DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs

Valentin Biryukov, 152
https://arxiv.org/pdf/1606.00915.pdf

Disclaimer

Yes, authors really use all of things from header.

It will be difficult (I myself am not sure what I understand)

Guide to presentation

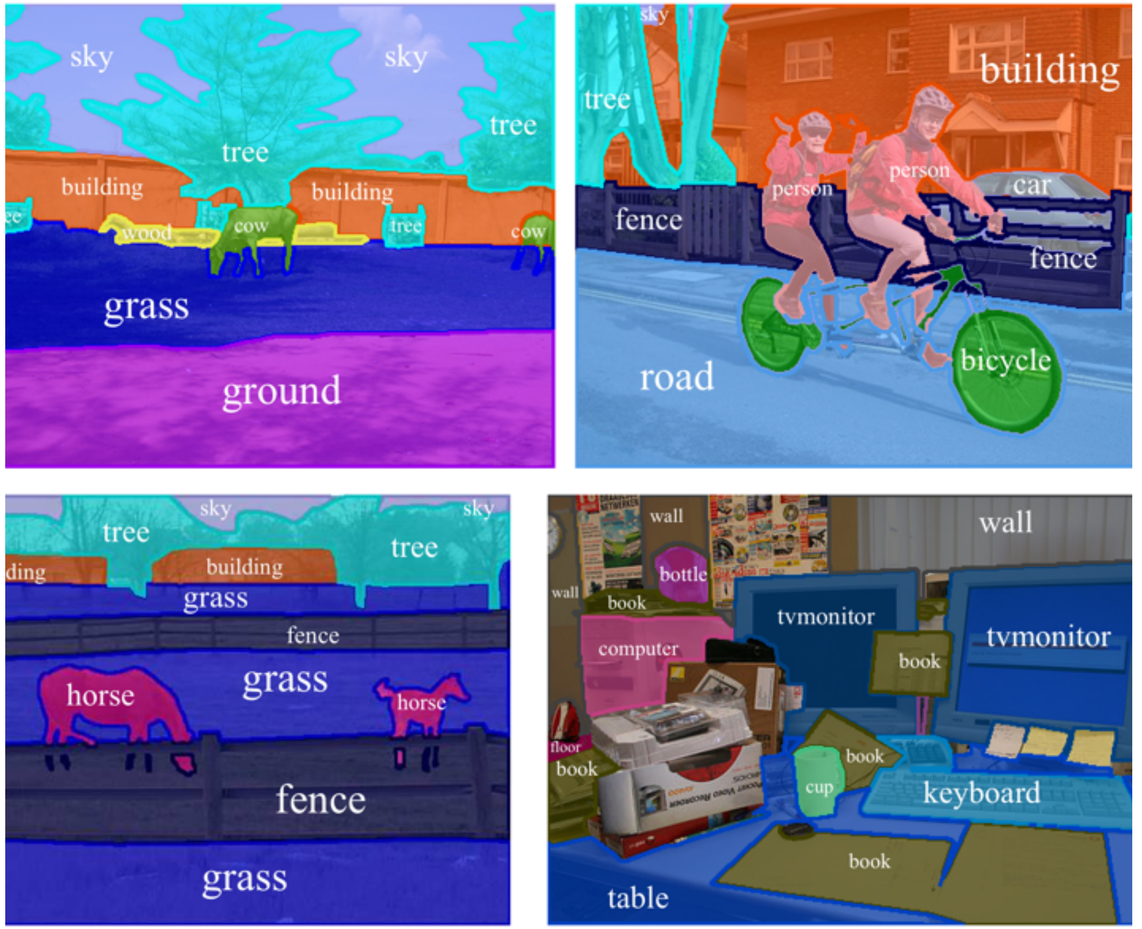
1) Task
2) Terms
3) Related works
4) Model OverView
5) Methods
5.1) AC for better feature extraction
5.2) Muliscale Image Reps via ASPP
5.3) Structured Prediction via fcCRF
6) Experimental results
7) ...
8) PROFIT!

Semantic image segmentation

«Segmentation" is a partition of an image into several "coherent" parts, but without any attempt at understanding what these parts represent. One of the most famous works (but definitely not the first) is Shi and Malik "Normalized Cuts and Image Segmentation" PAMI 2000. These works attempt to define "coherence" in terms of low-level cues such as color, texture and smoothness of boundary. You can trace back these works to the Gestalt theory.

On the other hand "semantic segmentation" attempts to partition the image into semantically meaningful parts, and to classify each part into one of the pre-determined classes. You can also achieve the same goal by classifying each pixel (rather than the entire image/segment). In that case you are doing pixel-wise classification, which leads to the same end result but in a slightly

- Shai, SOF



https://stackoverflow.com/questions/33947823/what-is-semantic-segmentation-compared-to-segmentation-and-scene-labeling

