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

Zhipeng Yan, Moyuan Huang, Hao Jiang

5/1/2017

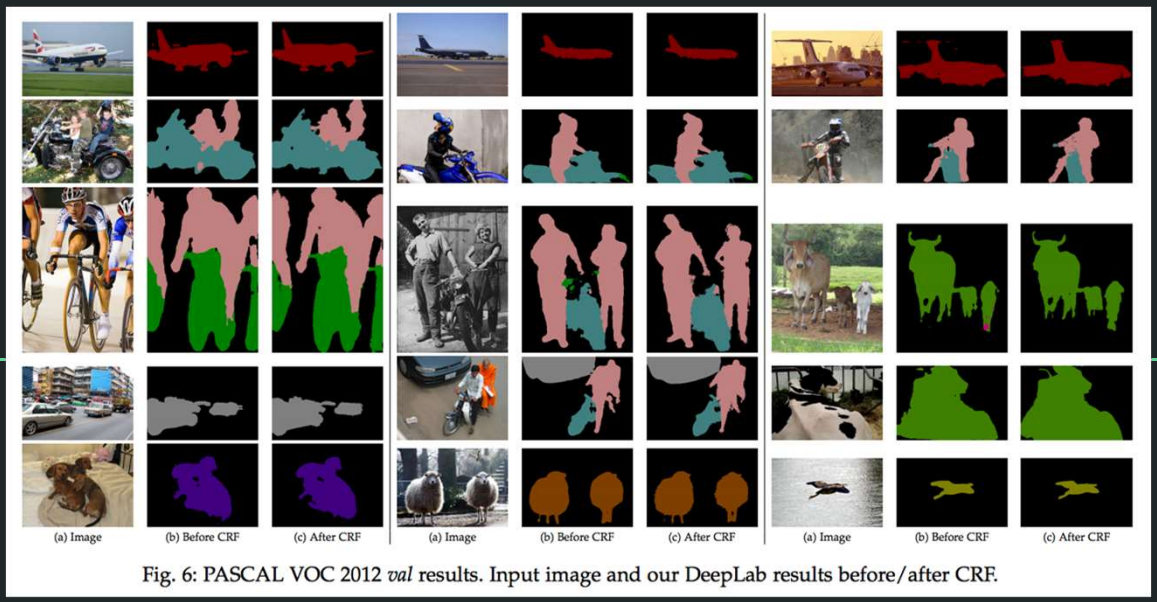


Fig. 6: PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.

Outline

- Background
  - semantic segmentation
    - Objective, Dataset, Basic Idea
  - Predecessor: Fully Convolutional network
    - Fully Convolution, Deconvolution, Skip path
- DeepLab
  - Atrous Convolution
  - ASPP
  - Fully Connected CRFs

Background: semantic segmentation

Semantic Segmentation

One of the most popular topics in Computer Vision

- Like image classification, object detection…

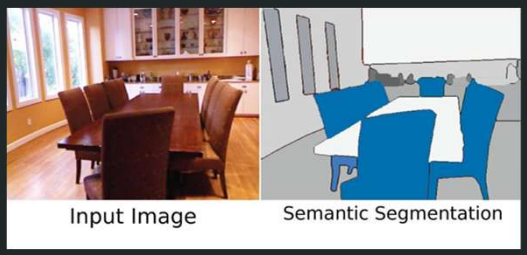
The objective

- Partition the image into segments according to different semantic groups

Semantic Segmentation

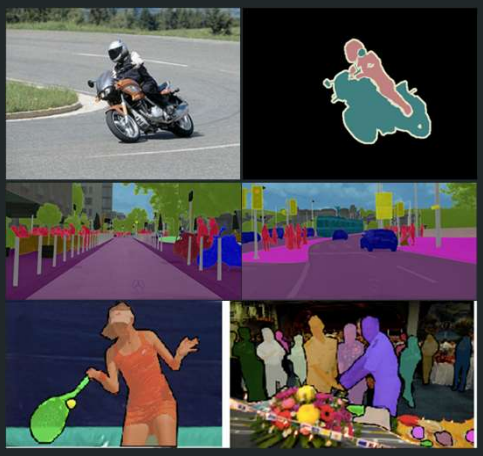
Example

- Split the image according to object boundaries
- Colors each object according to its semantic meanings

