

CSE 252D: Advanced Computer Vision

Manmohan Chandraker

Lecture 13: Semantic Segmentation



Virtual classrooms

- Virtual lectures on Zoom
 - Only host shares the screen
 - Keep video off and microphone muted
 - But please do speak up (remember to unmute!)
 - Slides uploaded on webpage just before class
- Virtual interactions on Zoom
 - Ask and answer plenty of questions
 - “Raise hand” feature on Zoom when you wish to speak
 - Post questions on chat window
 - Happy to try other suggestions!
- Lectures recorded and upload on Canvas
 - Available under “My Media” on Canvas

Overall goals for the course

- Introduce the state-of-the-art in computer vision
- Study principles that make them possible
- Get understanding of tools that drive computer vision
- Enable one or all of several such outcomes
 - Pursue higher studies in computer vision
 - Join industry to do cutting-edge work in computer vision
 - Gain appreciation of modern computer vision technologies
- This is a great time to study computer vision!

Papers for Fri, May 14

- ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation
 - <https://arxiv.org/abs/1606.02147>
- ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation
 - <https://ieeexplore.ieee.org/abstract/document/8063438>
- Fast-SCNN: Fast Semantic Segmentation Network
 - <https://arxiv.org/abs/1902.04502>
- Decoupled Deep Neural Network for Semi-Supervised Semantic Segmentation
 - <https://arxiv.org/abs/1506.04924>

Papers for Wed, May 19

- Context Encoding for Semantic Segmentation
 - <https://arxiv.org/abs/1803.08904>
- Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation
 - <https://arxiv.org/abs/1802.02611>
- High-Resolution Representations for Labeling Pixels and Regions
 - <https://arxiv.org/abs/1904.04514>
- MSeg: A Composite Dataset for Multi-Domain Semantic Segmentation
 - <https://ieeexplore.ieee.org/document/9157628>

Papers for Fri, May 21

- R-FCN: Object Detection via Region-based Fully Convolutional Networks
 - <https://arxiv.org/abs/1605.06409>
- Deformable Convolutional Networks
 - <https://arxiv.org/abs/1703.06211>
- Feature Pyramid Networks for Object Detection
 - <https://arxiv.org/abs/1612.03144>
- Training Region-Based Object Detectors with Online Hard Example Mining
 - <https://arxiv.org/abs/1604.03540>

Recap

Semantic Segmentation

Global Reasoning

Locally accurate
Boundaries

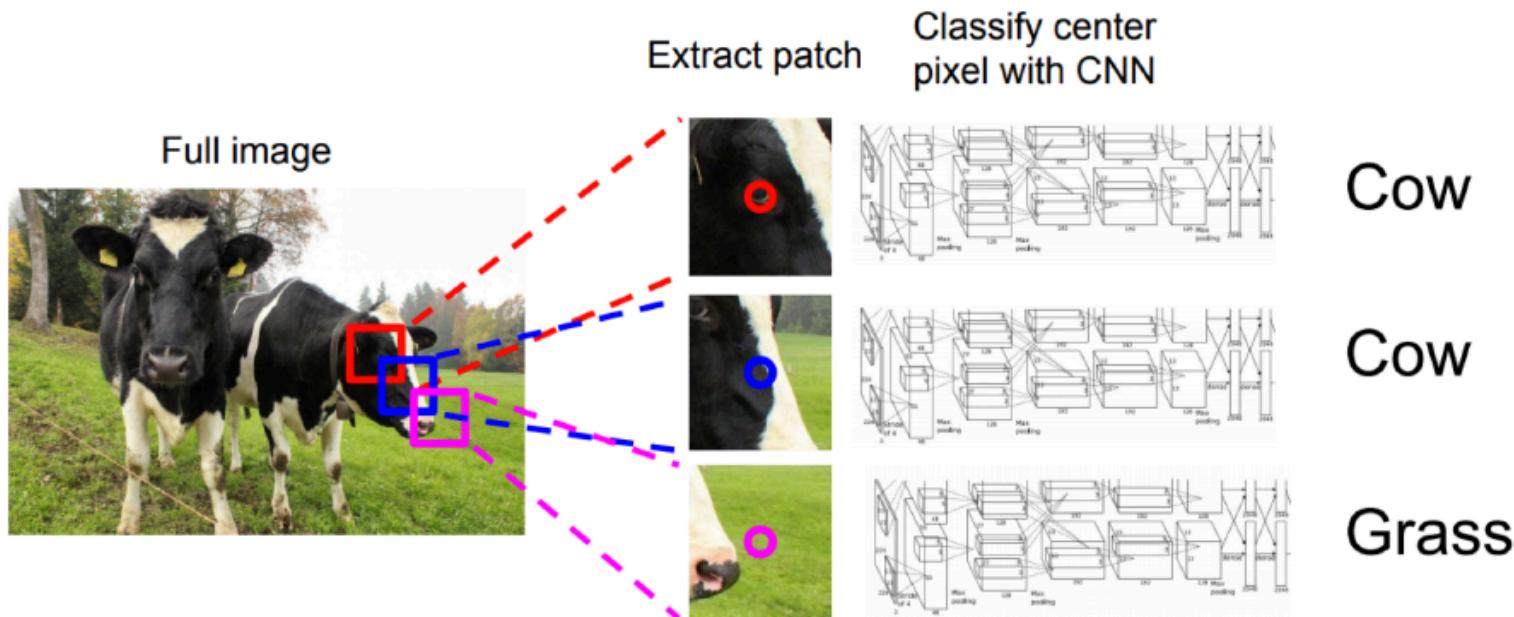


Figure from CityScapes Dataset

Per-Pixel Classification for Semantic Segmentation

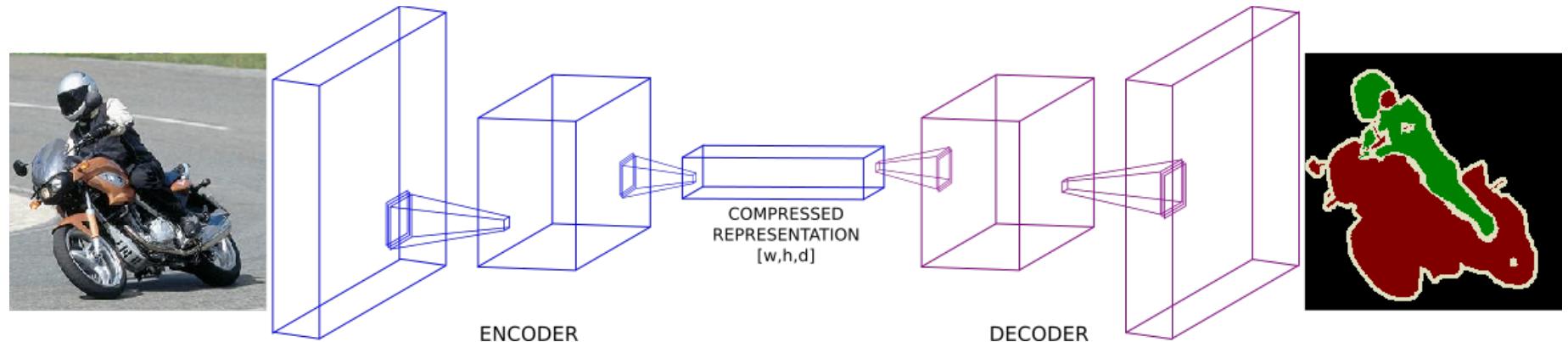
Approaches

What could be a Problem?