Tasks

(1) reduced feature resolution

(2) existence of objects at multiple scales

(3) reduced localization accuracy due to DCNN invariance

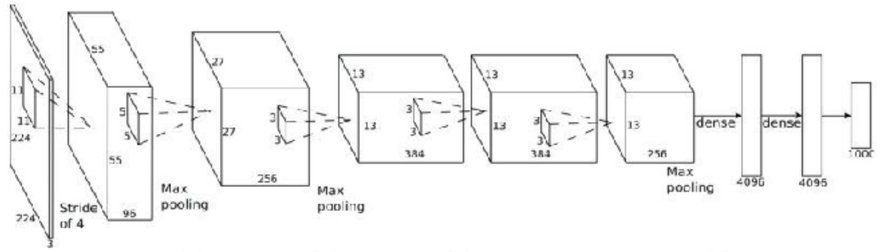
Terms

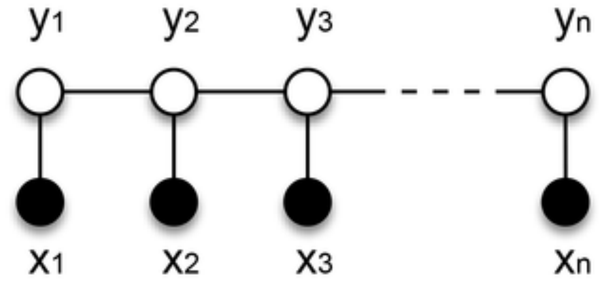
DCNN - Deep Convolutional Neural Networks

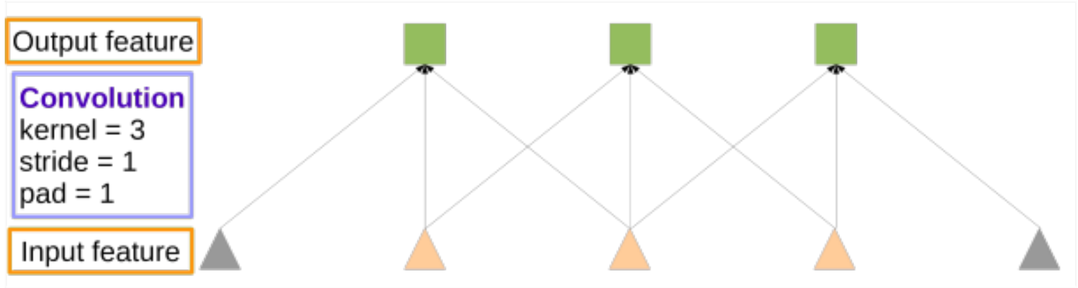
CRF - Conditional Random Fields

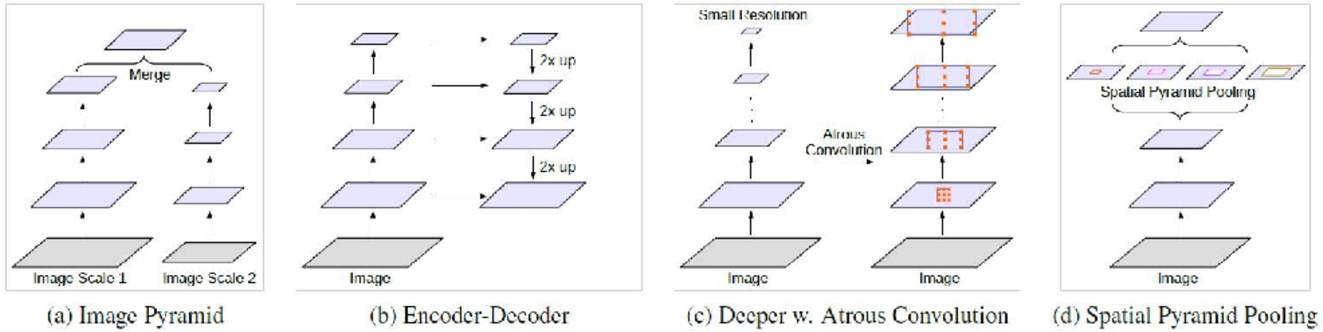
Atrous Convolution

ASPP - Atrous spatial pyramid pooling









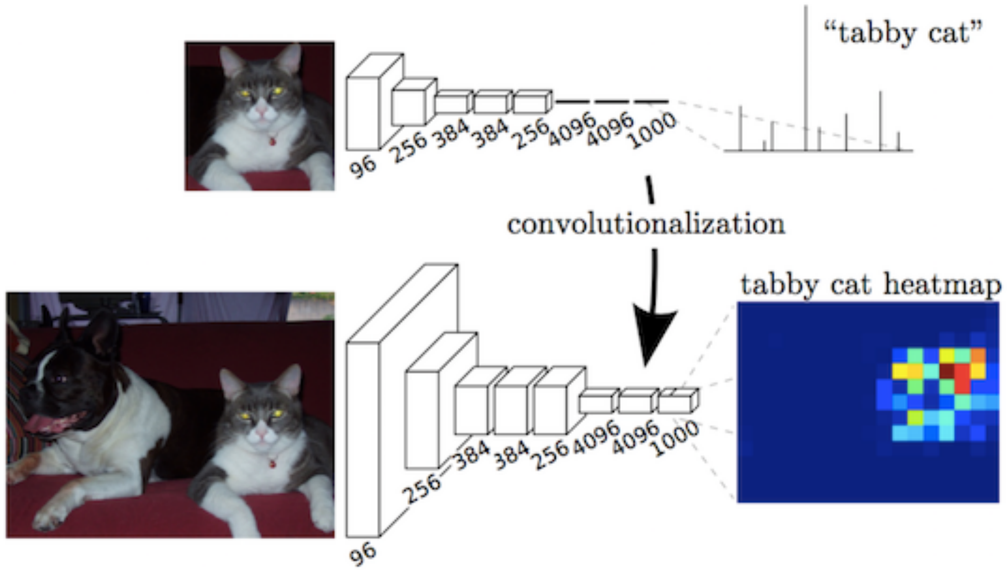
(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

