Adaptive Context Encoding Module for Semantic Segmentation

Congcong Wang[1], Faouzi Alaya Cheikh[1], Azeddine Beghdadi[2], and Ole Jacob Elle[3,4]

[1]Norwegian University of Science and Technology, Norway.

[2]University Paris 13, France.

[3]Oslo University Hospital, Norway

[4]University of Oslo, Norway

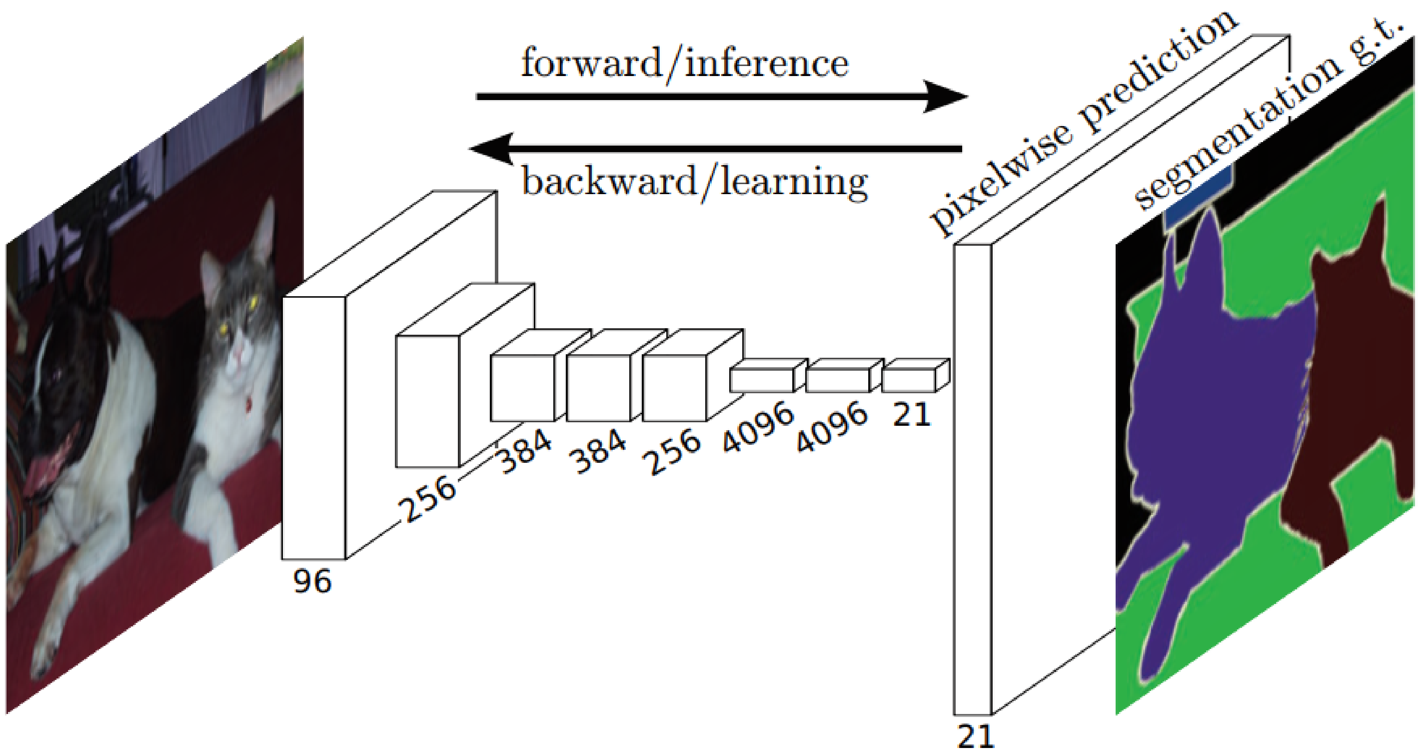
IS&T International Symposium on Electronic Imaging 2020 SCIENCE AND TECHNOLOGY

Semantic Segmentation

- Semantic Segmentation
  - Understanding an image at pixel level
  - Partitioning an image into regions of meaningful objects
  - Comprehensive scene description: object category, location, etc.

Zhang et al. Context Encoding for Semantic Segmentation

Fully Convolutional Network (FCN)



Long et al. Fully Convolutional Networks for Semantic Segmentation

- Pre-trained CNN + Decoder
- Trained end-to-end
- → Meta algorithm for semantic segmentation

What problem we are addressing?

- The existence of objects at multiple scales.

