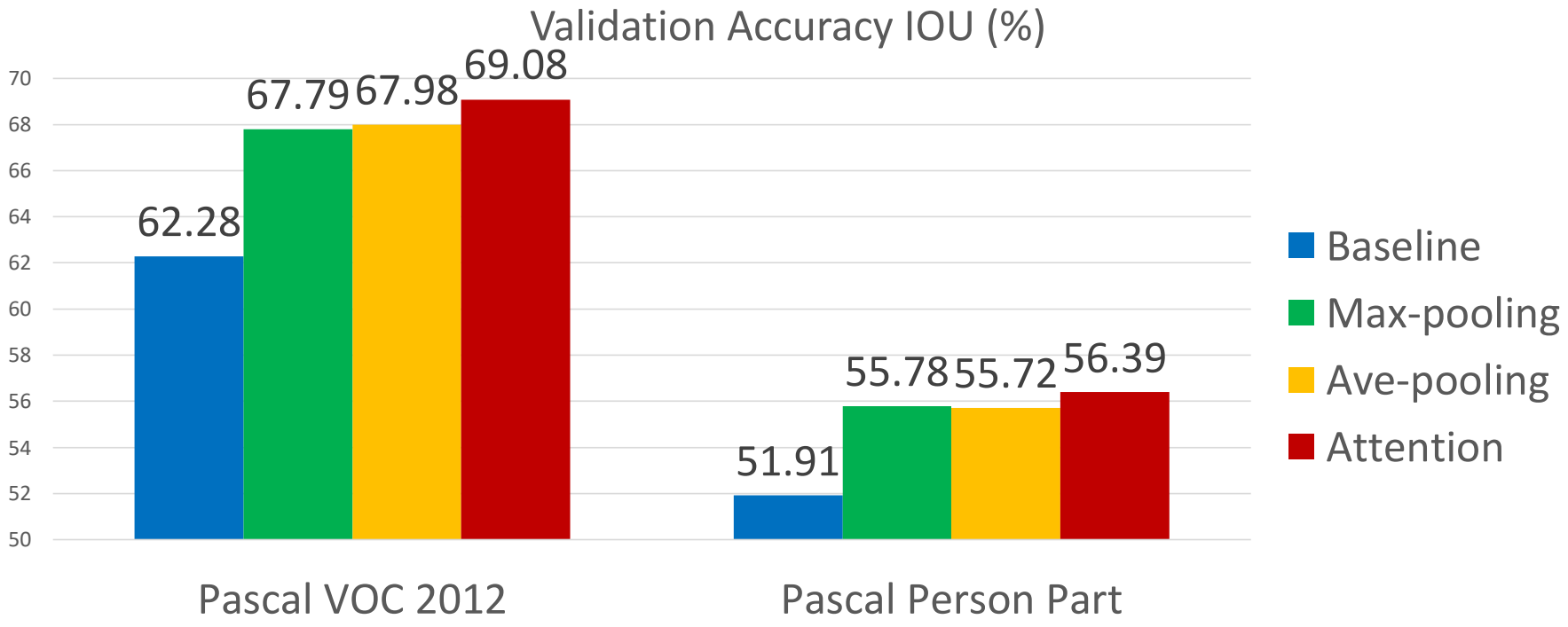
Experiments: Benchmark Datasets

- Dataset Statistics

Dataset	#Classes	#Train	#Validation
Pascal VOC 2012	20 + 1	1464	1449
Microsoft COCO	80 + 1	10000	1500
Pascal Human Part	6 + 1	1716	1817