Related

Fully Convolutional Networks for Semantic Segmentation

Submitted on 14 Nov 2014

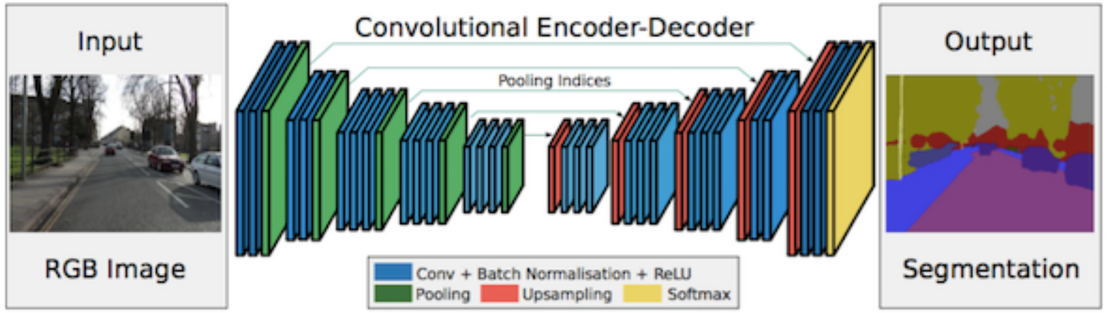
Benchmarks (VOC2012): 67.2 (62.2)



Related

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

Submitted on 2 Nov 2015



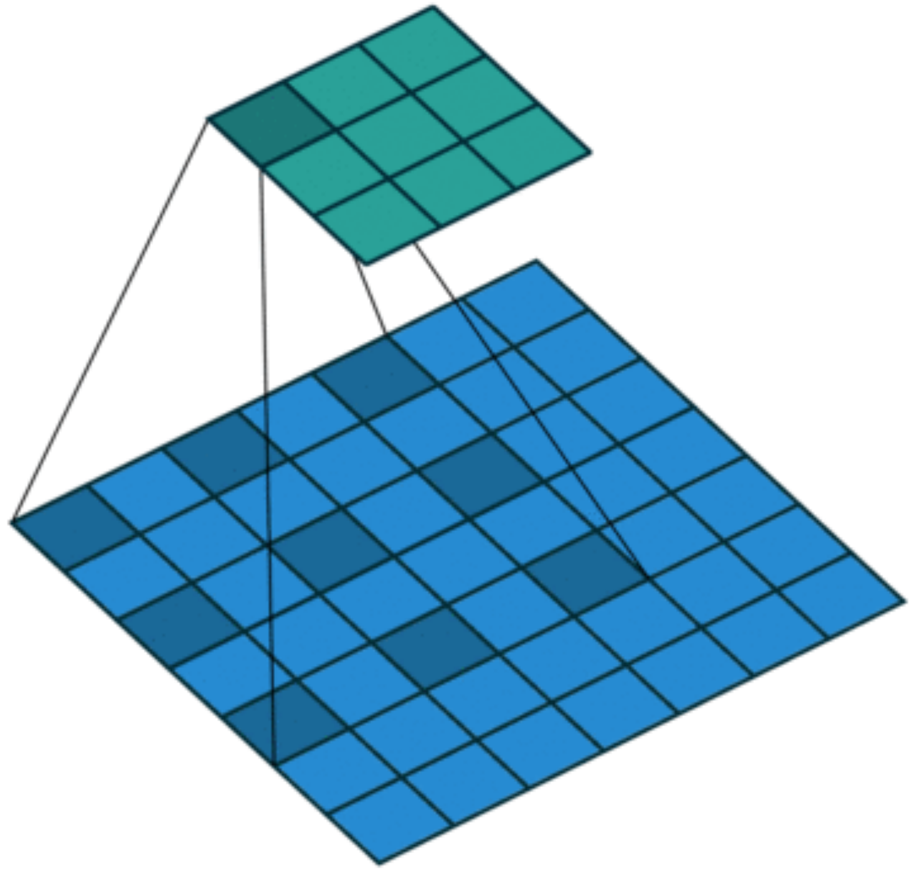
Benchmarks (VOC2012): 59.9

Related

Multi-Scale Context Aggregation by Dilated Convolutions

Submitted on 23 Nov 2015

Benchmarks (VOC2012): 75.3 (71.3)



This Work

v1 : Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs

Submitted on 22 Dec 2014

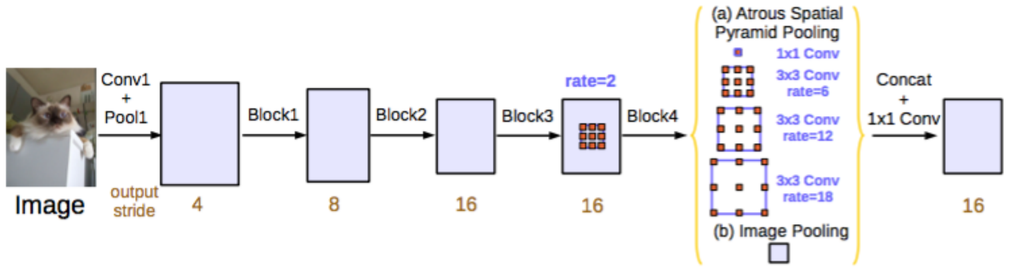
v2 : DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs

Submitted on 2 Jun 2016

v2 : Rethinking Atrous Convolution for Semantic Image Segmentation

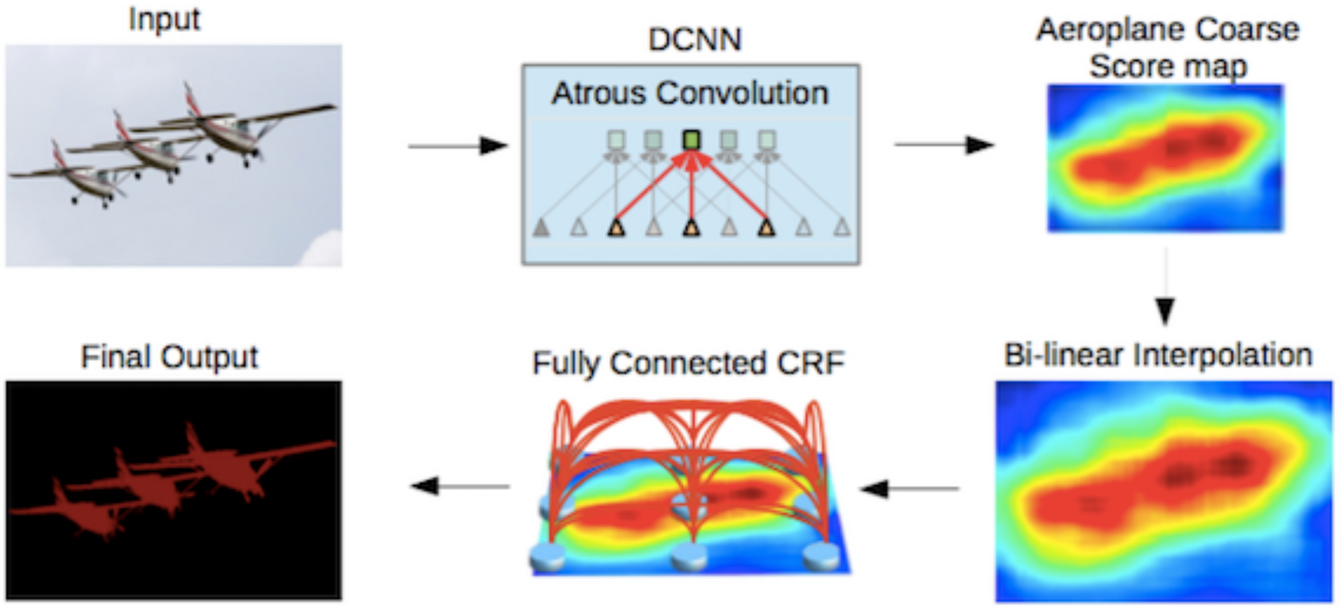
Submitted on 17 Jun 2017

Benchmarks (VOC2012): 85.7 (79.7 - v2)



Model OverView

A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries



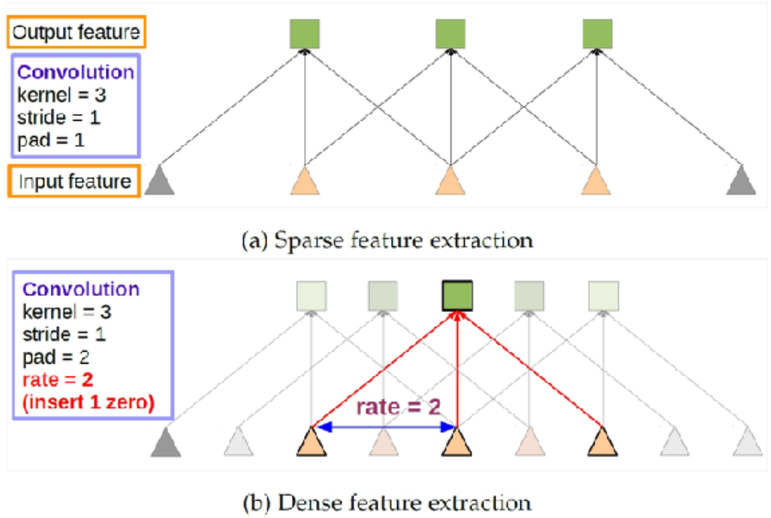
Atrous Convolution for Dense Feature Extraction and Field-of-View Enlargement

The output y[i] of atrous convolution:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k].$$

input signal x[i]

filter w[k] of length K



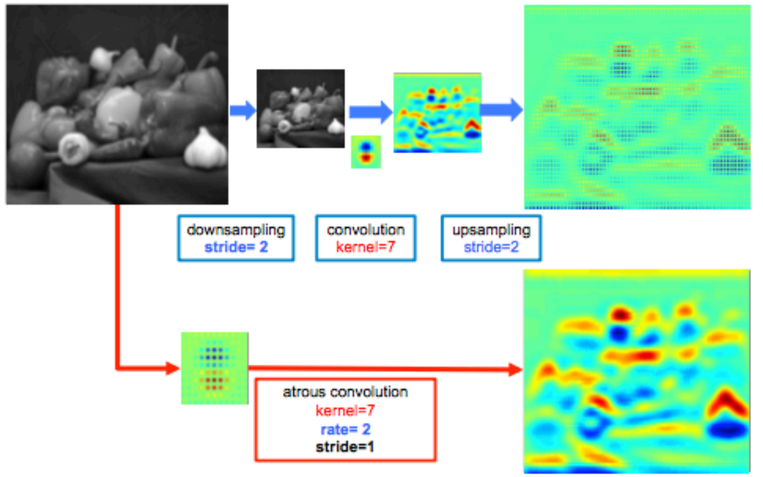
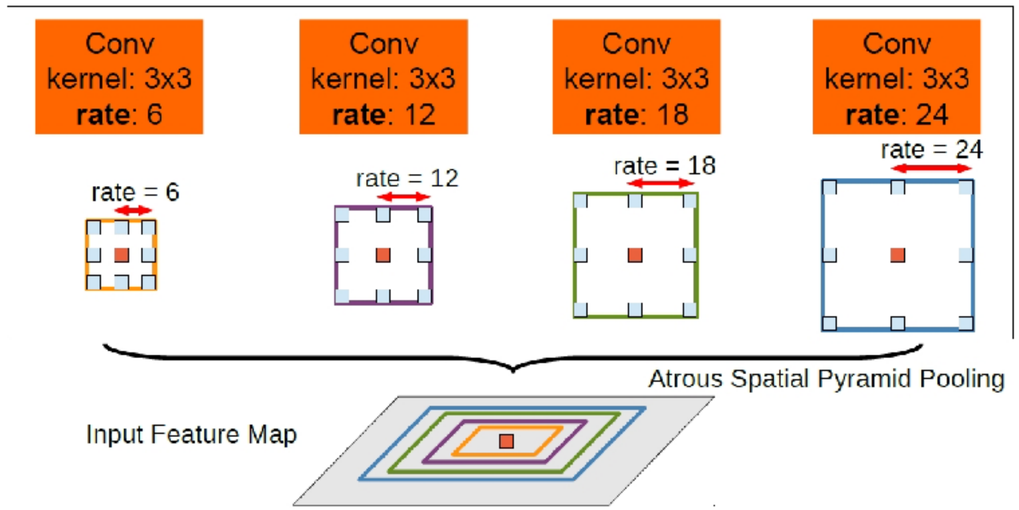