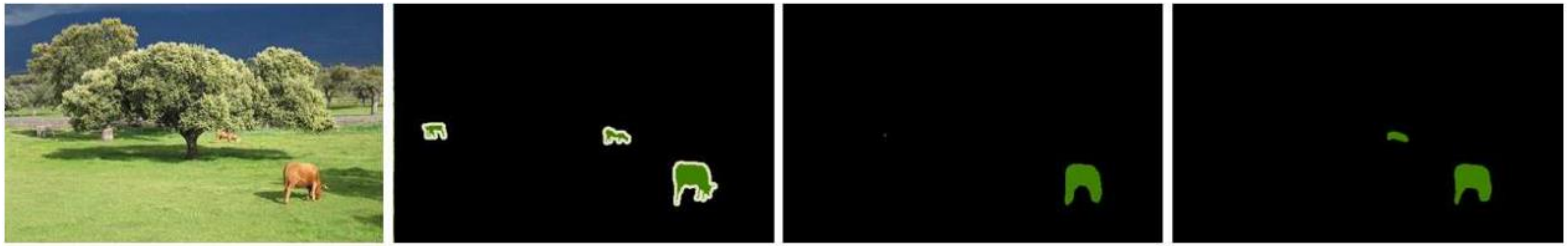
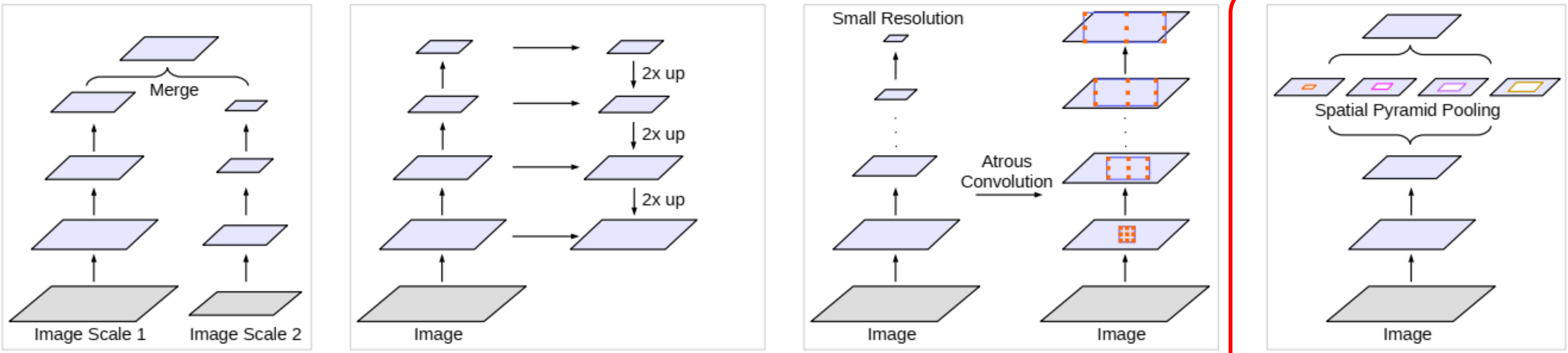


Image | Ground Truth | Result obtained by single scale | Result obtained by multi-scale

He et al. Adaptive Pyramid Context Network for Semantic Segmentation.

- How to capture multiple scale information?