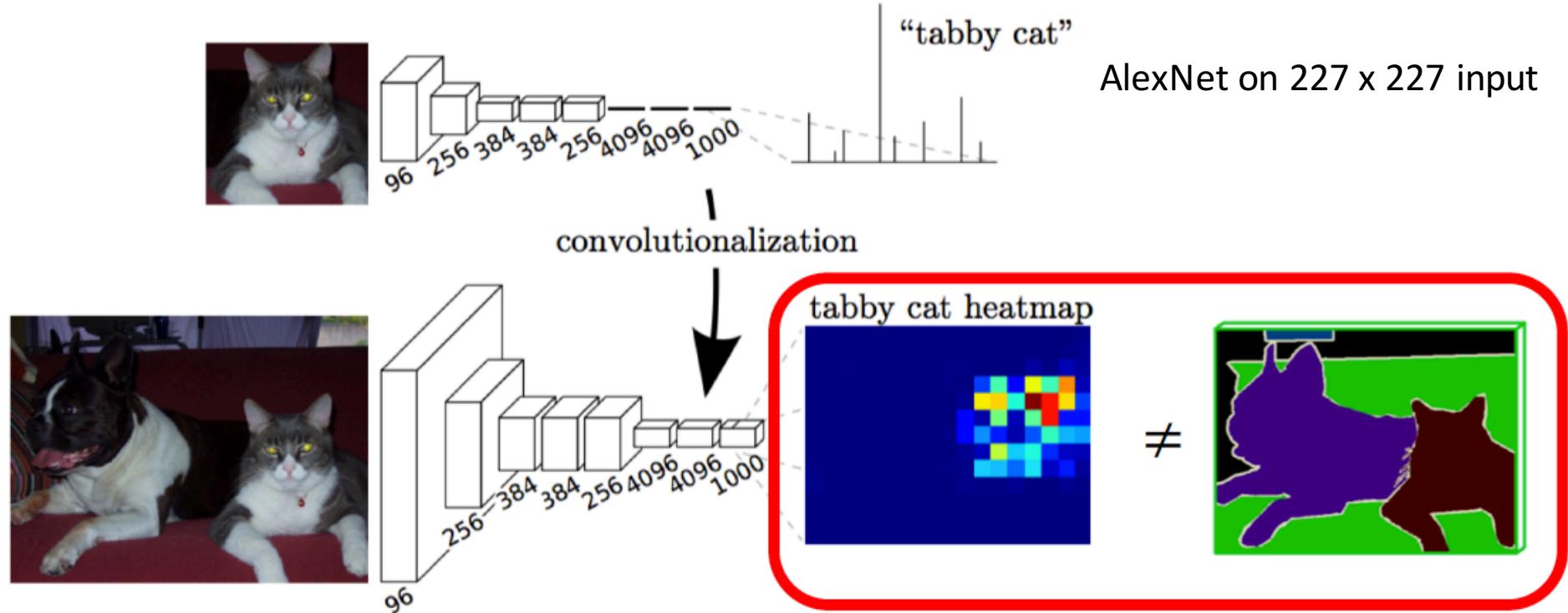
General Form of Segmentation Networks

- Most networks have similar encoders (inspired by classification networks)
 - Pooling and striding to downsample
 - Goal: do deeper convolutions without memory concerns
- Differences usually in decoder
 - Interpolation, transposed convolution, unpooling
 - Goal: recover spatial detail



Coarse Predictions

Fully-connected layer with k units = Convolution layer with k filters of size that covers input



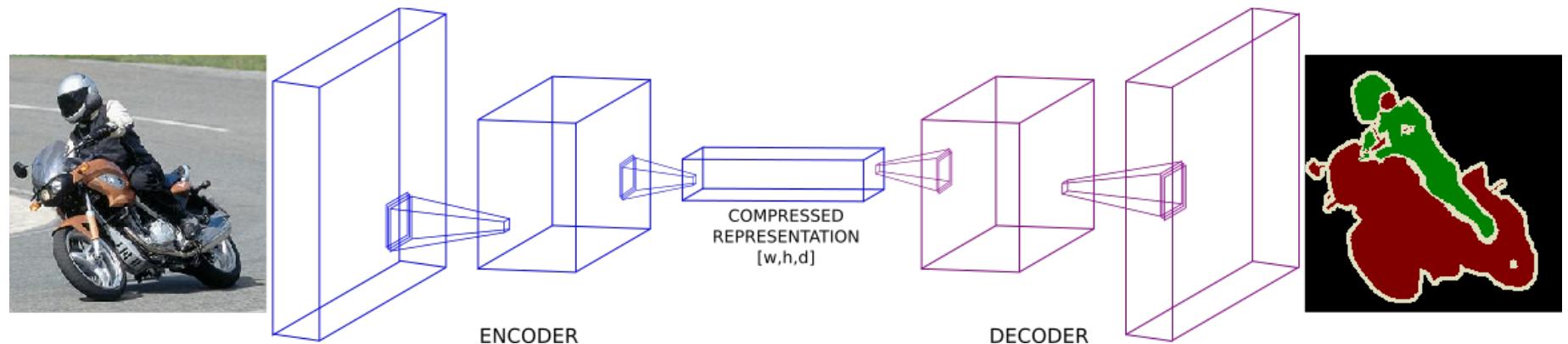
Given 500 x 500 image, slide FCN with stride 32 to get 10 x 10 output.

We want a segmentation output at image resolution.

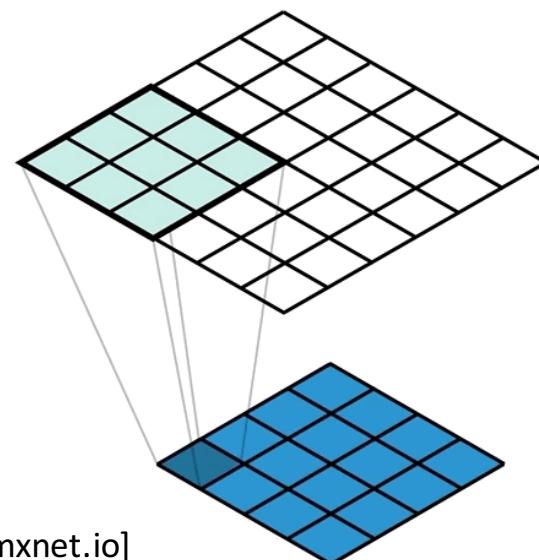
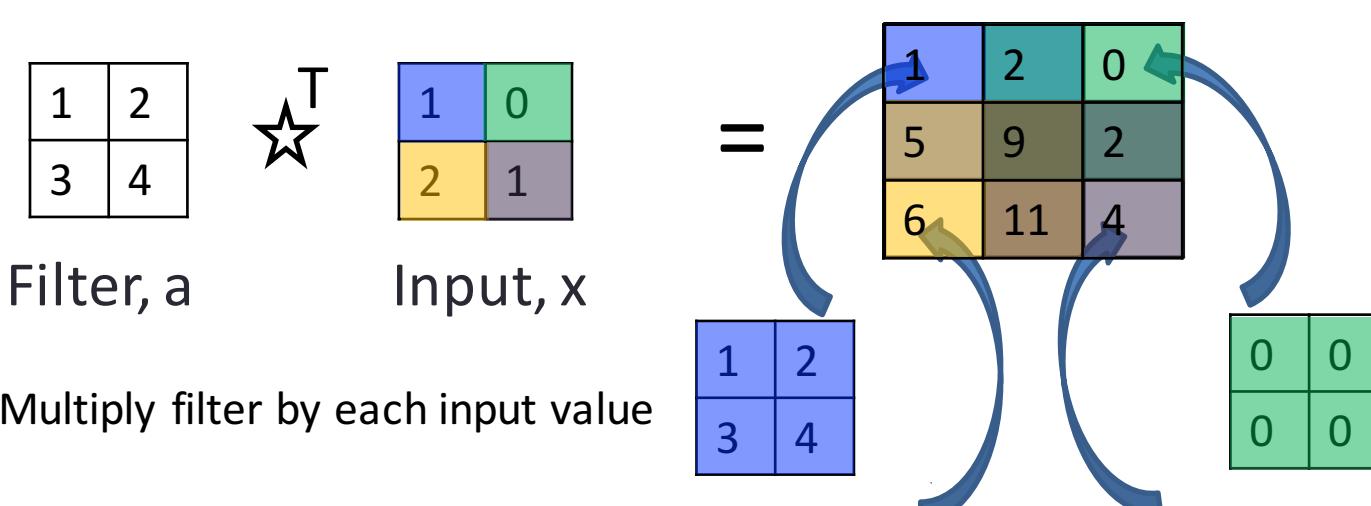
Output Going to Image Resolution