Fig. 3: Illustration of atrous convolution in 2-D. Top row: sparse feature extraction with standard convolution on a low resolution input feature map. Bottom row: Dense feature extraction with atrous convolution with rate $r = 2$, applied on a high resolution input feature map.

Multiscale Image Representations using Atrous Spatial Pyramid Pooling

Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.



Structured Prediction with Fully-Connected Conditional Random Fields for Accurate Boundary Recovery

X is the label assignment for pixels

potential $\theta_i(x_j) = -\log P(x_j)$ (label assignment probability at pixel i as computed by a DCNN)

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j)$$

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \left[w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2} \right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2} \right) \right] \quad (3)$$

b - Gaussian convolutions in bilateral space

σ - control the scale of Gaussian kernels

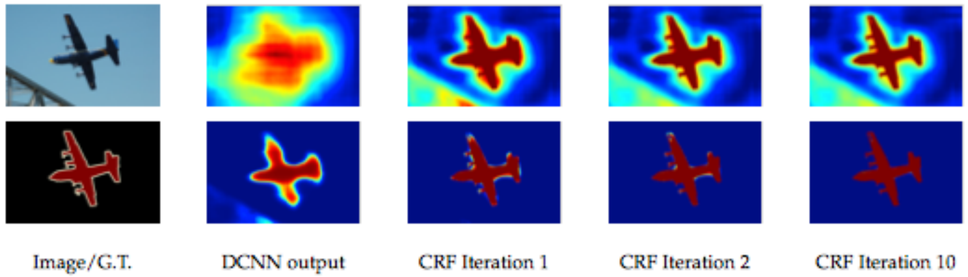


Fig. 5: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference.

Experimental 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

TABLE 1: Effect of Field-Of-View by adjusting the kernel size and atrous sampling rate r at 'fc6' layer. We show number of model parameters, training speed (img/sec), and *val* set mean IOU before and after CRF. DeepLab-LargeFOV (kernel size 3×3, $r = 12$) strikes the best balance.

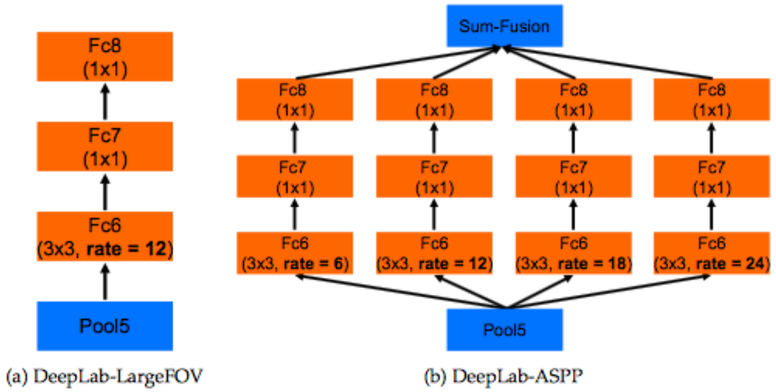
Learning policy	Batch size	Iteration	mean IOU
step	30	6K	62.25
poly	30	6K	63.42
poly	30	10K	64.90
poly	10	10K	64.71
poly	10	20K	65.88

TABLE 2: PASCAL VOC 2012 *val* set results (%) (before CRF) as different learning hyper parameters vary. Employing "poly" learning policy is more effective than "step" when training DeepLab-LargeFOV.

Fig. 6: PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.

http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

Experimental results. We Can - We do

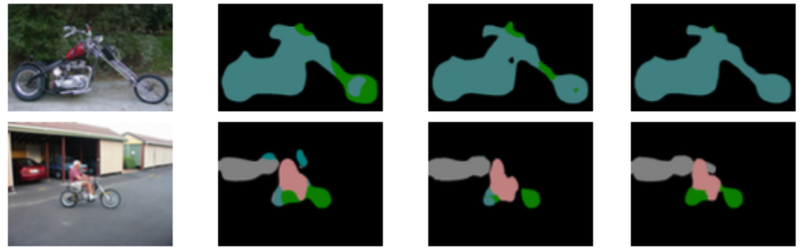


(a) DeepLab-LargeFOV (b) DeepLab-ASPP

Fig. 7: DeepLab-ASPP employs multiple filters with different rates to capture objects and context at multiple scales.