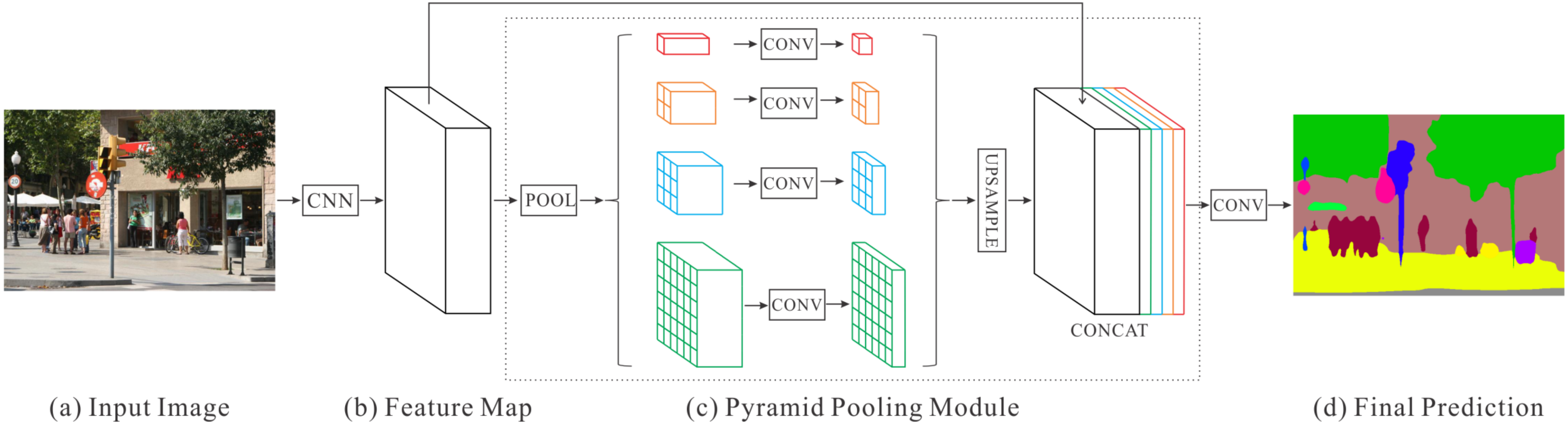


(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

Chen et al. Rethinking Atrous Convolution for Semantic Image Segmentation

Prior work

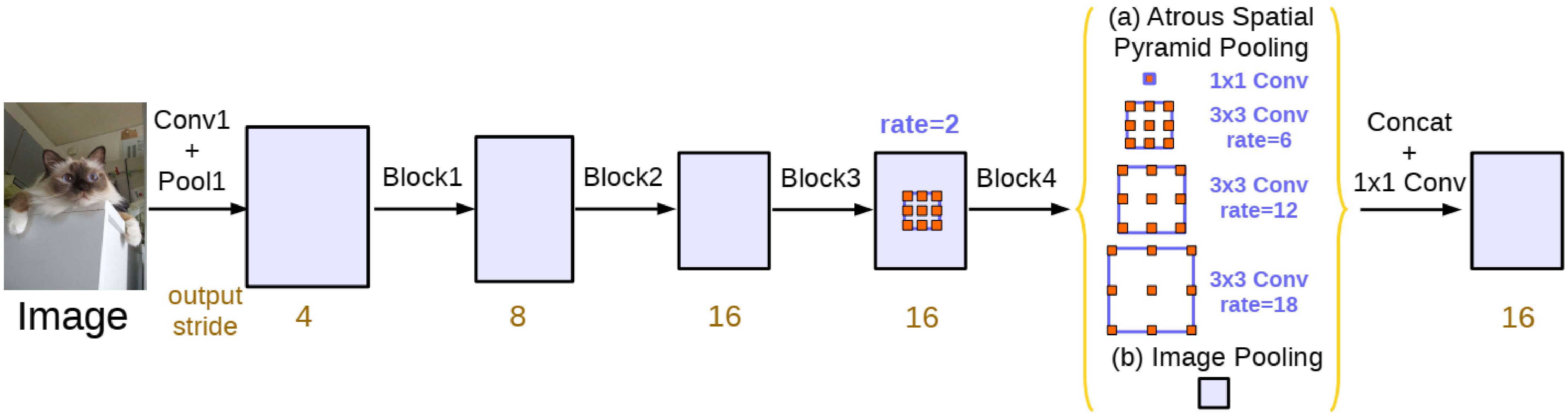
- Spatial Pyramid Pooling strategy
  - Pyramid Pooling Module (PPM) from PSPNet



Zhao et al. Pyramid Scene Parsing Network

Prior work

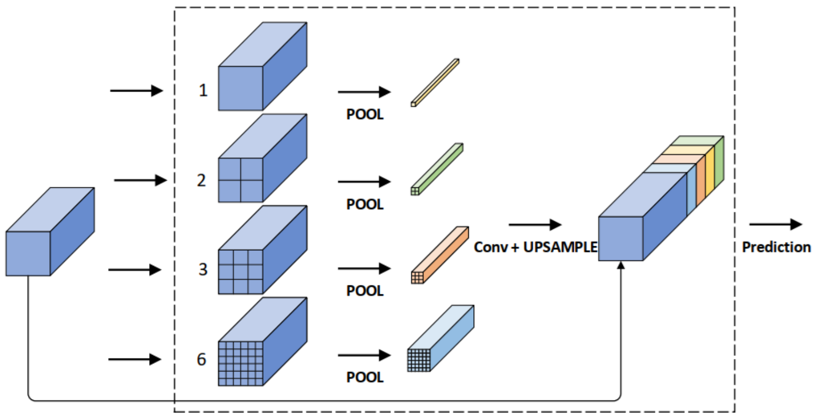
- Spatial Pyramid Pooling strategy
  - Atrous Spatial Pyramid Pooling from Deeplabs.

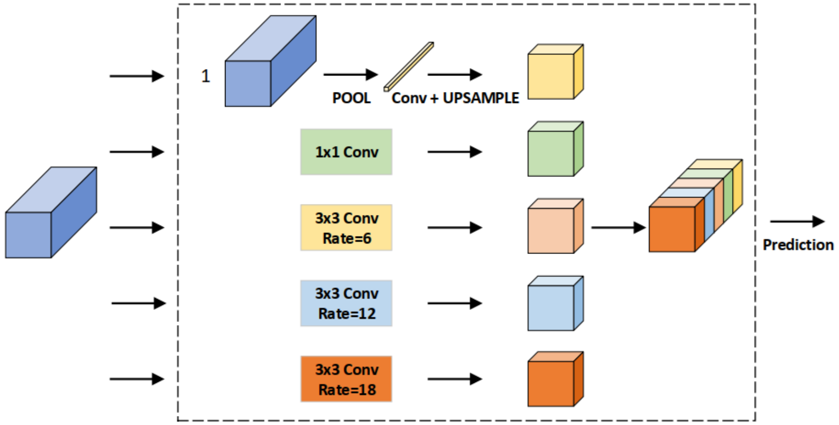


Chen et al. Rethinking Atrous Convolution for Semantic Image Segmentation

Prior work

- Two problems:
  - (1) The numbers of sub-region of PPM in PSPNet and the atrous rates of ASPP module from Deeplabs are selected empirically.
  - (2) PPM and ASPP both extract the context information by sampling from rigid rectangular regions which contain pixels from different object categories.