- Encoder aggressively pools and subsamples image
- Necessary to capture context information which is necessary for segmentation
- But spatial detail is also necessary

- Goal for decoder: obtain output at image resolution
- Goal for decoder: recover detail in encoder feature maps before subsampling



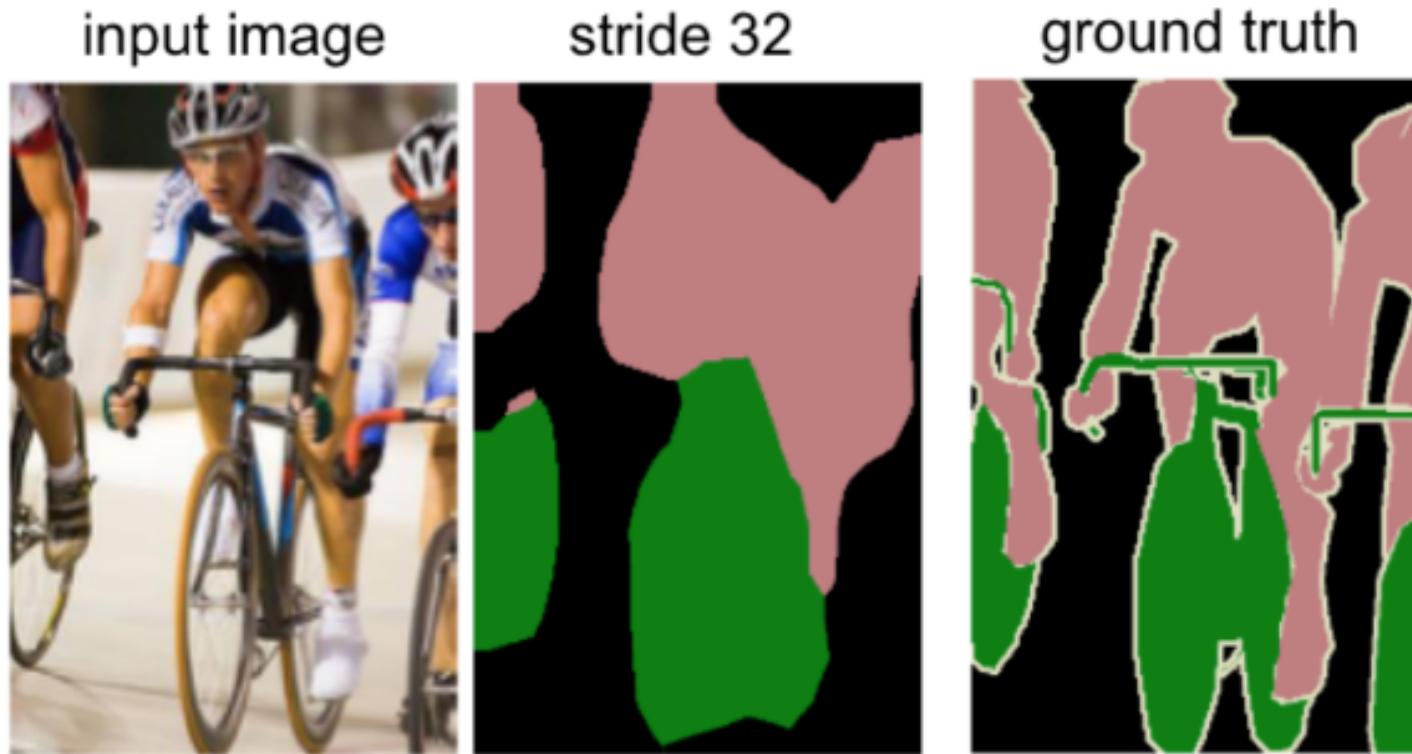
Transposed Convolutions



[Thom Lane, mxnet.io]

Tile the scaled filter at output locations
Add overlapping values

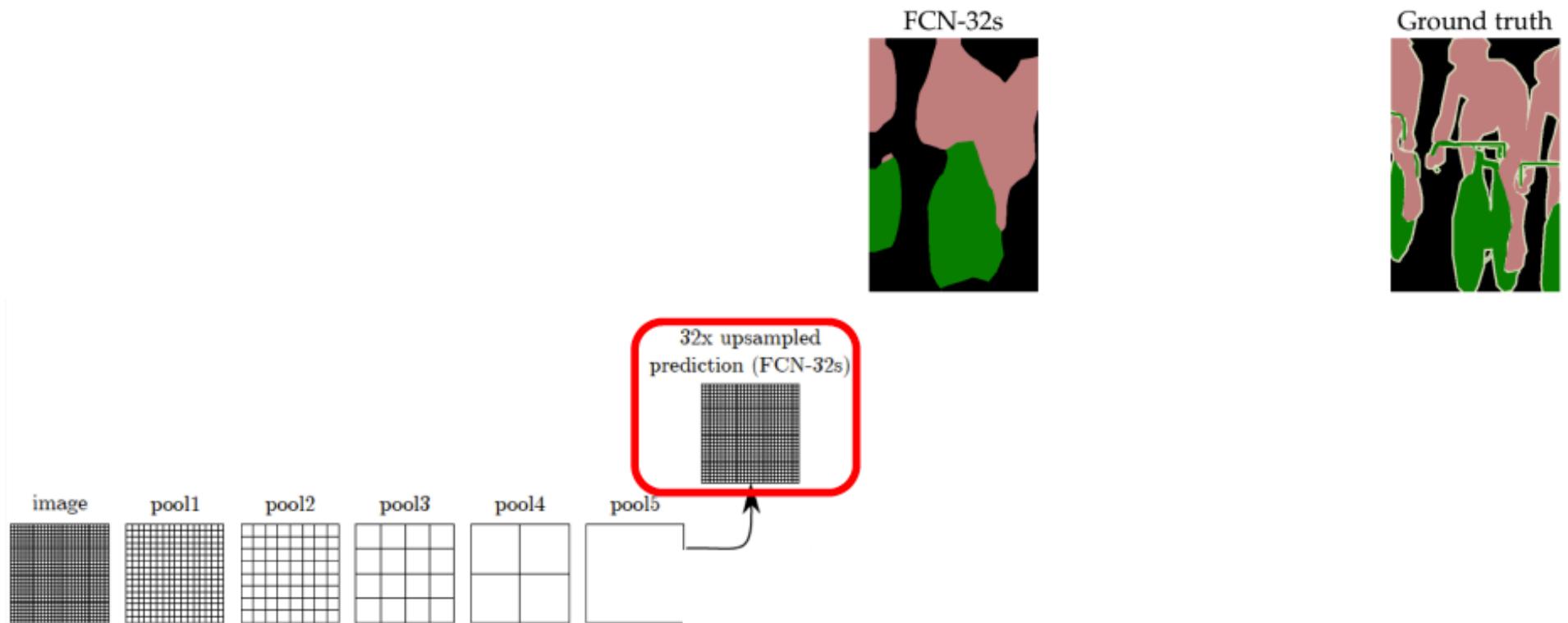
Output Going to Image Resolution



Output Going to Image Resolution

- Encoder aggressively pools and subsamples image
- Necessary to capture context information which is necessary for segmentation
- But spatial detail is also necessary
- Goal for decoder: obtain output at image resolution
- Goal for decoder: recover detail in encoder feature maps before subsampling
- Option 1
 - Use transposed convolution to upsample to image resolution
 - Concatenate encoder features to upsampled features during decoding