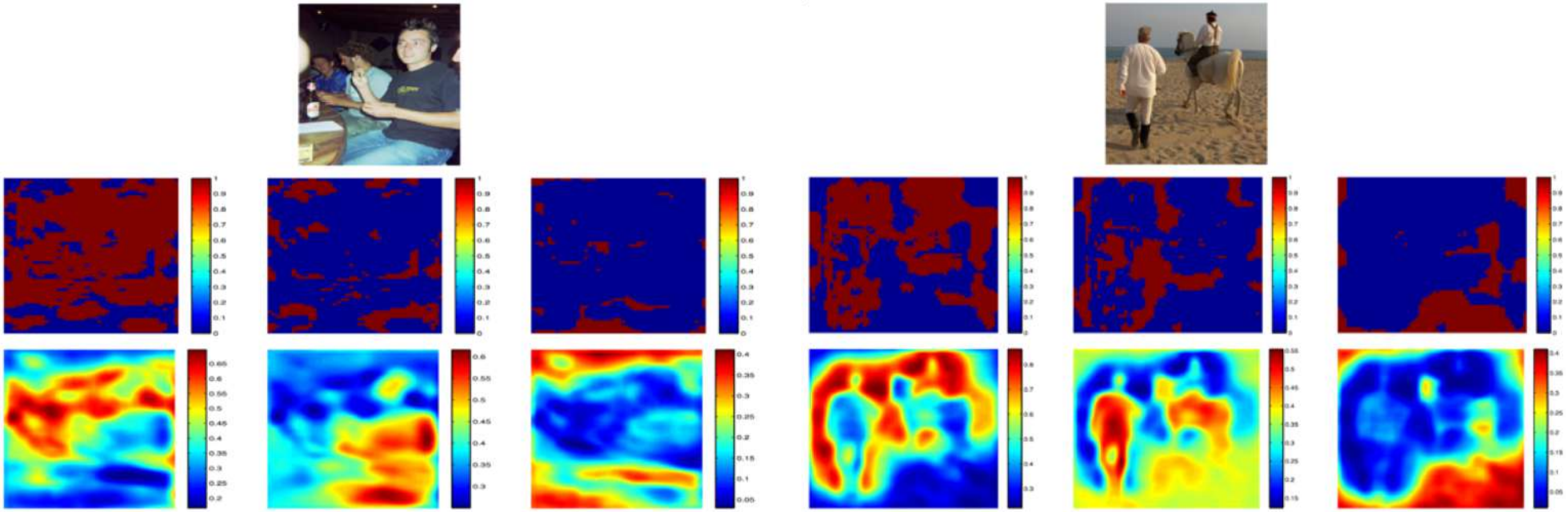


Attention v.s. Avg-Pool and Max-Pool



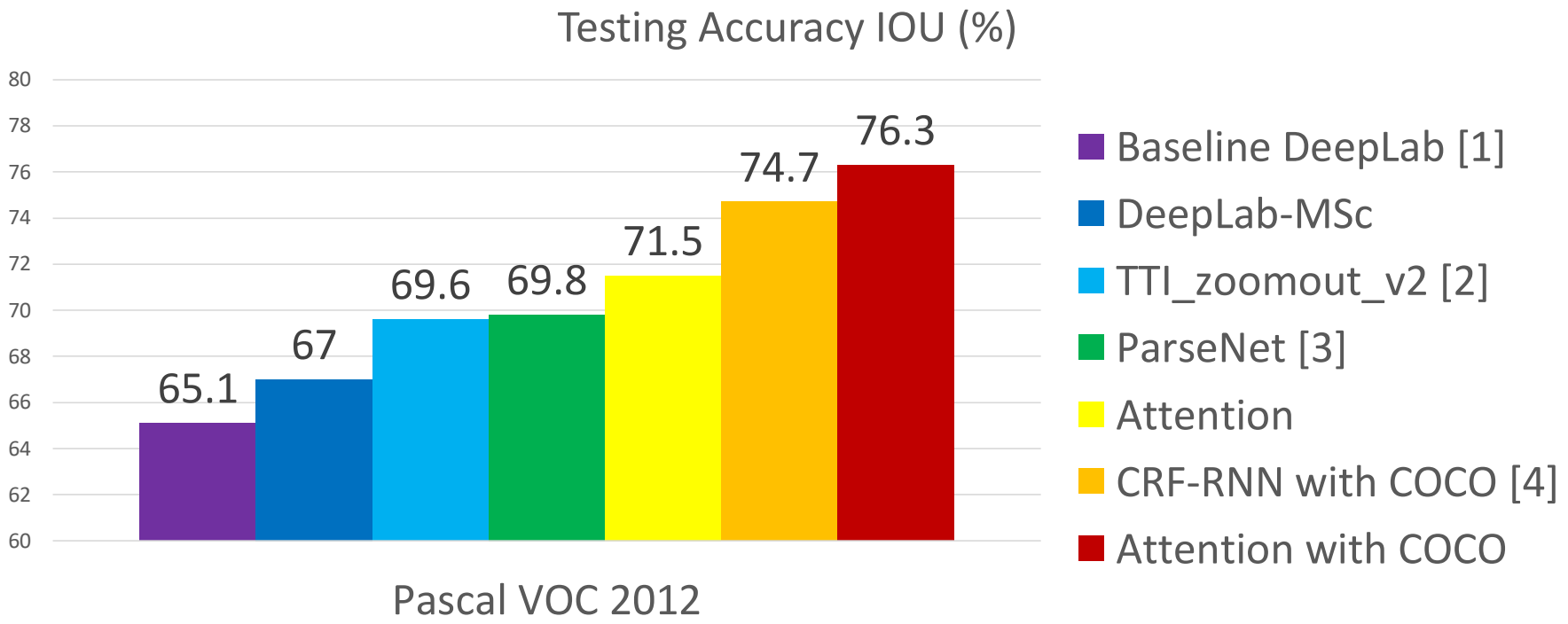
Visualization of Feature Importance

- Compared to max-pooling, the attention maps look more meaningful.