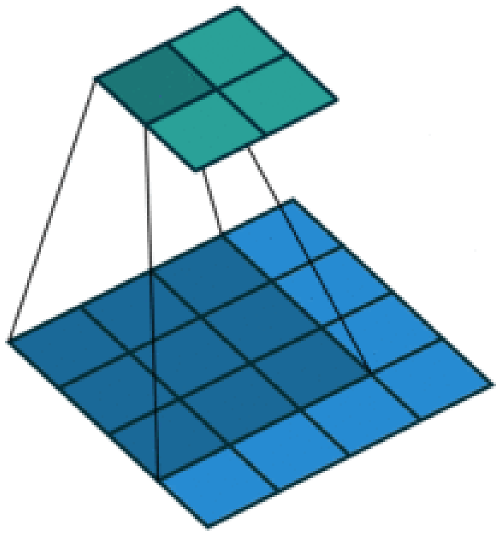




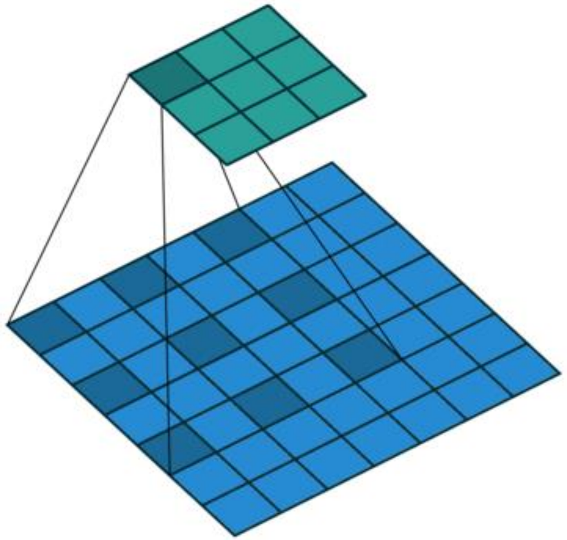
- How to solve the above two problem?

Rethinking ASPP

- Convolution operation (no padding, no strides)

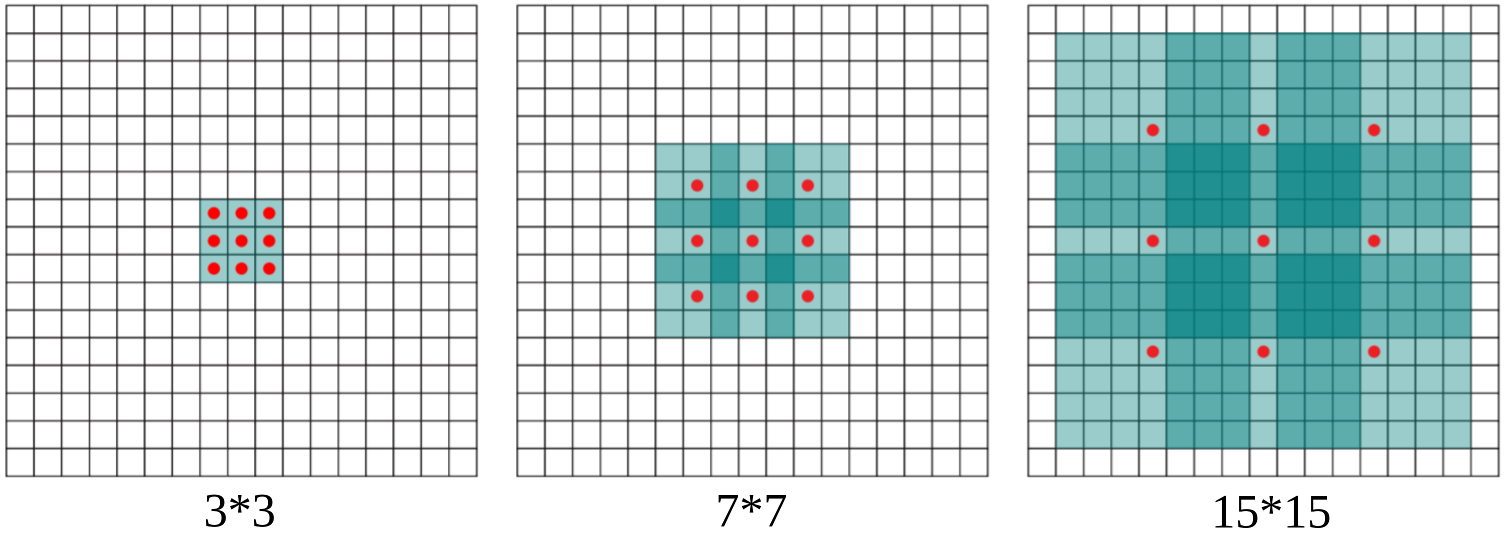


- Atrous convolution operation (atrous rate = 1)