Combine Global and Local Information



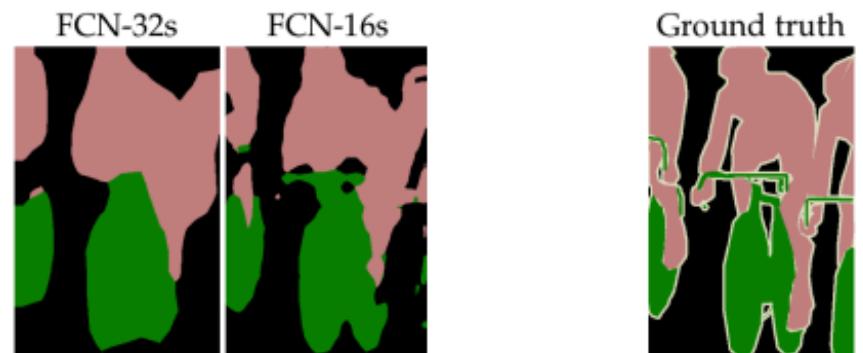
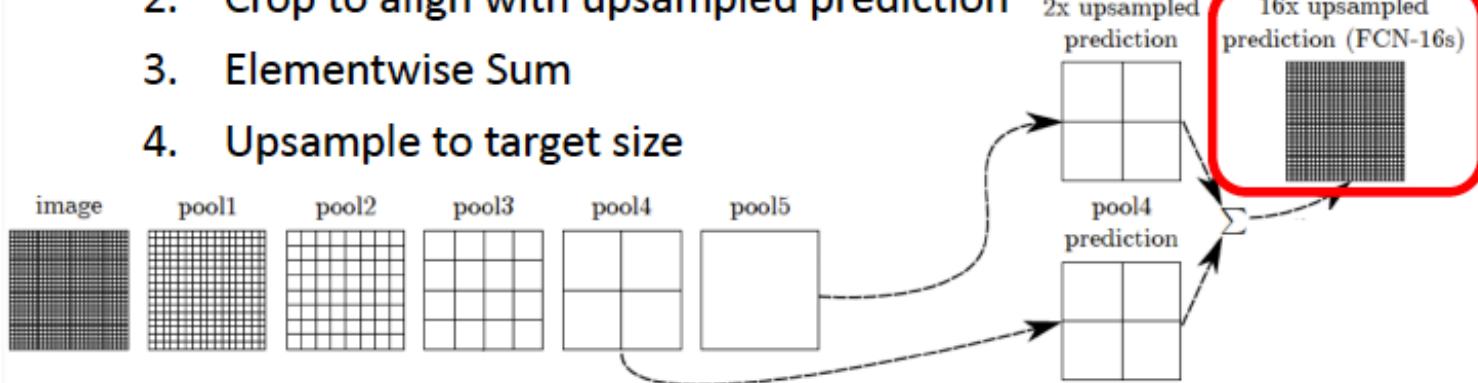
Combine Global and Local Information

Fuse coarse semantic information with local appearance

Similar to: skip connections.

Process:

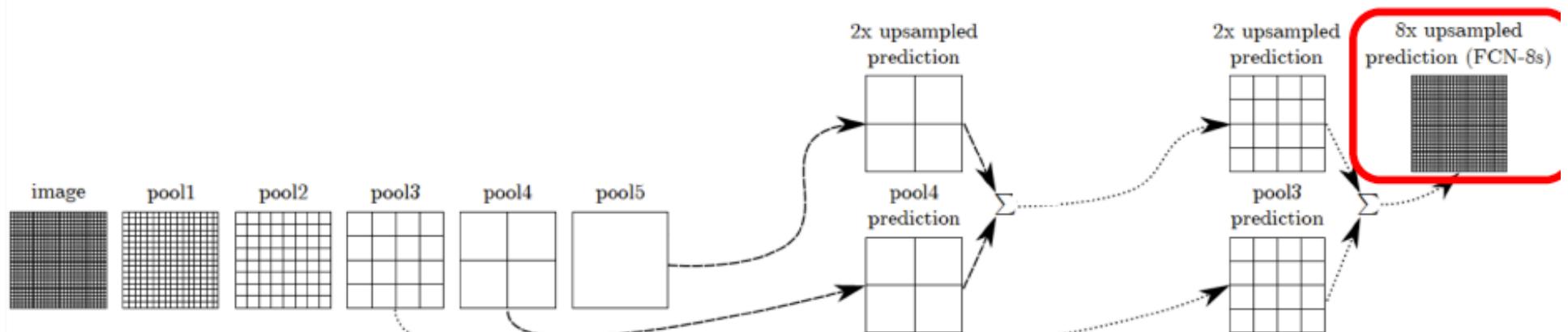
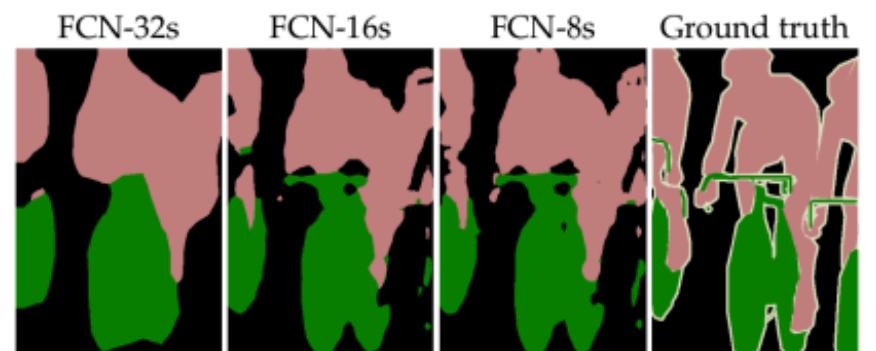
1. 1x1 Convolution on pool4 output
2. Crop to align with upsampled prediction
3. Elementwise Sum
4. Upsample to target size



Combine Global and Local Information

Fuse coarse semantic information with local appearance

Similar to: skip connections.



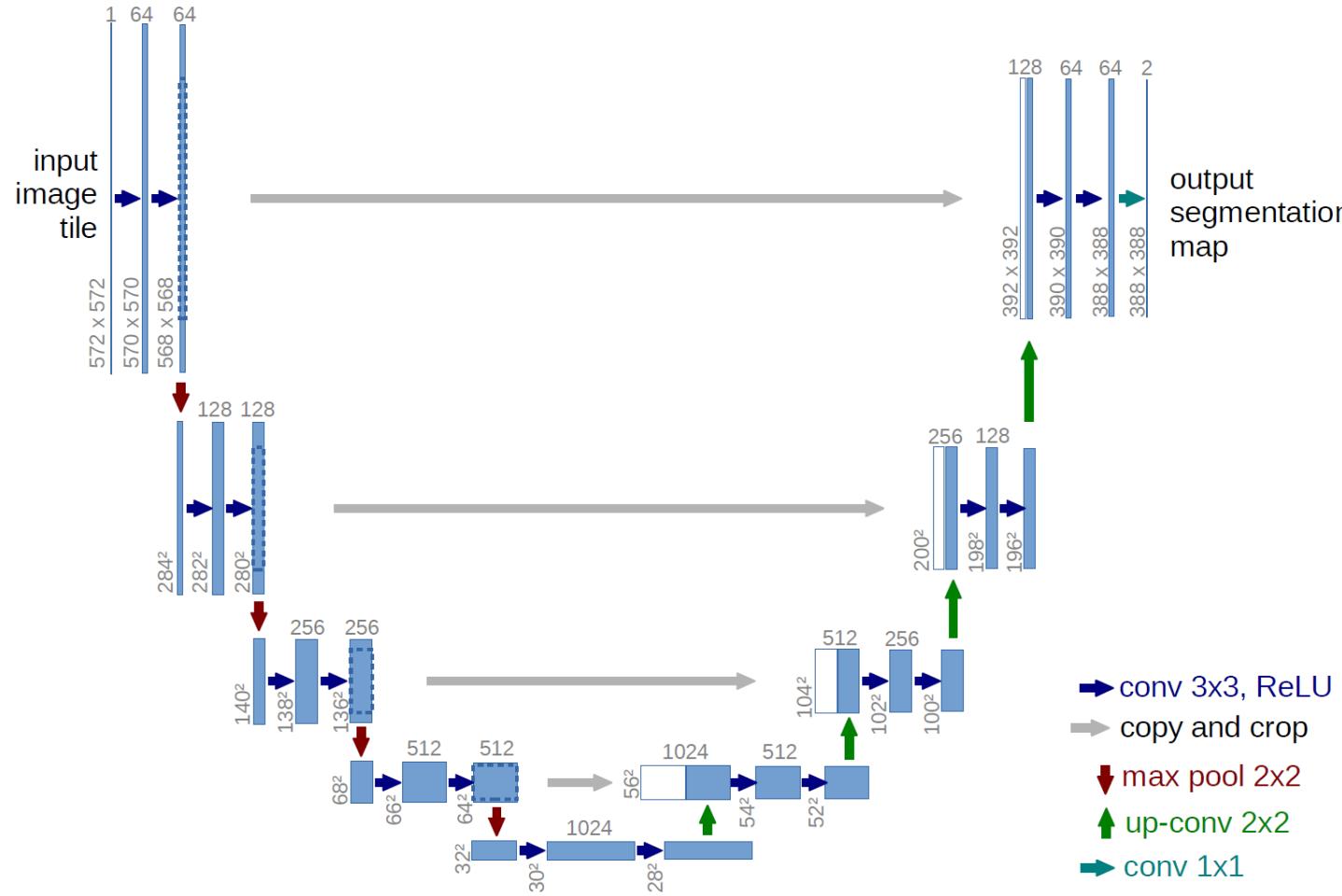
Figures adapted from paper author: <http://people.eecs.berkeley.edu/~jonlong/>