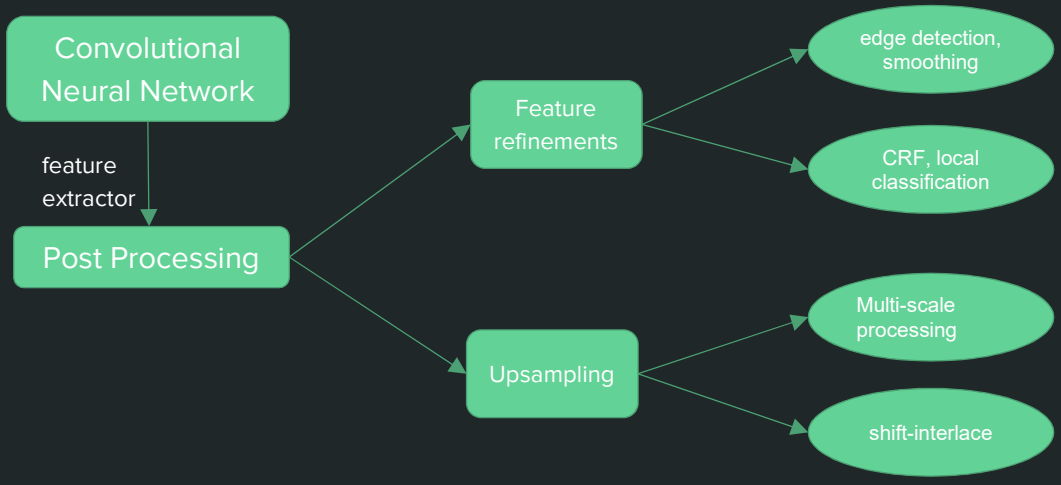


Semantic Segmentation - Dataset

- PASCAL VOC
- Cityscapes
- Microsoft COCO



Semantic Segmentation - Idea

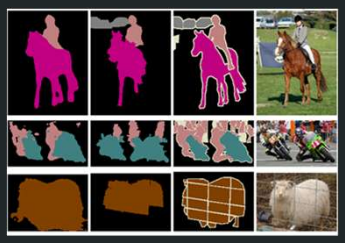


Semantic Segmentation - Idea

Problems:

- Not end-to-end
- Fixed size input & output
- Small/restricted Field of View
- Low performance
- Time consuming

Solved at once by **Fully Convolutional Network (J. Long, E. Shelhamer, T. Darrell**



Fully Convolutional Networks

Jonathan Long et al.
CVPR 2015 & PAMI 2016

Previous CNN

Previous CNN downsized the output layer-by-layer in order to reduce parameter.

Very harmful for dense classification tasks like image segmentation: Lot of information

Can design a network specifically for the Segmentation task?

Fully Convolutional Networks

Input size can be arbitrary.
- Remove fully connected layer
- Practical use

30% relative improvement on previous state-of-the-art
- Avoid fully connected layers in traditional CNNs which cause the loss of spatial information

5x efficiency
- Reduced parameters
- End-to-end training

Fully Convolutional Networks

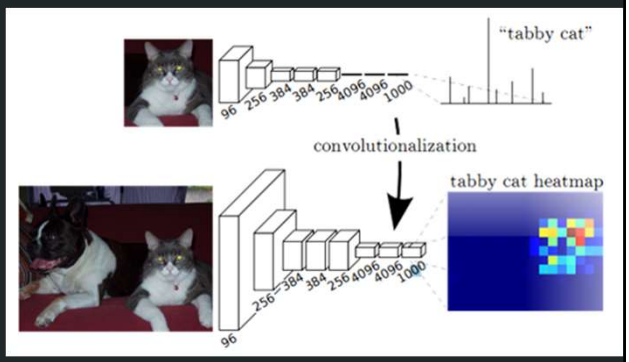
3 Major innovations on network architecture

- Removal of fully connected layers
- Deconvolution
- Skip path

Fully Convolutional Networks

Removal of fully connected layers

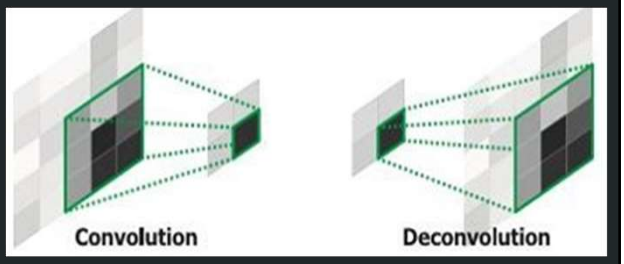
- Dense output with relative size to the input
- Replace with 1 x 1 convolutions to transform feature maps to class-wise predictions
- Respons map with “tabby cat” kernel



Fully Convolutional Networks

Deconvolution

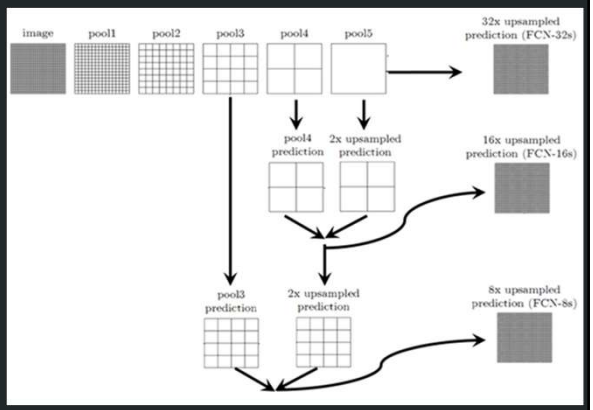
- Used for retrieving information, can be regarded as a reverse of convolution
- Stride size so as to avoid overlap: though NNs can adjust the corresponding weight to avoid them, it's really a struggle to avoid it completely.



Fully Convolutional Networks

Skip path

- Concatenate low level features with high level features to handle multiscale objects
- Provide options for different output sizes