(a) Scale-1 Attention (b) Scale-0.75 Attention (c) Scale-0.5 Attention (a) Scale-1 Attention (b) Scale-0.75 Attention (c) Scale-0.5 Attention

Quantitative Results



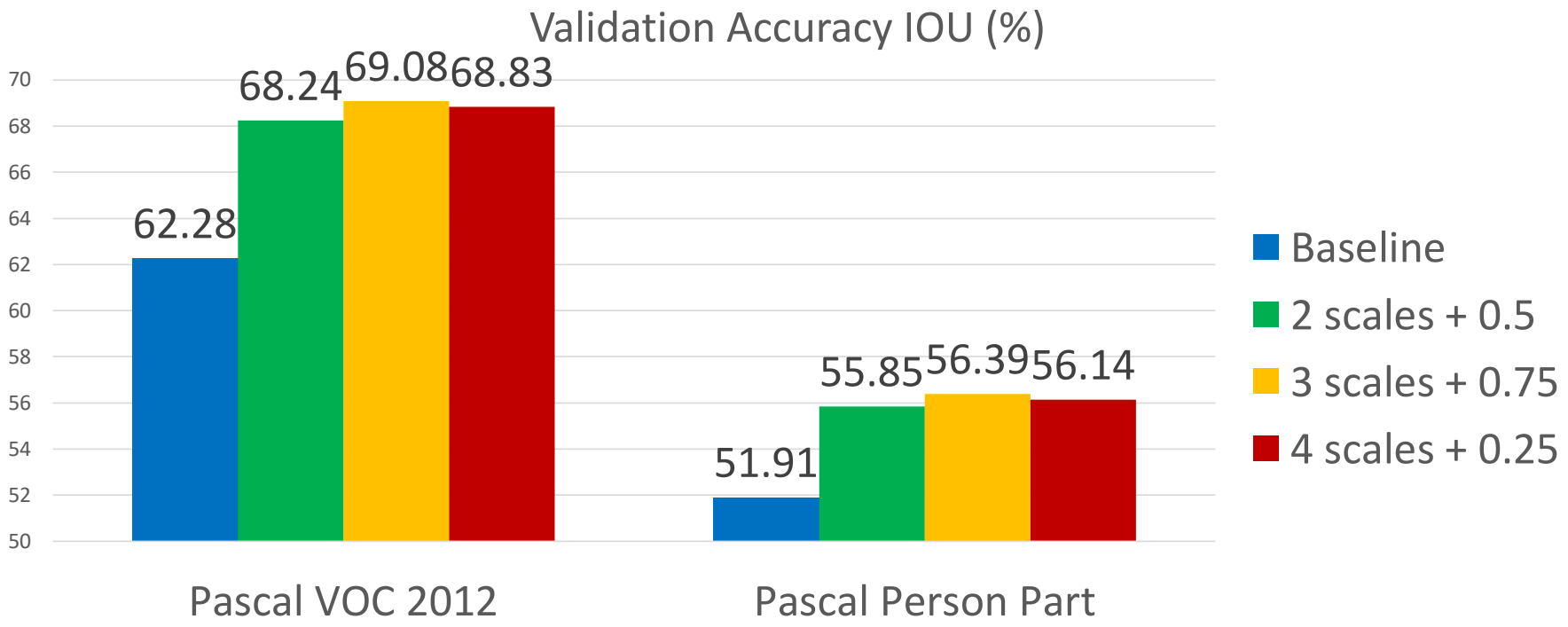
[1] Chen et al, DeepLab: Semantic Image Segmentation with Atrous Convolution, PAMI 2017