Output Going to Image Resolution

- Encoder aggressively pools and subsamples image
- Necessary to capture context information which is necessary for segmentation
- But spatial detail is also necessary
- Goal for decoder: obtain output at image resolution
- Goal for decoder: recover detail in encoder feature maps before subsampling
- Option 1
 - Use transposed convolution to upsample to image resolution
 - Concatenate encoder features to upsampled features during decoding
- Option 2
 - Similar as above, but more symmetric upsampling to image resolution

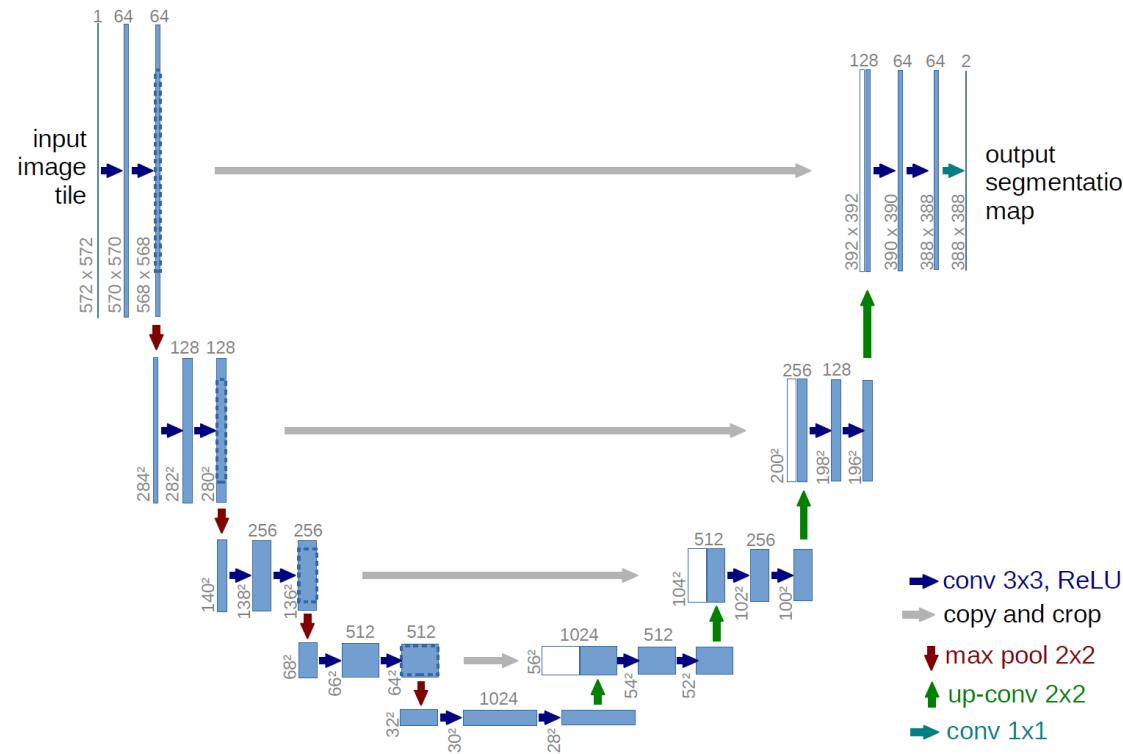
U-Net

- Similar to FCN, but symmetric architecture



U-Net

- Encoder: Modules of two 3×3 convolution, followed by 2×2 pooling
- Decoder:
 - Upsample with 2×2 transposed convolution
 - Concatenate feature map from corresponding encoder level
 - Two 3×3 convolutions



Output Going to Image Resolution

- Encoder aggressively pools and subsamples image
- Necessary to capture context information which is necessary for segmentation
- But spatial detail is also necessary
- Goal for decoder: obtain output at image *resolution*
- Goal for decoder: recover *detail* in encoder feature maps before subsampling
- Option 1
 - Use transposed convolution to upsample to image resolution
 - Concatenate encoder features to upsampled features during decoding
- Option 2
 - Similar as above, but more symmetric upsampling to image resolution
- Option 3
 - Unpool based on encoder locations and convolve to densify

Refinement: Unpooling followed by convolution

Max Pooling

Remember which element was max!