Dumoulin et al. CNN operation demo A guide to convolution arithmetic for deep learning

Rethinking ASPP

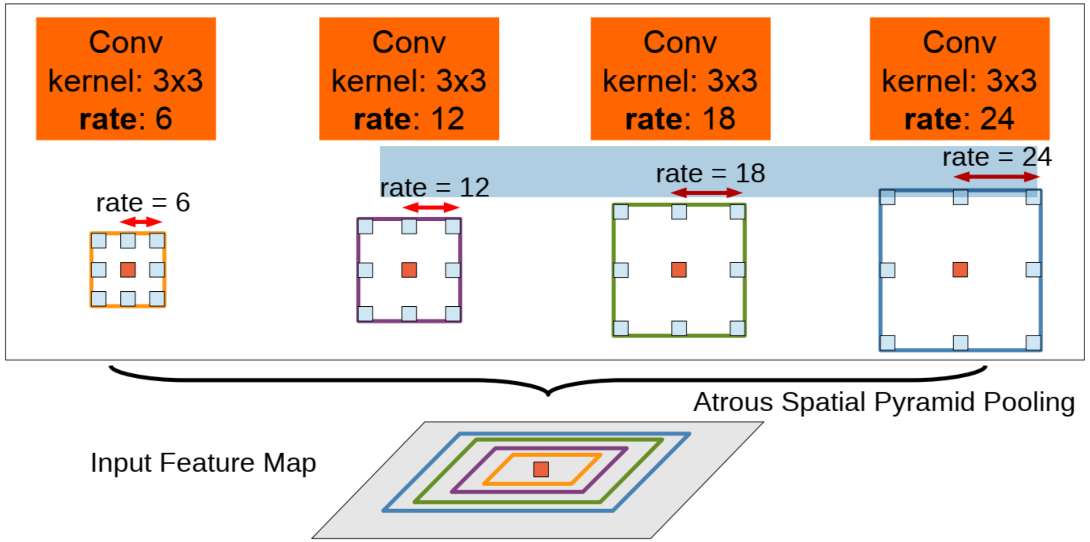


Yu et al. Multi-scale context aggregation by dilated convolutions

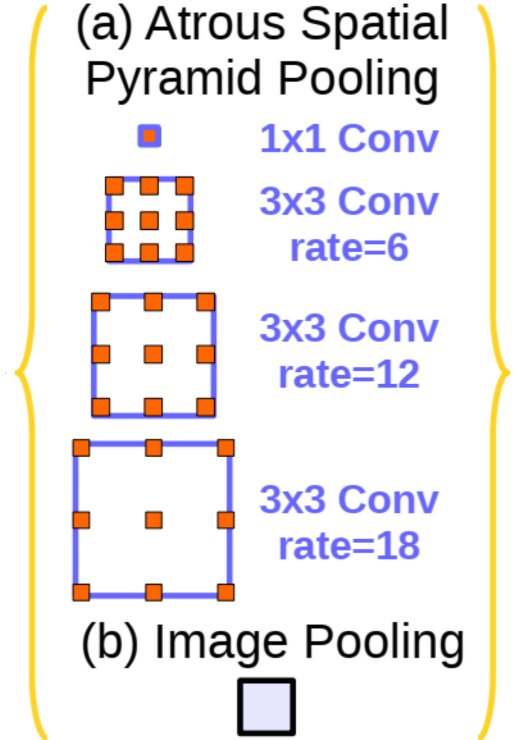
Small atrous rate → Small field of view → accurate localization
Large atrous rate → Large field of view → context assimilation
➔
Different values of atrous rate → different sample locations of the convolution operation → multiple effective fields of view → capturing multi-scale context information

Rethinking ASPP

- **Problem:** Existence of objects at multiple scales
- **Solution:** Different values of atrous rate → different sample locations of the convolution operation → multiple effective fields of view → capturing multi-scale context information