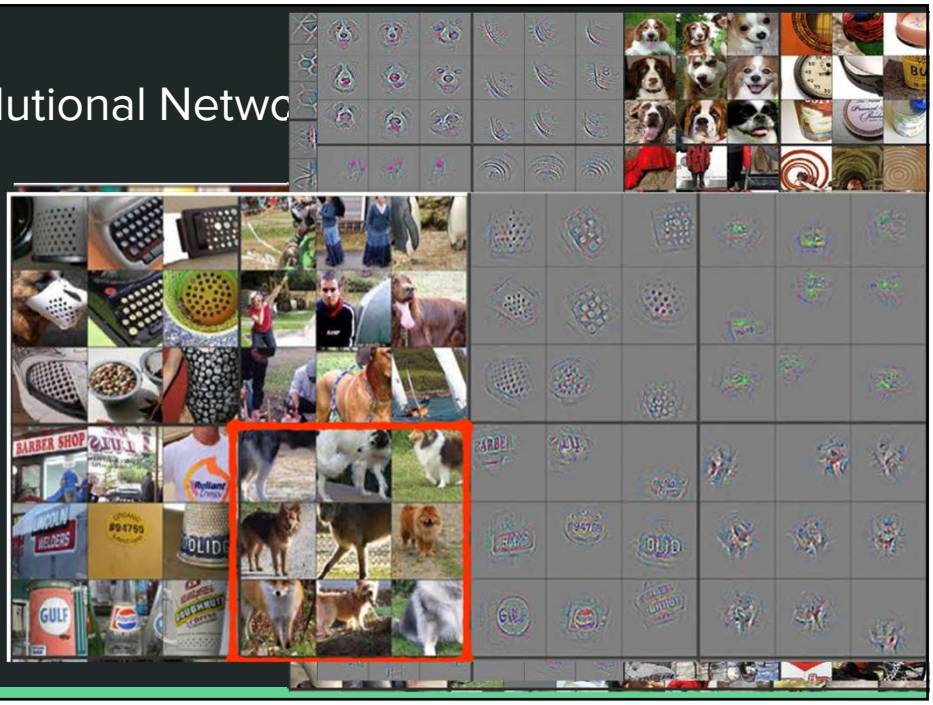


Skip layer visualization

Fully Convolutional Netwo

Layer 2

Layer 3

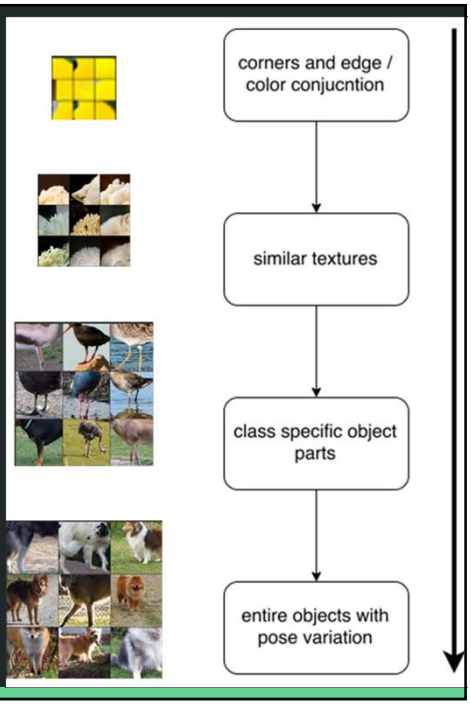


different layer provide us with different levels of information.

increasing locality of perceptual field

Different level of generality

detailed visualization can be found on https://arxiv.org/pdf/1311.2901.pdf.



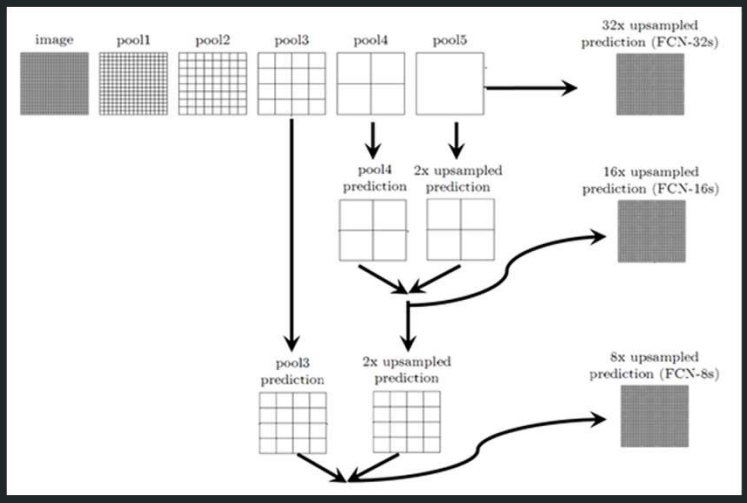
Skip Layer

The skip layer technique is widely used in many popular deep networks such as ResNet, Inside-Outside Net, HED. The advantage as well as motivation is that it allows more lower level information to reach top level.

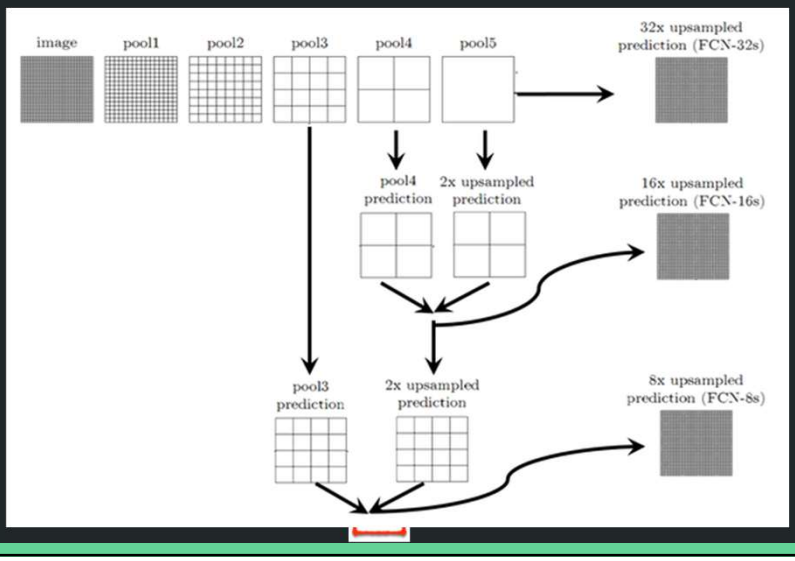
With skip layer, we can get a rather finer pixel output, instead of a small coarse one.