1	2	6	3
3	5	2	1
1	2	2	1
7	3	4	8

Input: 4 x 4

5	6
7	8

Output: 2 x 2

Max Unpooling

Use positions from pooling layer

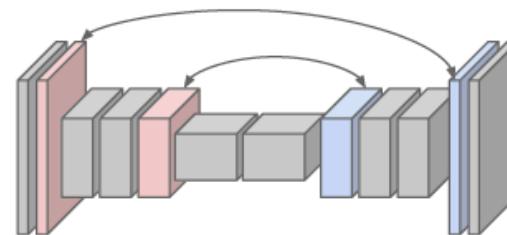
1	2
3	4



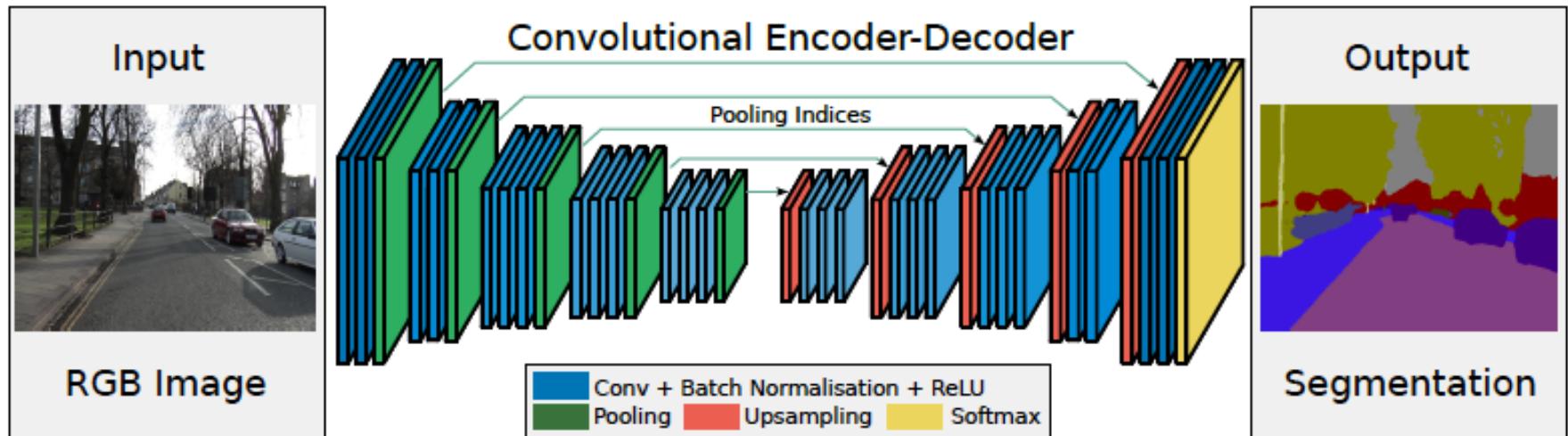
0	0	2	0
0	1	0	0
0	0	0	0
3	0	0	4

Output: 4 x 4

Corresponding pairs of
downsampling and
upsampling layers



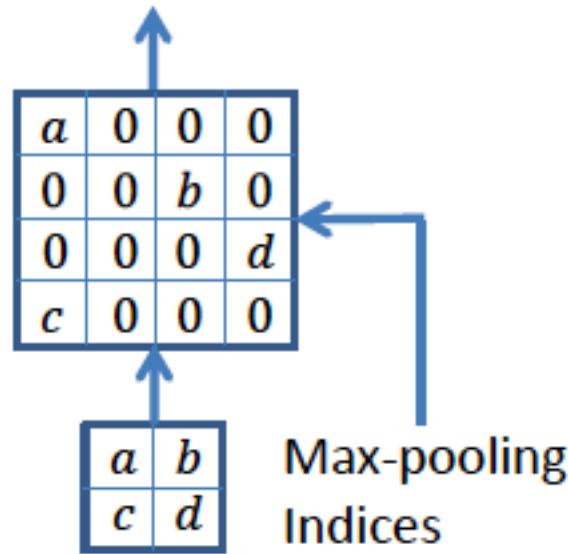
Output Going to Image Resolution



- In U-Net, encoder feature maps need to be stored in memory for concatenation
- In SegNet, only store the indices where max-pool is active
- Store 2x2 pooling index with 2 bits, as opposed to floating point feature map
- Significant memory reduction at inference time

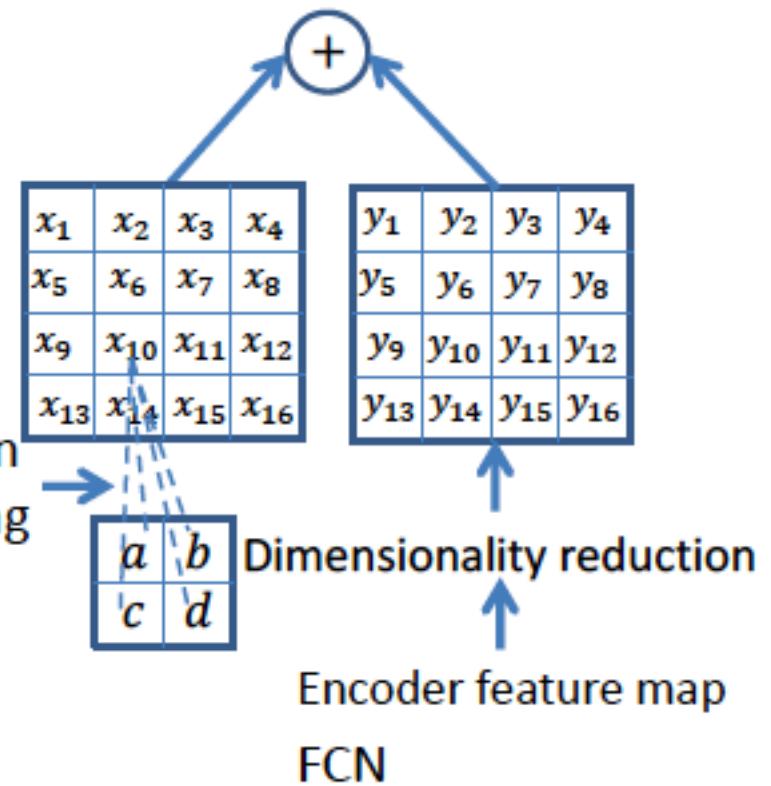
Comparison of two approaches to upsample