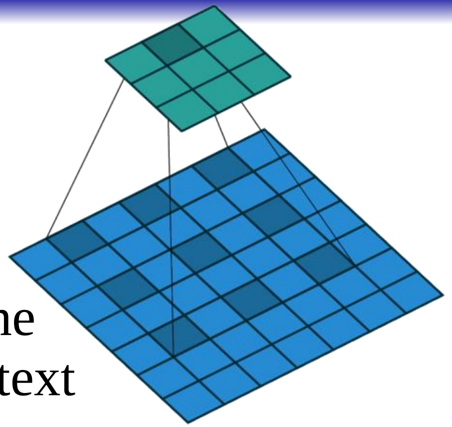


Chen et al. Deeplabv2

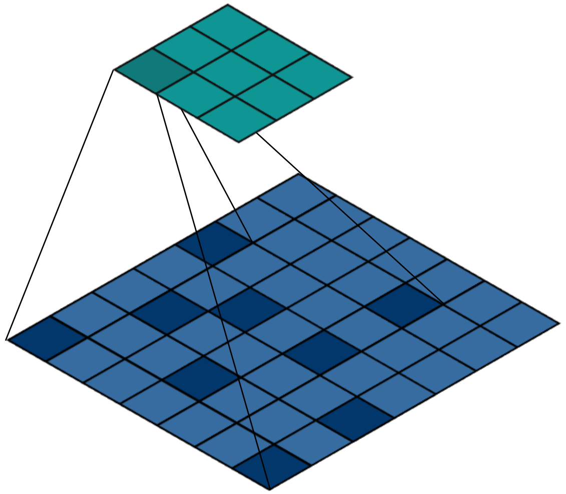


Chen et al. Deeplabv3

Main idea



- Problem: existence of objects at multiple scales
- Solution of ASPP: Different values of atrous rate → different sample locations of the convolution operation → multiple effective fields of view → capturing multi-scale context information
- Our idea: Learned sample locations of the convolution operation → learned effective fields of view → capturing multi-scale context information adaptively
- Learn the sample locations
  - No need to decide the atrous rate manully – solving issue 1
  - No need to sample the pixel in a rectangular – solving issue 2
  - Tool: Deformable convolution



Dai et al. Deformable ConvNets v1
Zhou et al. Deformable ConvNets v2

Main idea

- ❑ Convolutional Kernel (3*3):

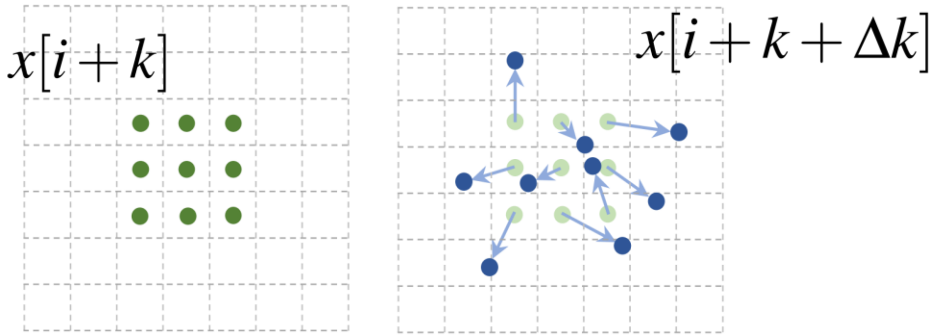
$$K = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$$

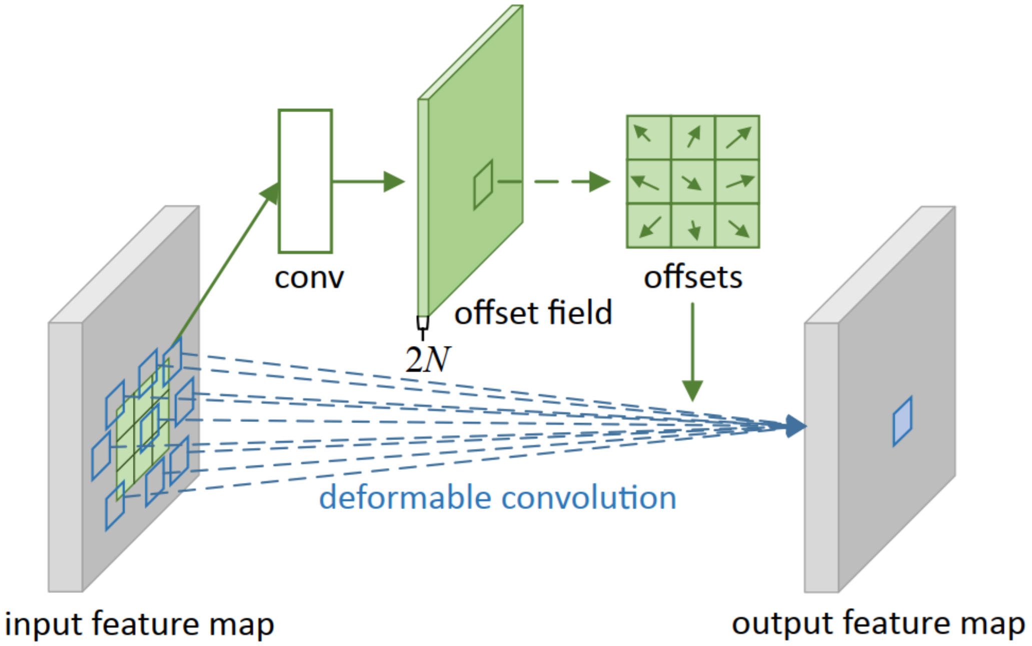
- ❑ Standard convolution operation:

$$y[i] = \sum_{k \in K} x[i+k] \cdot w[k]$$

- ❑ Deformable convolution operation:

$$y[i] = \sum_{k \in K} x[i+k+\Delta k] \cdot w[k]$$

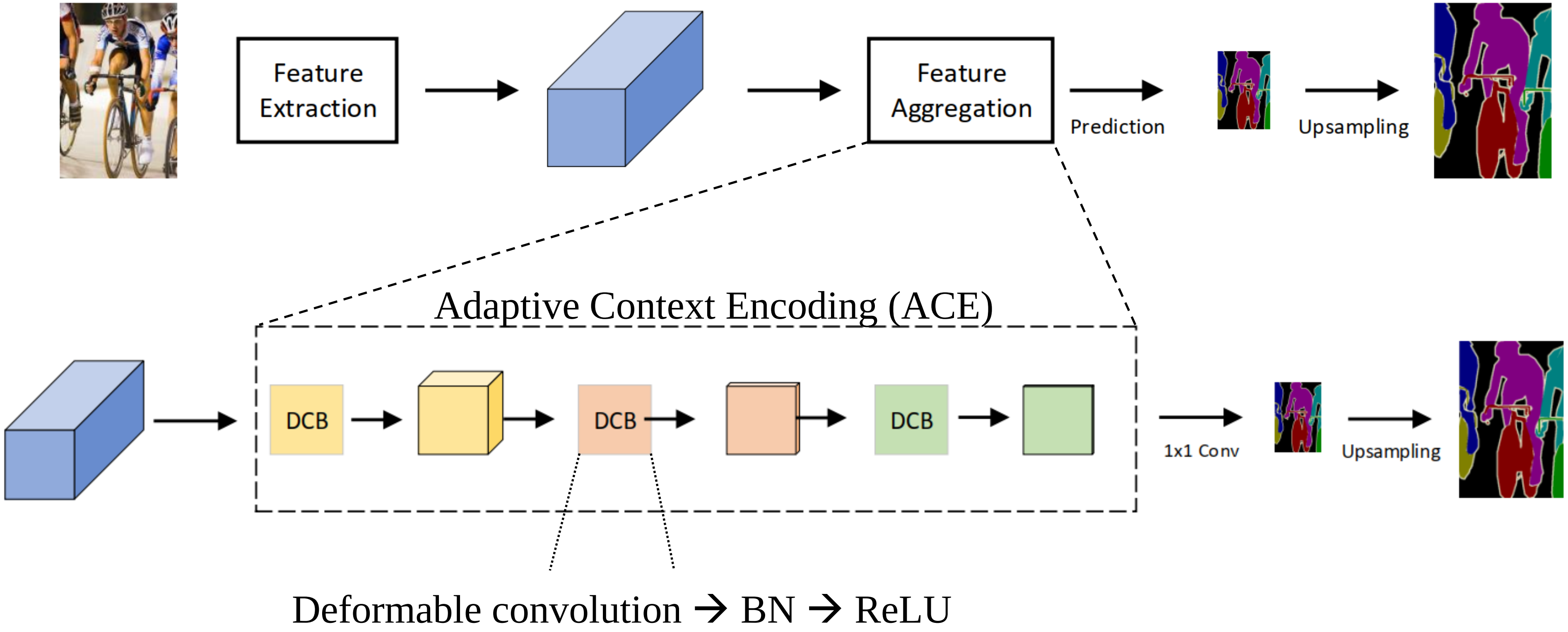




Dai et al. Deformable ConvNets v1

Main idea

- Network Architecture



Comparison to PSP and Deeplab

- Pascal-Context
  - 4998 training images
  - 5105 testing images
  - 59 object classes with background