Upsampling is used to resolve the size incompatible problem between different layer. Combining is done by simple sum operation.

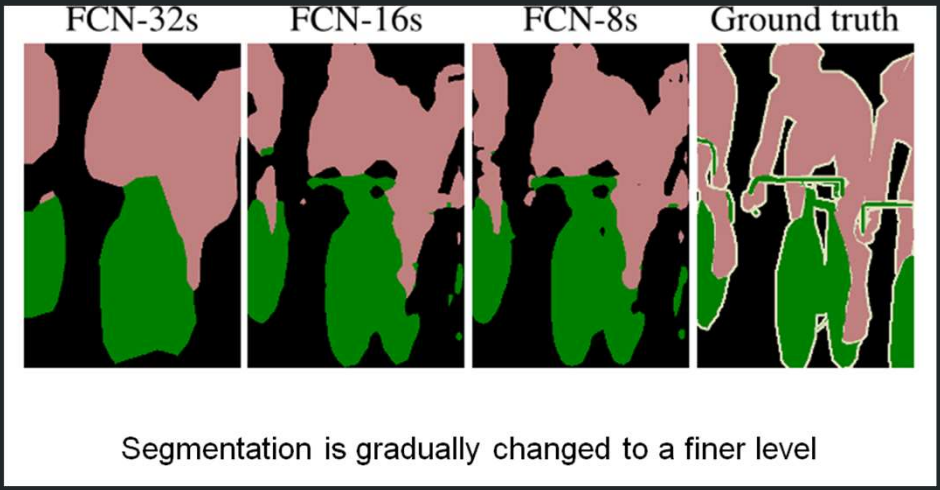
Skip Layer Model In detail



Skip Layer Model In detail



Skip layer on different scales



Segmentation is gradually changed to a finer level

Some Visualization

- FCN architecture visualization with Caffe
- A distill dynamic paper

Part II:
DeepLab - Semantic Image Segmentation with DCNN, ASPP and Fully Connected FRFs

DeepLab

Unresolved challenges:

- Reduced feature resolution
  - 8 - 32 times downsample
- Poor prediction on multi-scale objects
- High uncertainty near object boundaries
  - Trucks sometimes can have size of half the image
  - Intrinsic problem of CNN

DeepLab: resolution

Why feature resolution reduced?

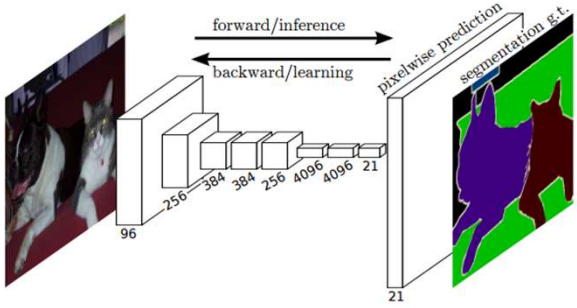


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

DeepLab: Challenge 1

How to solve reduced resolution?

Do not downsample

Convolution on large images ⇒ Small FOV

Enlarge kernel size

Use **Atrous Convolution.**

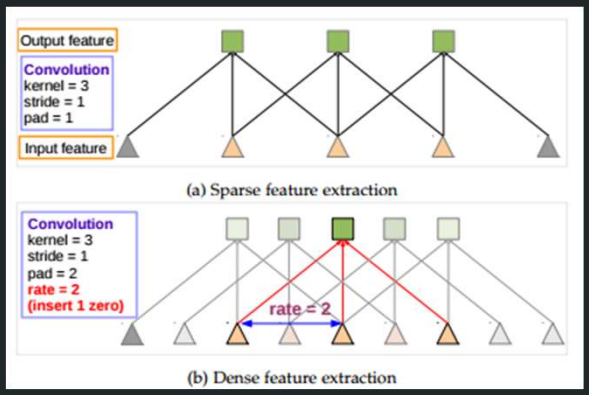
$O(n^2)$ more parameters ⇒ getting close to fully connected layer, slow training, overfitting…

- a.k.a Dilated convolution
- Large FOV with little parameters ⇒ Kill two birds with one stone!

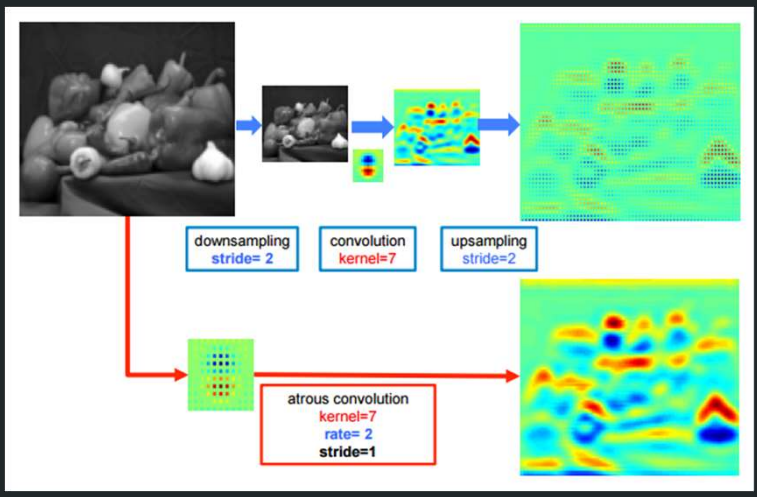
DeepLab: Challenge 1

Atrous Convolution

- "Insert hole" into convolution kernel
- Large receptive field with 'sparse' parameters



DeepLab: Challenge 1



DeepLab: Challenge 2

Multi-scale object problem?