Convolution with trainable decoder filters



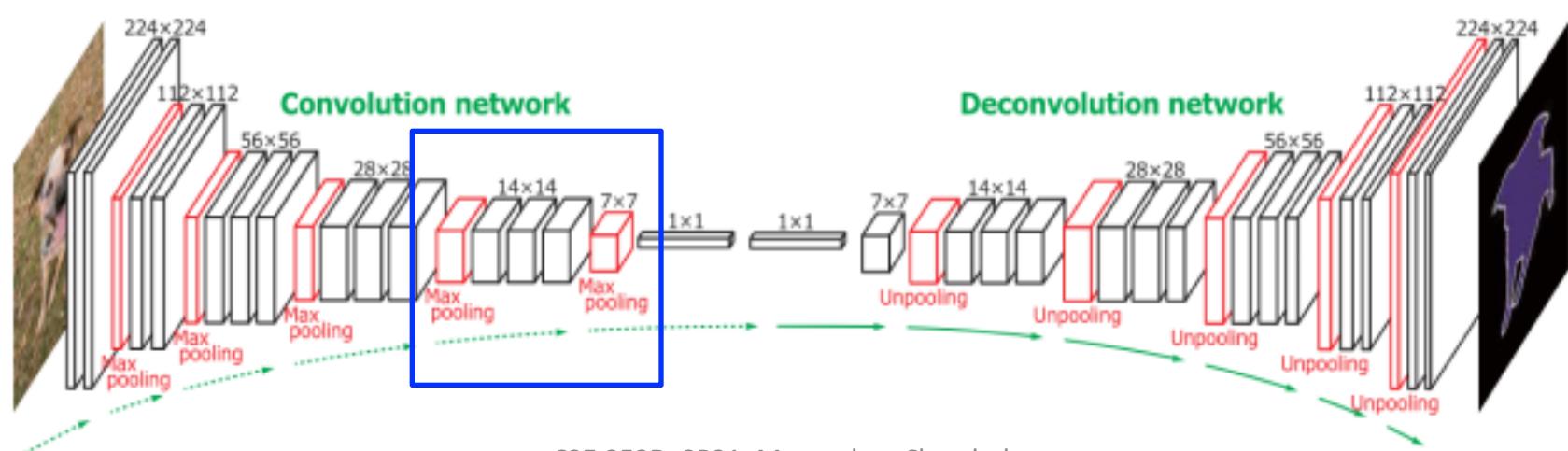
SegNet

Deconvolution
for upsampling



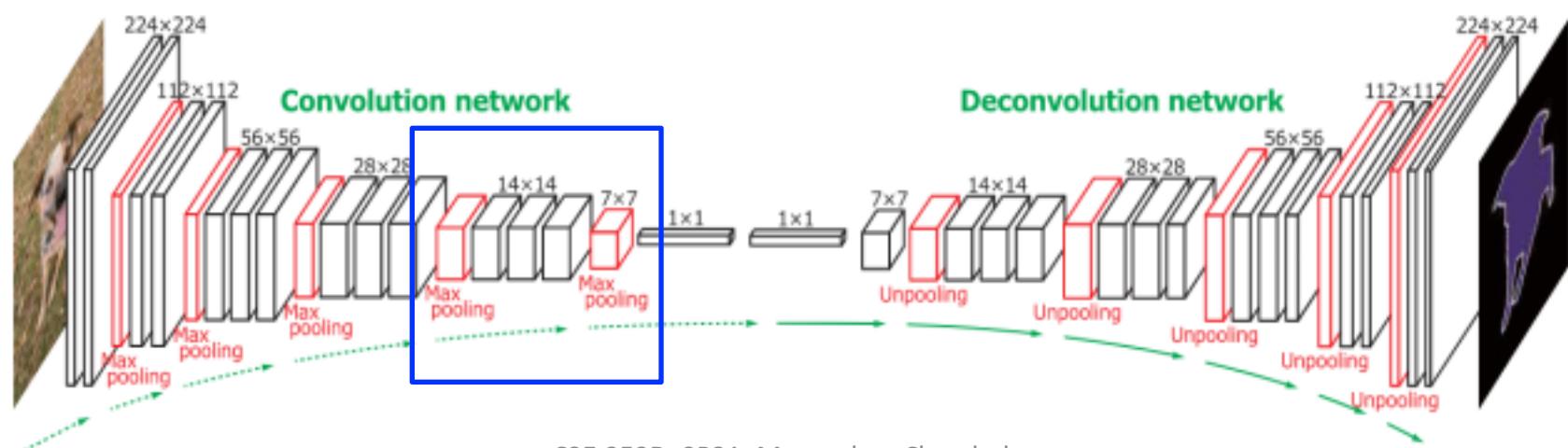
Wide Receptive Fields

- Most networks have similar encoders (inspired by classification networks)
- Downsample to save memory and obtain large receptive fields



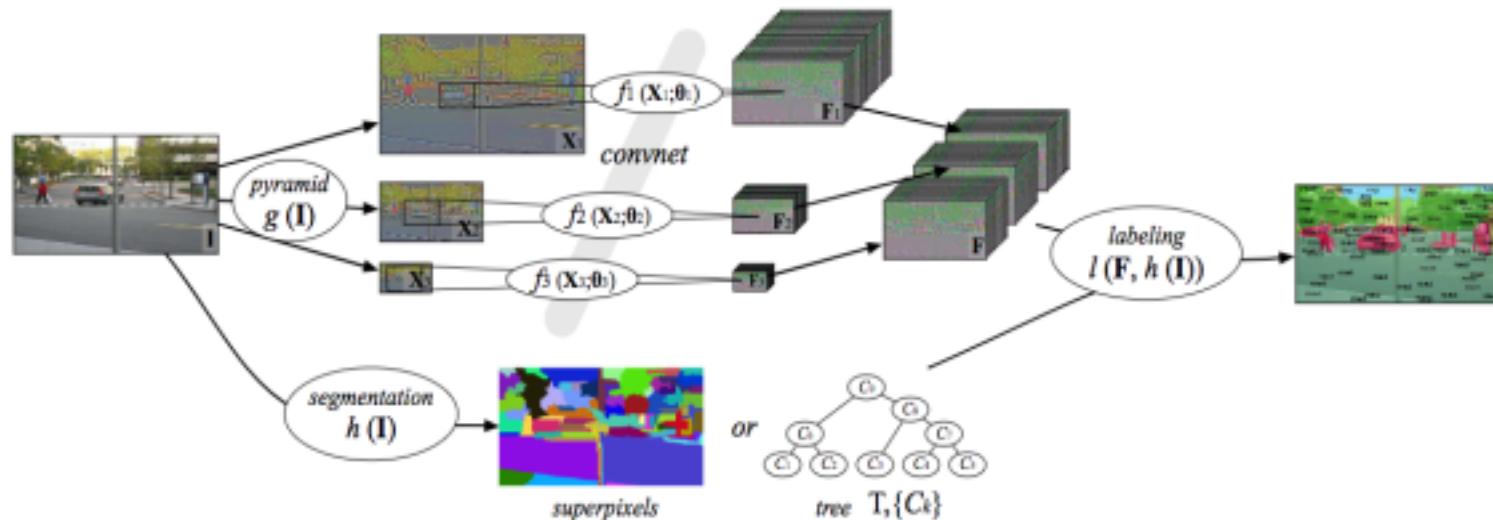
Wide Receptive Fields

- Most networks have similar encoders (inspired by classification networks)
- Downsample to save memory and obtain large receptive fields
- Consider not downsampling features, but still achieve large receptive fields
 - Once downsampled, signal might be lost for small objects
 - Hard to recover by subsequent layers during training



Wide Receptive Fields

- Most networks have similar encoders (inspired by classification networks)
- Downsample to save memory and obtain large receptive fields
- Consider not downsampling features, but still achieve large receptive fields
 - Once downsampled, signal might be lost for small objects
 - Hard to recover by subsequent layers during training
- Option 1:
 - Can process input image at multiple scales and combine the predictions



Wide Receptive Fields

- Most networks have similar encoders (inspired by classification networks)
- Downsample to save memory and obtain large receptive fields
- Consider not downsampling features, but still achieve large receptive fields
 - Once downsampled, signal might be lost for small objects
 - Hard to recover by subsequent layers during training
- Option 1:
 - Can process input image at multiple scales and combine the predictions
- Option 2:
 - Do not let encoder subsample too much in the first place
 - Challenge: maintain spatial resolution along with a large receptive field
 - Solution: use dilated convolutions

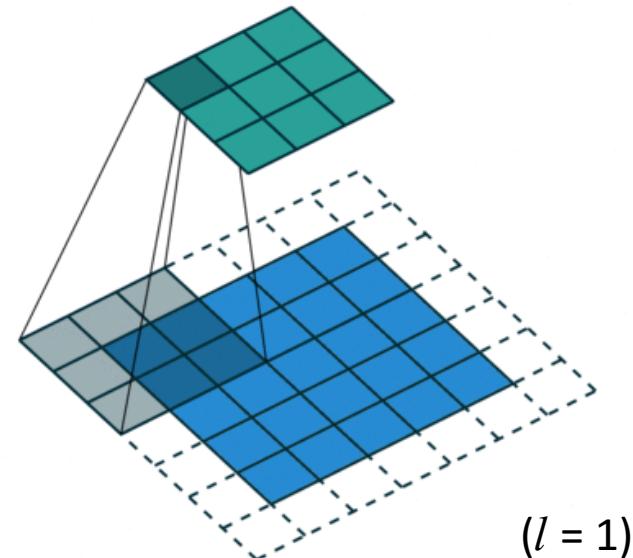
Wide Receptive Fields

- Most networks have similar encoders (inspired by classification networks)
- Downsample to save memory and obtain large receptive fields
- Consider not downsampling features, but still achieve large receptive fields
 - Once downsampled, signal might be lost for small objects
 - Hard to recover by subsequent layers during training
- Option 1:
 - Can process input image at multiple scales and combine the predictions
- Option 2:
 - Do not let encoder subsample too much in the first place
 - Challenge: maintain spatial resolution along with a large receptive field
 - Solution: use dilated convolutions
- Design goals:
 - A module specifically for dense prediction
 - Rectangular prism of convolutional layers, instead of pyramid
 - No pooling or subsampling.