[2] Mostajabi et al, Feed- forward semantic segmentation with zoom-out features , CVPR 2015

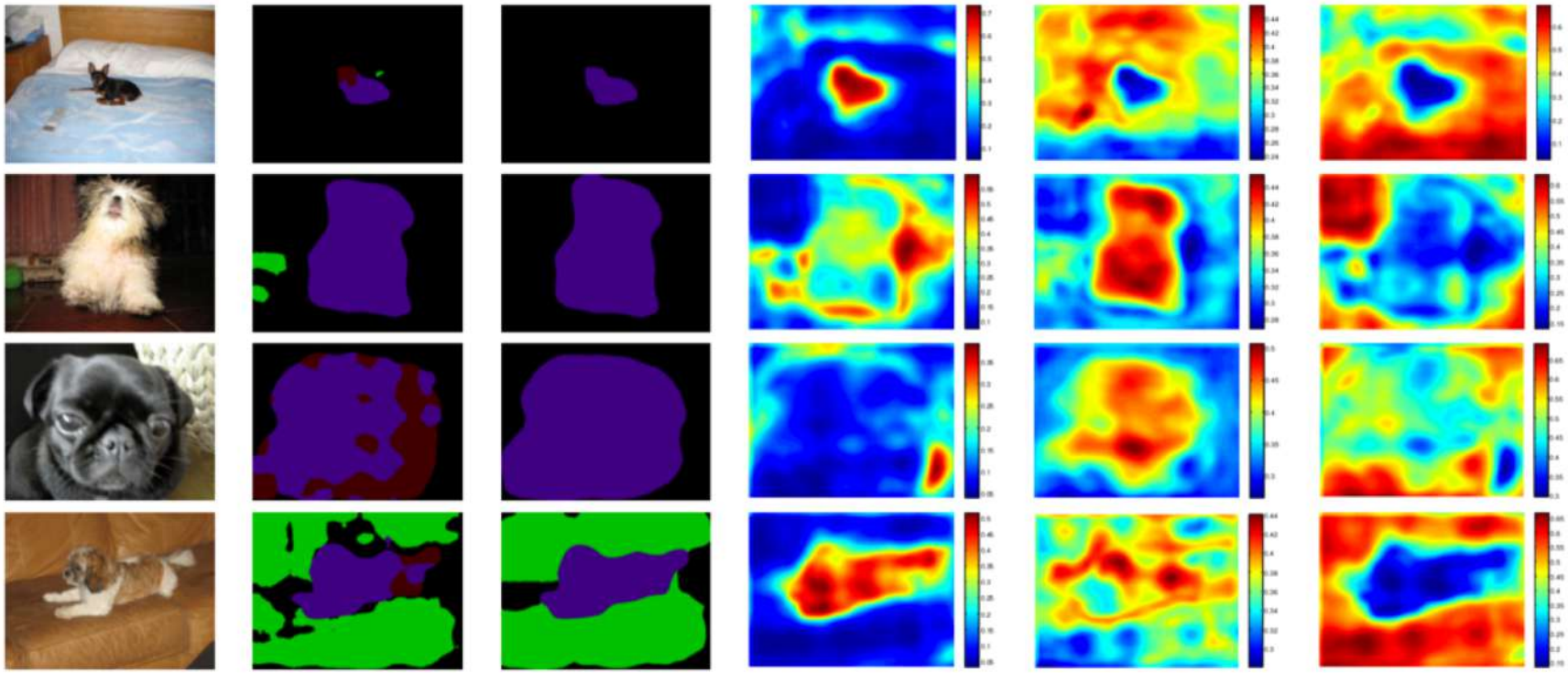
[3] Liu et al, Parsenet: Looking wider to see better., ICLR 2016W

[4] Zheng et al, Conditional random fields as recurrent neural networks, ICCV 2015

Ablation Study – Number of Scales



Qualitative Results



(a) Image (b) Baseline (c) Our model (d) Scale-1 Attention (e) Scale-0.75 Attention (f) Scale-0.5 Attention

Cherry-Pick Examples