- Objects of the same type can be hugely different in size
- Previous CNN (AlexNet, VGGnet, FCN) models yield bad result

How to get features with different scale?

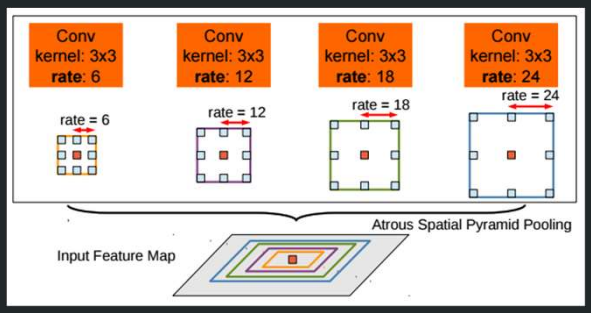
- Multi scale training, or
- Spatial Pyramid Pooling

Combined with Atrous Convolution, we have **Atrous Spatial Pyramid Pooling**

DeepLab: Challenge 2

Atrous Spatial Pyramid Pooling

- Motivated by the spatial pyramid pooling
- Filter maps with multiple scales controlled by **dilation rate**



DeepLab: Challenge 3

Poor performance near object boundaries

- Intrinsic CNN problem
- Convolution gives smooth outputs in small neighborhoods
- But ground truth varies rapidly at boundaries

How to improve results near boundary?

- **Conditional Random Fields**

DeepLab: Challenge 3

Conditional Random Field (CRF) is a graphical model where nodes are locally connected. It calculates the output probabilities at each node given neighborhood information w.r.t. some predefined energy function

$$p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w \neq v) = p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w \sim v)$$

~ means that u and v are connected in G.

Local approximation - only information at adjacent nodes is taken into consideration

DCNN's outputs probability vectors at each pixel, then CRF refines the output

DeepLab: Challenge 3

Training of CRF:

Energy definition (Identical to [Krahenbuhl et al])

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

Minimization through iterative **compatibility transform**

DeepLab: Challenge 3

Able to capture edge details and iteratively refine the prediction

Efficient CRF [Krahenbuhl et al] achieves 0.5 sec/image on PASCAL VOC.

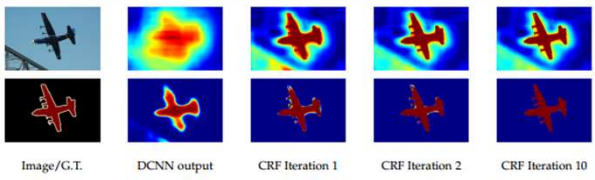


Fig. 5: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference.

DeepLab: Effectiveness of ASPP and CRF

Experiments with

- Different kernel sizes
- Before / After CRF

Results

- Larger dilation ⇒ higher performance, less parameters

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

DeepLab Whole Model

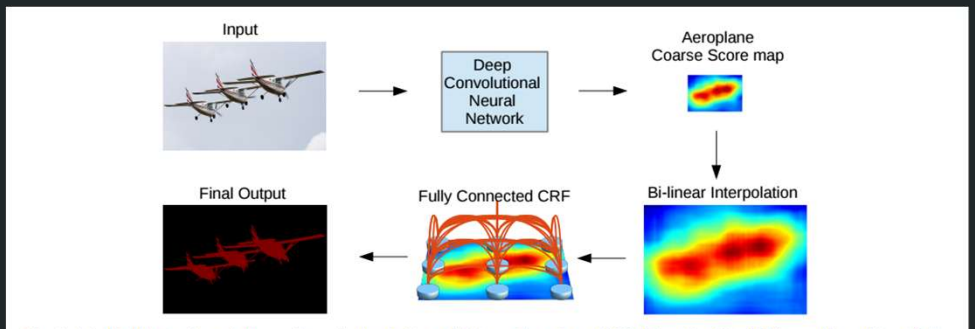


Fig. 1: Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries.

DeepLab Experiment Detail 1

"poly" learning rate policy

More effective way to change the learning rate → $(1-\frac{iter}{max\_iter})^{power})$

Yields better result

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

DeepLab Experiment Detail 2

Various training tricks

MSC	COCO	Aug	LargeFOV	ASPP	CRF	mIOU
						68.72
✓						71.27
✓	✓					73.28
✓	✓	✓				74.87
✓	✓	✓	✓			75.54
✓	✓	✓		✓		76.35
✓	✓	✓		✓	✓	77.69

TABLE 4: Employing ResNet-101 for DeepLab on PASCAL VOC 2012 *val* set. **MSC**: Employing mutli-scale inputs with max fusion. **COCO**: Models pretrained on MS-COCO. **Aug**: Data augmentation by randomly rescaling inputs.

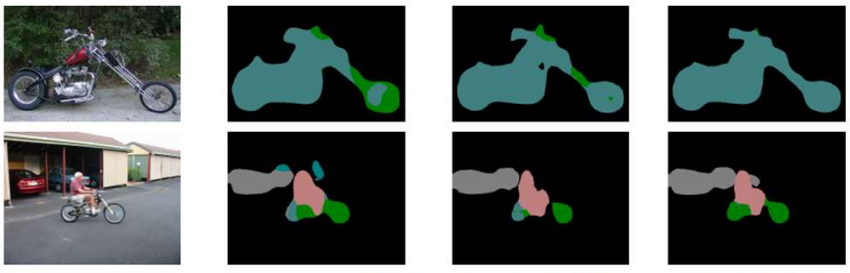
DeepLab: Results on PASCAL VOC 2012

Atrous Spatial Pyramid Pooling

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, $r = \{6, 12, 18, 24\}$.