Method	before CRF	after CRF
LargeFOV	65.76	69.84
ASPP-S	66.98	69.73
ASPP-L	68.96	71.57

TABLE 3: Effect of ASPP on PASCAL VOC 2012 *val* set performance (mean IOU) for VGG-16 based DeepLab model. **LargeFOV**: single branch, $r = 12$. **ASPP-S**: four branches, $r = \{2, 4, 8, 12\}$. **ASPP-L**: four branches, $r = \{6, 12, 18, 24\}$.



(a) Image (b) LargeFOV (c) ASPP-S (d) ASPP-L

Fig. 8: Qualitative segmentation results with ASPP compared to the baseline LargeFOV model. The **ASPP-L** model, employing multiple *large* FOVs can successfully capture objects as well as image context at multiple scales.

http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

Experimental results. Gonna be on top

Method	mIOU
DeepLab-CRF-LargeFOV-COCO [58]	72.7
MERL_DEEP_GCRF [88]	73.2
CRF-RNN [59]	74.7
POSTECH_DeconvNet_CRF_VOC [61]	74.8
BoxSup [60]	75.2
Context + CRF-RNN [76]	75.3
QO_4^{mres} [66]	75.5
DeepLab-CRF-Attention [17]	75.7
CentraleSuperBoundaries++ [18]	76.0
DeepLab-CRF-Attention-DT [63]	76.3
H-ReNet + DenseCRF [89]	76.8
LRR_4x_COCO [90]	76.8
DPN [62]	77.5
Adelaide_Context [40]	77.8
Oxford_TVG_HO_CRF [91]	77.9
Context CRF + Guidance CRF [92]	78.1
Adelaide_VeryDeep_FCN_VOC [93]	79.1
DeepLab-CRF (ResNet-101)	79.7

TABLE 5: Performance on PASCAL VOC 2012 *test* set. We have added some results from recent arXiv papers on top of the official leadearboard results.

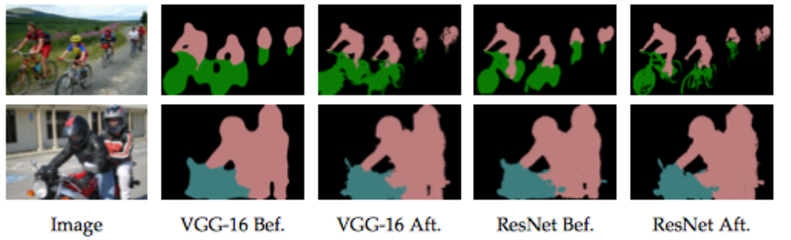


Fig. 9: DeepLab results based on VGG-16 net or ResNet-101 before and after CRF. The CRF is critical for accurate prediction along object boundaries with VGG-16, whereas ResNet-101 has acceptable performance even before CRF.

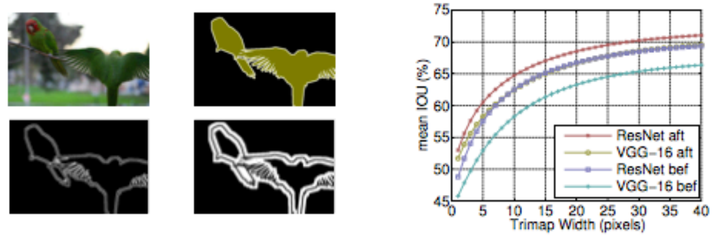


Fig. 10: (a) Trimap examples (top-left: image. top-right: ground-truth. bottom-left: trimap of 2 pixels. bottom-right: trimap of 10 pixels). (b) Pixel mean IOU as a function of the band width around the object boundaries when employing VGG-16 or ResNet-101 before and after CRF.