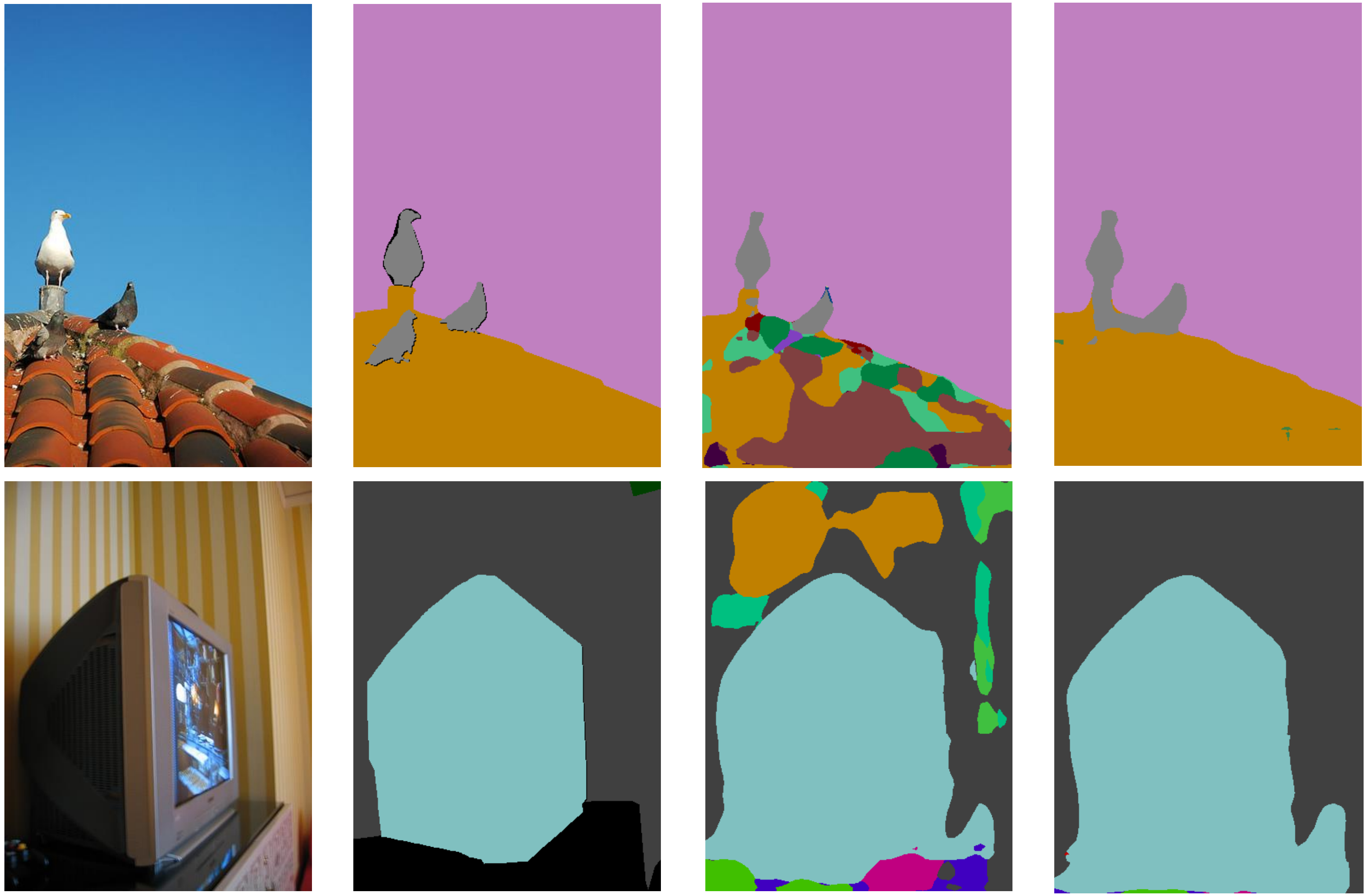
Batch Size	Head	pixAcc%	mIoU%
4	ASPP	75.42	43.62
	PPM	75.58	45.68
	Proposed	**77.68**	**48.07**
6	ASPP	77.19	46.53
	PPM	77.45	48.32
	Proposed	**78.35**	**49.36**
16	ASPP	78.68	49.04
	PPM	78.41	49.54
	Proposed	**78.85**	**50.35**

- ADE20k
  - 150 object classes
  - 20k images for training
  - 2k/3k images validation and testing

Batch Size	Head	pixAcc%	mIoU%
4	ASPP	78.11	37.11
	PPM	77.39	37.80
	Proposed	**78.62**	**38.51**

Some visual results

Original Images | Ground Truth | Deeplabv3* | Proposed



Some visual results

Original Images	Ground Truth	Deeplabv3*	Proposed

Comparison to state-of-the-art

- Pascal-Context

Method	mIoU%
FCN-8s [1]	37.8
ParseNet [26]	40.4
Piecewise [8]	43.3
Deeplabv2 (Res101-COCO) [3]	45.7
RefineNet (Res152) [9]	47.3
PSPNet (Res101) [13]	47.8
EncNet (Res101) [33]	51.7
DANet (Res101) [34]	52.6
FastFCN (Res101,EncNet)* [22]	53.1
Proposed (Res101)	**53.6**

* FastFCN backbone with EncNet *head*.

Comparison to state-of-the-art

- ADE20K

Method	pixAcc%	mIoU%
FCN [1]	71.32	29.39
SegNet [4]	71.00	21.64
DilatedNet [35]	73.55	32.31
CascadeNet [18]	74.52	34.90
RefineNet (Res152) [9]	-	40.7
PSPNet (Res101) [13]	81.39	43.29
EncNet (Res101) [33]	**81.69**	**44.65**
FastFCN (Res101,EncNet)* [22]	80.99	44.34
Proposed	81.07	43.81

* FastFCN backbone with EncNet *head*.

Discussion & Conclusion

- An ACE module is proposed for semantic segmentation to capture multiscale context information.
  - ❑ More robust and better performance is achieved compared to ASPP and PPM modules.
  - ❑ State-of-the-art results on Pascal-Context and encouraging results on ADE20K is shown.
- The proposed context aggregation module can be easily embedded into other semantic segmentation networks for further improvement.

- Future consideration
  - ❑ Is deformable convolution operation the right tool?
  - ❑ Visual result indicates that the network prone to output smooth result and easy ignore small detail. How to explain and improve it?

Thank you