DeepLab: Results on PASCAL VOC 2012



(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.

Performance on PASCAL VOC 2012 - Qualitative

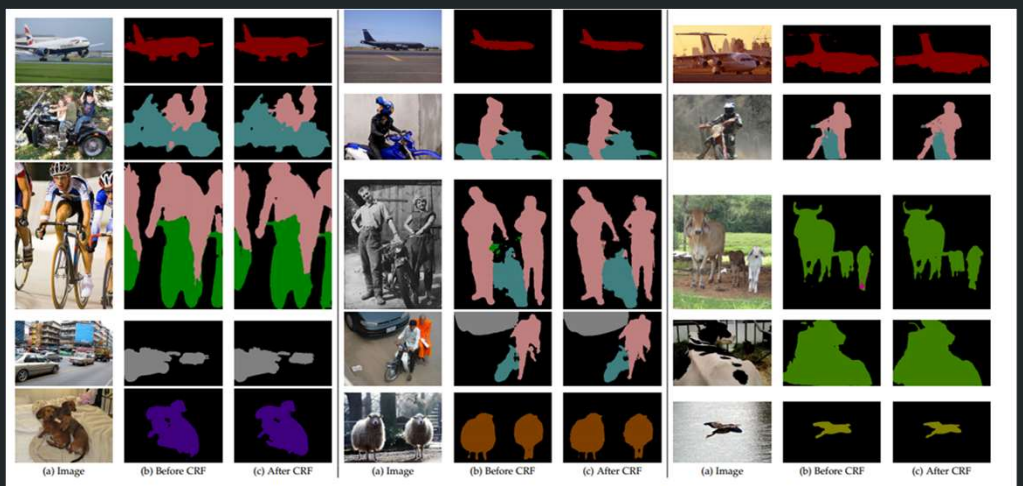


Fig. 6: PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.

Performance on PASCAL VOC 2012 - Quantitative

Method	mIOU
DeepLab-CRF-LargeFOV-COCO [58]	72.7
MERL_DEEP_GCRF [89]	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 [90]	76.8
LRR_4x_COCO [91]	76.8
DPN [62]	77.5
Adelaide_Context [40]	77.8
Oxford_TVG_HO_CRF [88]	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.

Backbone network: VGG-16 vs. ResNet-101

DeepLab based on ResNet-101 delivers better segmentation results

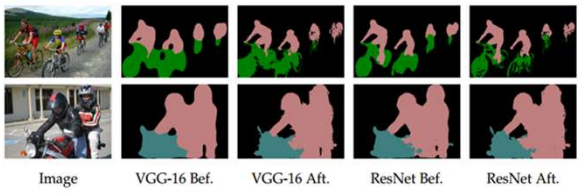


Fig. 9: DeepLab results based on VGG-16 net or ResNet-101 before and after CRF. The CRF is critical for accurate prediction along object boundaries with VGG-16, whereas ResNet-101 has acceptable performance even before CRF.

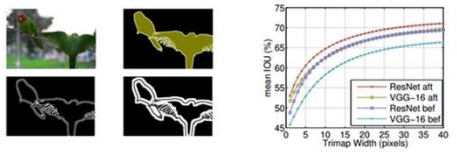


Fig. 10: (a) Trimap examples (top-left: image. top-right: ground-truth. bottom-left: trimap of 2 pixels. bottom-right: trimap of 10 pixels). (b) Pixel mean IOU as a function of the band width around the object boundaries when employing VGG-16 or ResNet-101 before and after CRF.

DeepLab Experiment Results

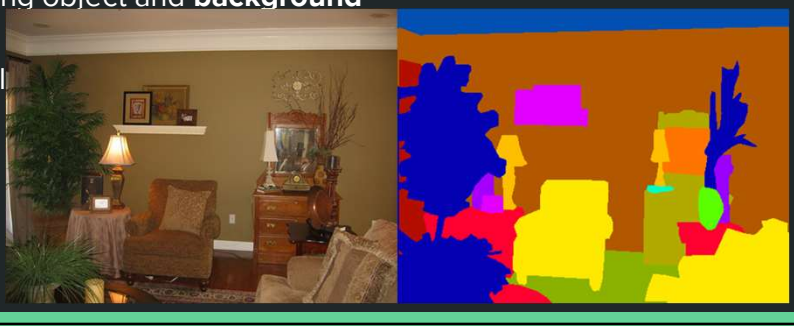
Dataset: PASCAL-Context

59 classes, approx. 5000 images

Provides detailed semantic labels for the whole scene

Including object and **background**

Ex. ceil



PASCAL-Context: Qualitative results

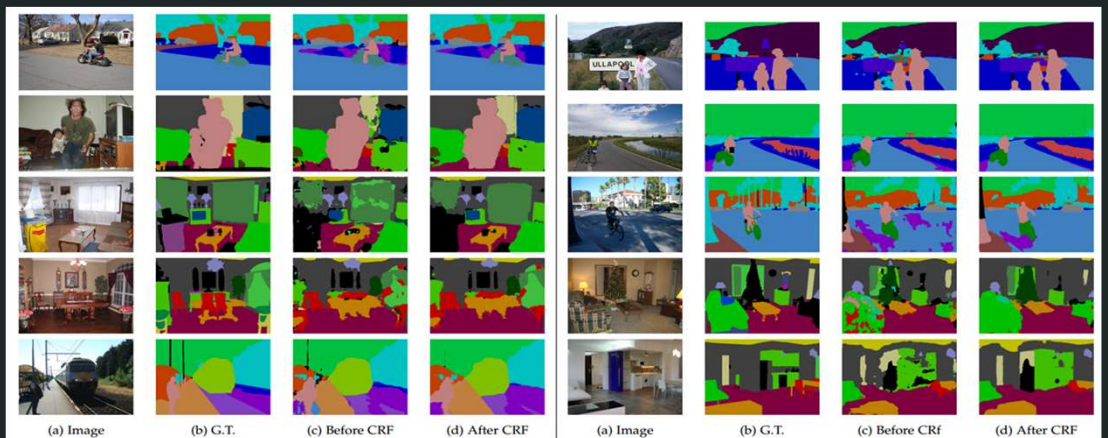


Fig. 11: PASCAL-Context results. Input image, ground-truth, and our DeepLab results before/after CRF.