http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

Some pics

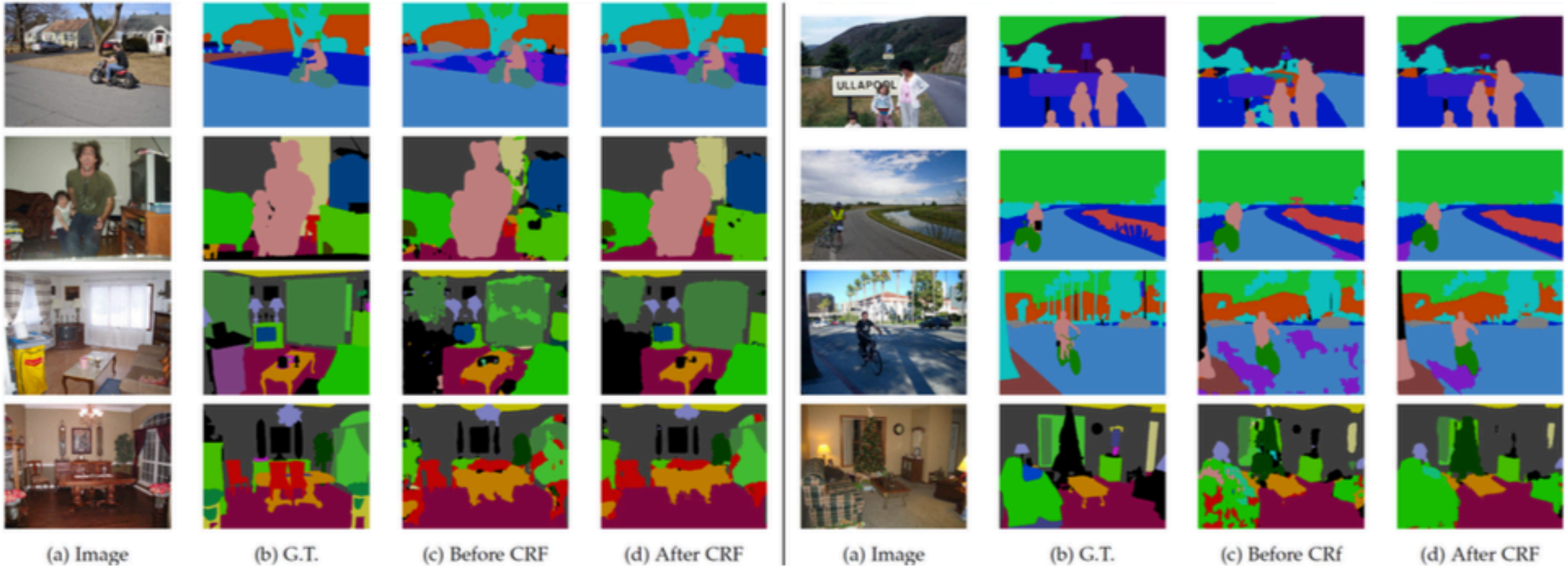
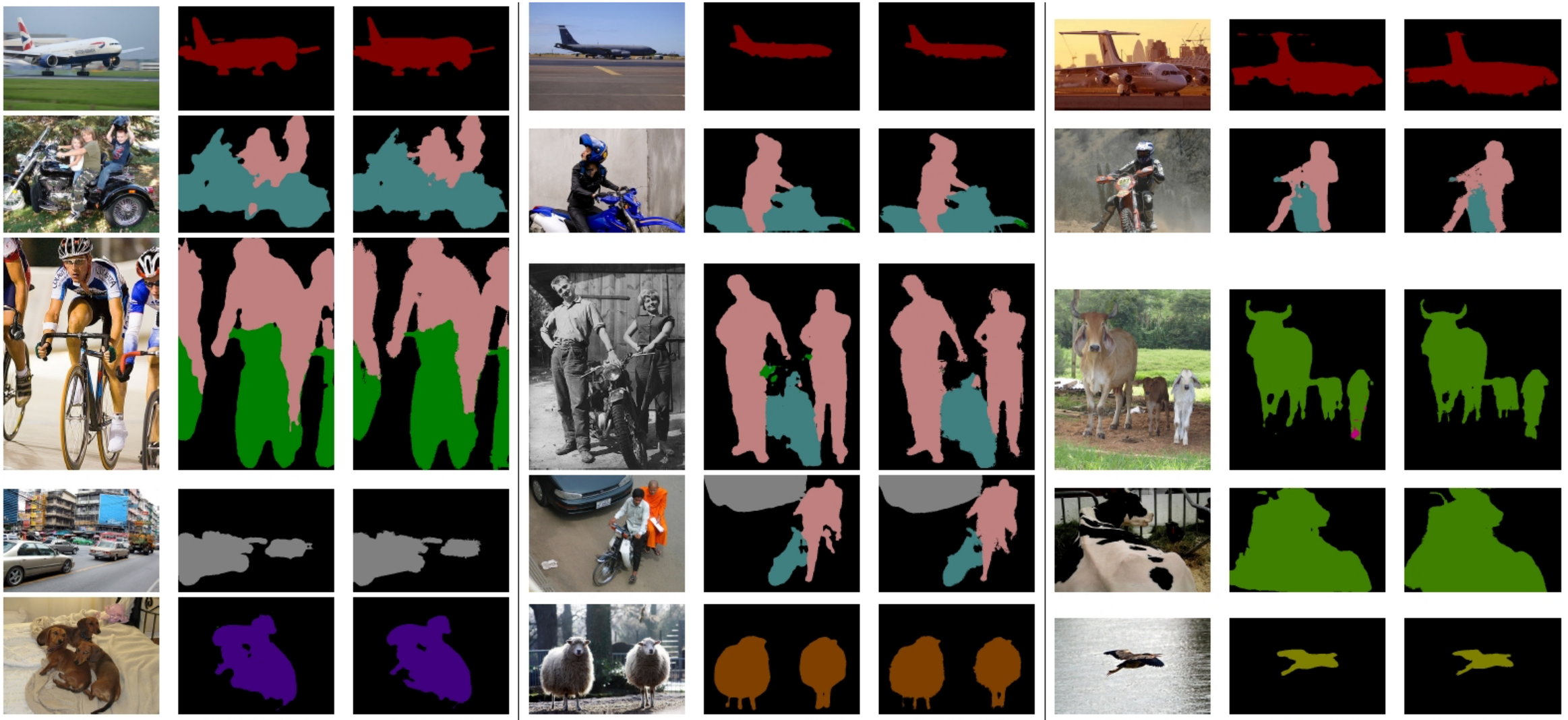


Fig. 11: PASCAL-Context results. Input image, ground-truth, and our DeepLab results before/after CRF.