PASCAL-Context: Quantitative results

Evaluation Results:

- ResNet-101 is better
- ASPP is more efficient than large FOV
- Using CRF improves the score

Method	MSC	COCO	Aug	LargeFOV	ASPP	CRF	mIOU
VGG-16							
DeepLab [38]				✓			37.6
DeepLab [38]				✓		✓	39.6
ResNet-101							
DeepLab							39.6
DeepLab	✓		✓				41.4
DeepLab	✓	✓	✓				42.9
DeepLab	✓	✓	✓	✓			43.5
DeepLab	✓	✓	✓		✓		44.7
DeepLab	✓	✓	✓		✓	✓	45.7
O_2P [45]							18.1
CFM [51]							34.4
FCN-8s [14]							37.8
CRF-RNN [59]							39.3
ParseNet [86]							40.4
BoxSup [60]							40.5
HO_CRF [88]							41.3
Context [40]							43.3
VeryDeep [93]							44.5

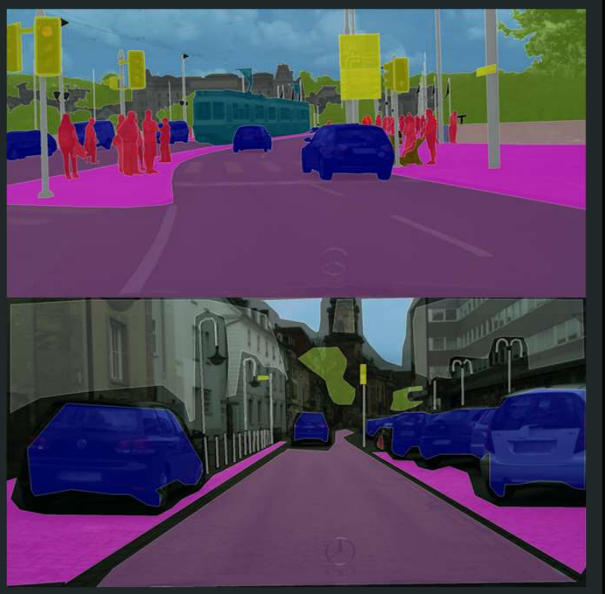
TABLE 6: Comparison with other state-of-art methods on PASCAL-Context dataset.

Dataset: Cityscapes

Street views of 50 different cities

Large-scale dataset

- Image size 1024 x 2048
- Approx. 3000 training images and hundreds of validation images
- 20,000 augmented images with coarse label



Cityscapes qualitative result:

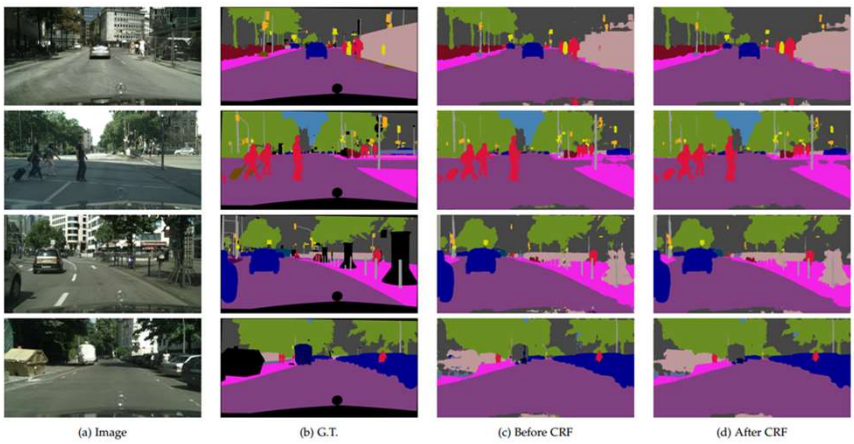


Fig. 13: Cityscapes results. Input image, ground-truth, and our DeepLab results before/after CRF.

Cityscapes Quantitative results

3rd place on the leaderboard

63.1% and 64.8% on pre-release

and official versions resp.

Method	mIOU
pre-release version of dataset	
Adelaide_Context [40]	66.4
FCN-8s [14]	65.3
DeepLab-CRF-LargeFOV-StrongWeak [58]	64.8
DeepLab-CRF-LargeFOV [38]	63.1
CRF-RNN [59]	62.5
DPN [62]	59.1
Segnet basic [100]	57.0
Segnet extended [100]	56.1
official version	
Adelaide_Context [40]	71.6
Dilation10 [76]	67.1
DPN [62]	66.8
Pixel-level Encoding [101]	64.3
DeepLab-CRF (ResNet-101)	70.4

TABLE 8: Test set results on the Cityscapes dataset, comparing our DeepLab system with other state-of-art methods.

Cityscapes val set result:

Full	Aug	LargeFOV	ASPP	CRF	mIOU
VGG-16					
		✓			62.97
		✓		✓	64.18
✓		✓			64.89
✓		✓		✓	65.94
ResNet-101					
✓					66.6
✓		✓			69.2
✓			✓		70.4
✓	✓		✓		71.0
✓	✓		✓	✓	71.4

TABLE 9: Val set results on Cityscapes dataset. **Full**: model trained with full resolution images.