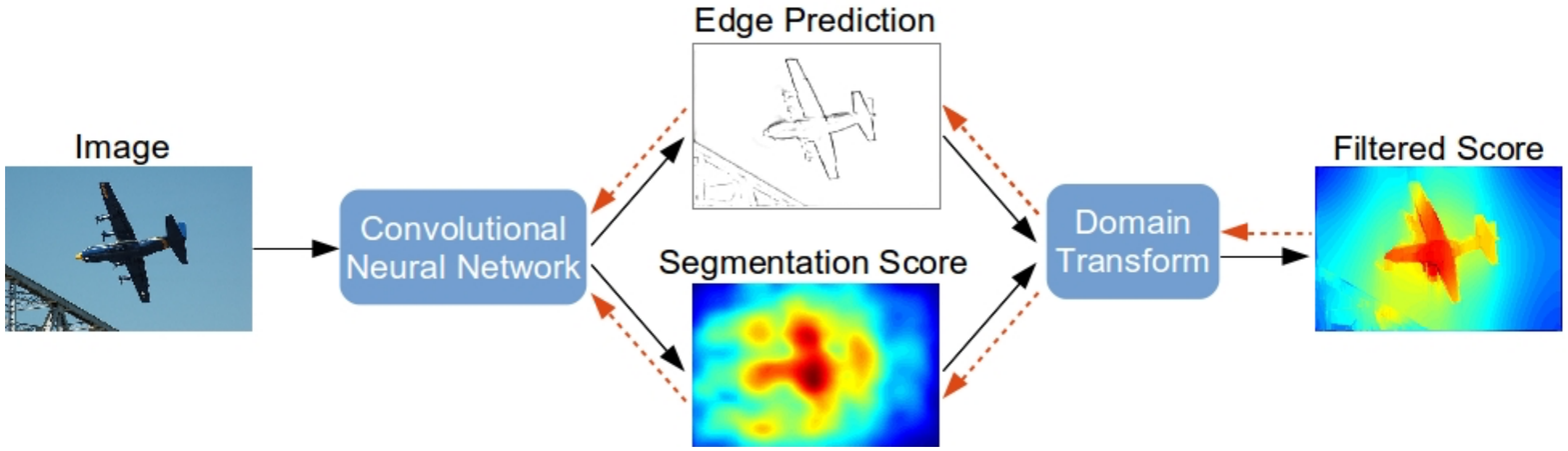
Some pics



(a) Image (b) Before CRF (c) After CRF (a) Image (b) Before CRF (c) After CRF (a) Image (b) Before CRF (c) After CRF

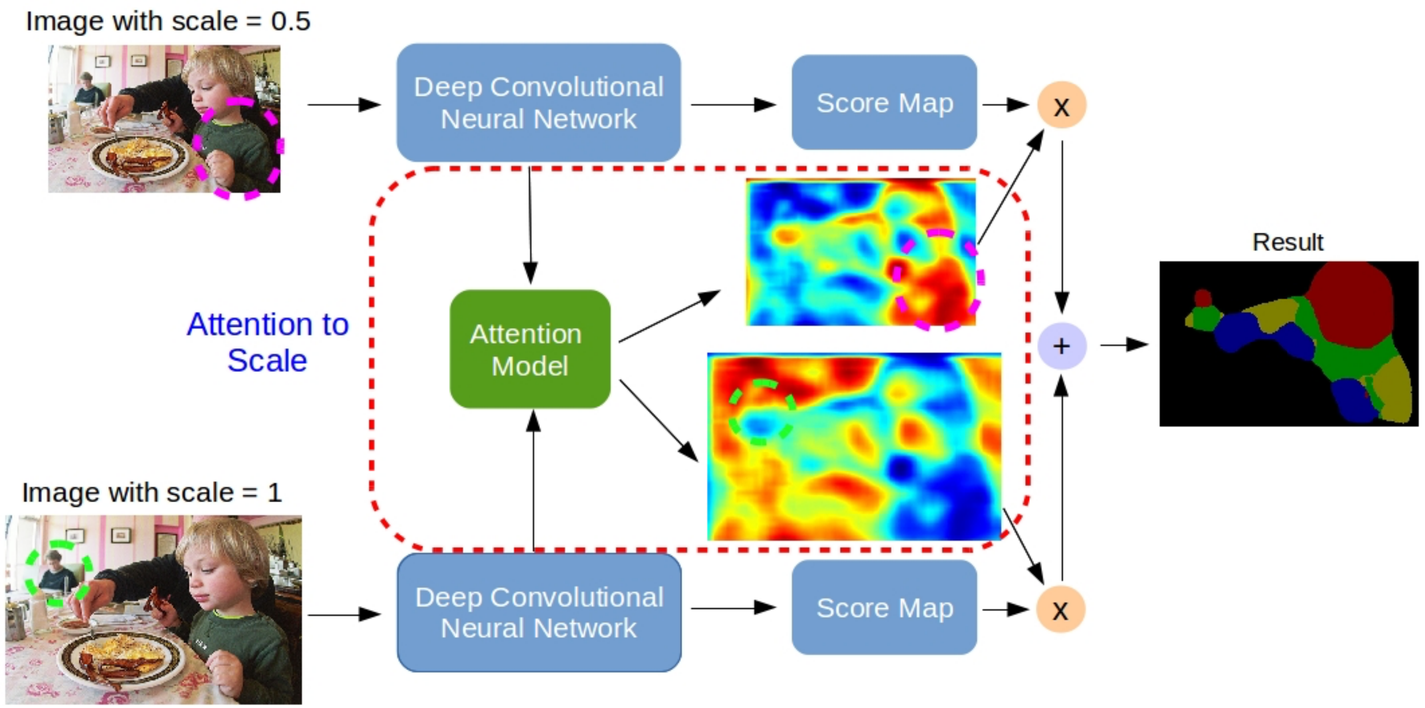
Fans. Projects

Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform



Fans. Projects

Attention to Scale: Scale-aware Semantic Image Segmentation



Use machine learning carefully =)