Summary

DeepLab's model significantly advances the state-of-art in several challenging datasets

Successful combine the ideas from the deep convolutional neural networks and fully-connected conditional random field to produce accurate predictions and detailed segmentation maps

Q&A

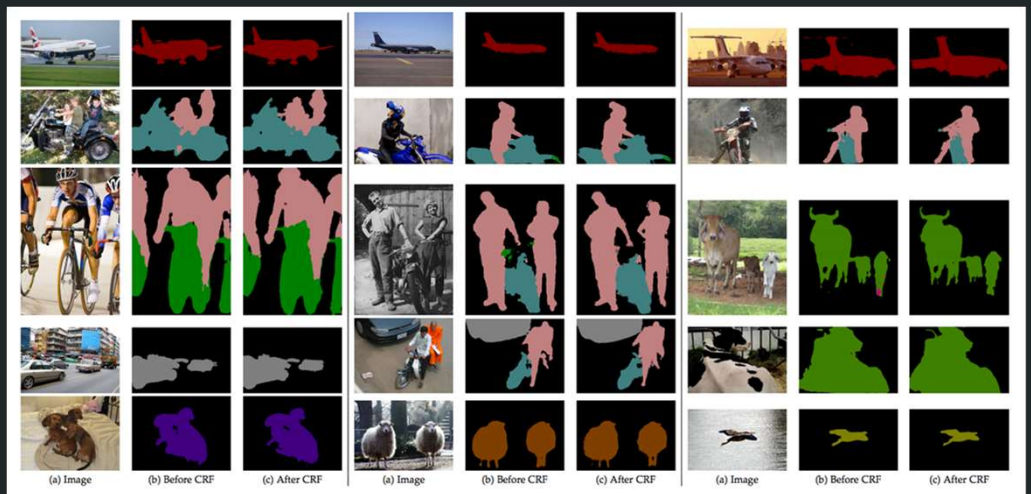


Fig. 6: PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.

Reference:

- L-C Chen, G. Papandreou, I. Kokkions, K. Murphy, and A.L. Yuille. Semantic Image Segmentation with Deep Convolutional Neural Networks. International Conference on Learning Representations. 2015.
- Jon Long, Evan Shelhamer, Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. Computer Vision and Pattern Recognition. 2015
- P Krähenbühl, V Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. Neural Information Processing Systems. 2011

Dataset: PASCAL-Person-Part

- Contains detailed part annotations for each person, including eyes, nose.
- Merge the annotations to be head, torso, upper/lower arms and upper. Lower legs
- Resulting in 6 person part class and one background

Dataset: PASCAL-Person-Part

Method	MSC	COCO	Aug	LFOV	ASPP	CRF	mIOU
ResNet-101							
DeepLab							58.90
DeepLab	✓		✓				63.10
DeepLab	✓	✓	✓				64.40
DeepLab	✓	✓	✓			✓	64.94
DeepLab	✓	✓	✓	✓			62.18
DeepLab	✓	✓	✓		✓		62.76
Attention [17]							56.39
HAZN [95]							57.54
LG-LSTM [96]							57.97
Graph LSTM [97]							60.16

TABLE 7: Comparison with other state-of-art methods on PASCAL-Person-Part dataset.

Dataset: PASCAL-Context

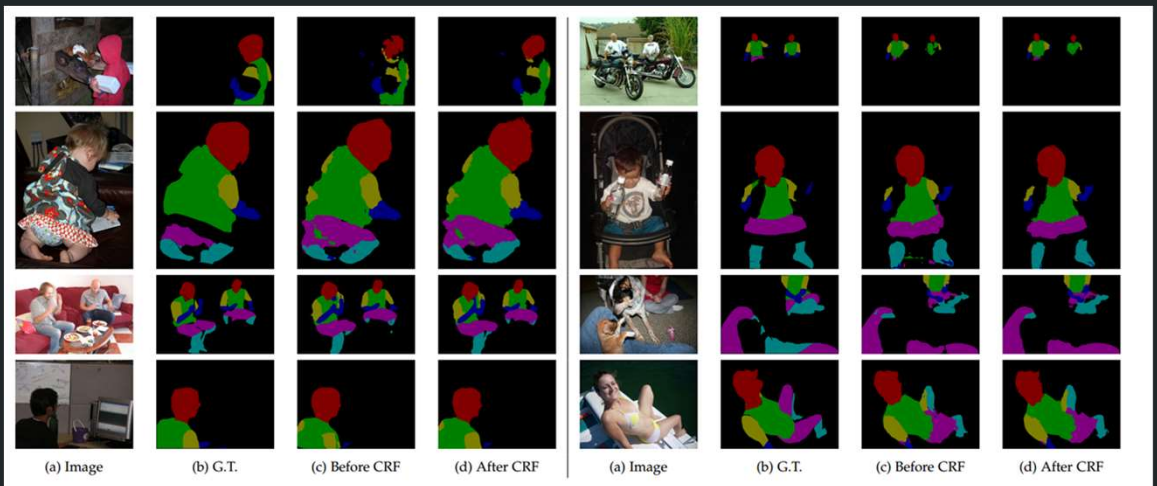


Fig. 12: PASCAL-Person-Part results. Input image, ground-truth, and our DeepLab results before/after CRF.

Dataset 1: PASCAL VOC 2012

- 20 object classes, one background class
- Thousands of images
- Augmented by extra annotations