Dilated Convolutions

- Regular convolution:

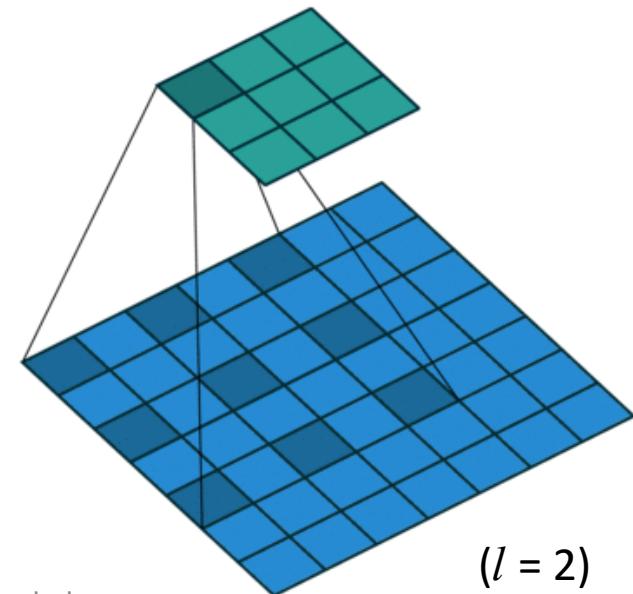
$$(F * k)(\mathbf{p}) = \sum_{\mathbf{s} + \mathbf{t} = \mathbf{p}} F(\mathbf{s}) k(\mathbf{t}).$$



($l = 1$)

- Dilated convolution:

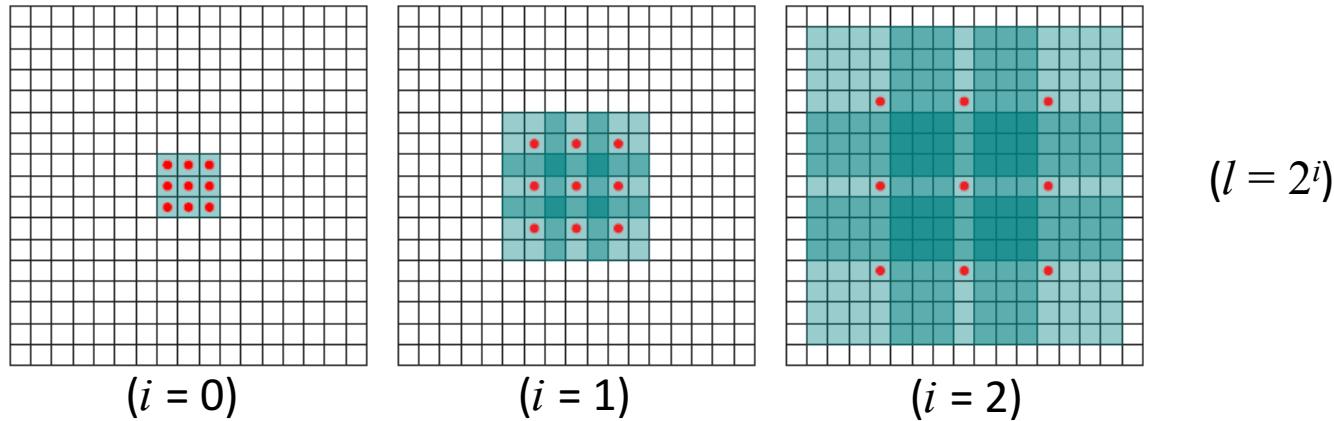
$$(F *_l k)(\mathbf{p}) = \sum_{\mathbf{s} + l\mathbf{t} = \mathbf{p}} F(\mathbf{s}) k(\mathbf{t}).$$



($l = 2$)

Dilated Convolutions

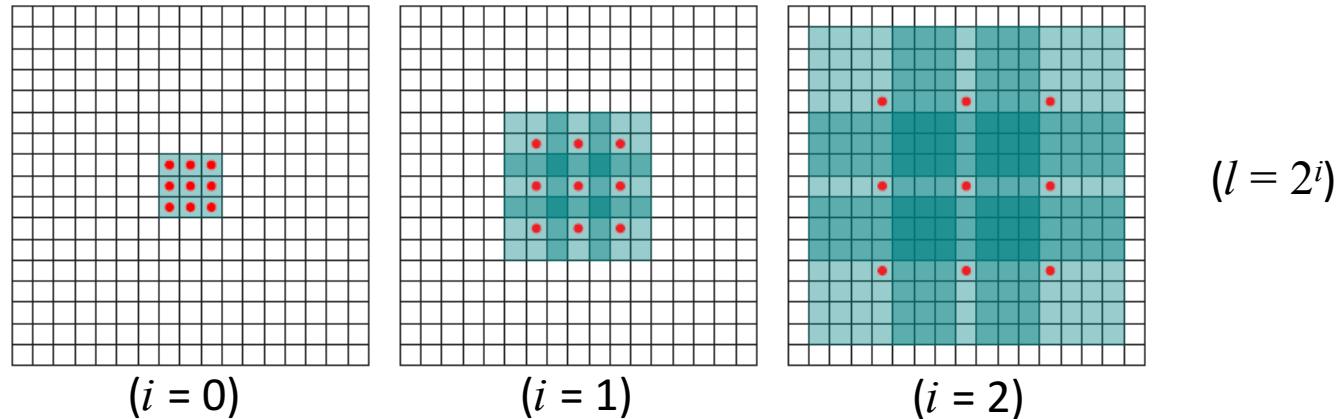
- Receptive field increases with greater dilation factor



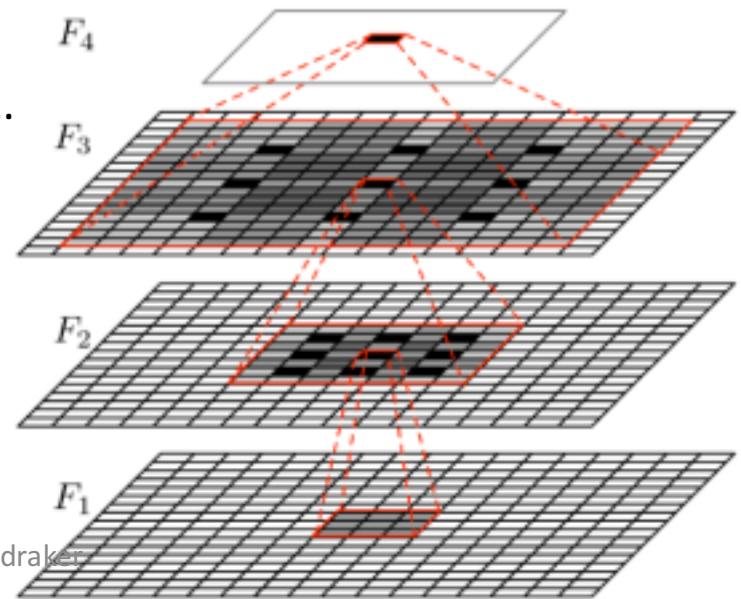
- Number of parameters remains the same

Dilated Convolutions

- Receptive field increases with greater dilation factor



- Number of parameters remains the same
- F1: dilation = 1, F2: dilation = 2, F3: dilation = 4,
- **Exponential RF** on stacking with dilations 1, 2, 4,
$$F_{i+1} = F_i *_{2^i} k_i \quad \text{for } i = 0, 1, \dots, n-2.$$
- (i+1) layer receptive field:
$$(2^{i+2} - 1) \times (2^{i+2} - 1)$$
- **Multiscale and full resolution!**



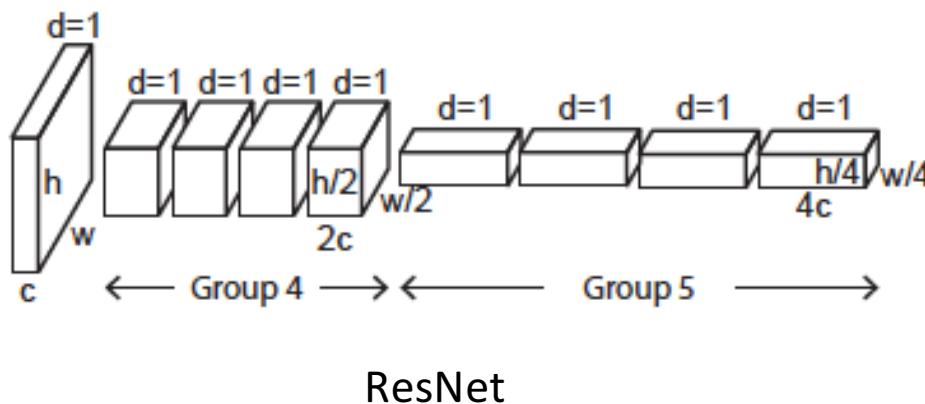
Dilated Context Module

- Goals:
 - Equip feature maps with multiscale context to improve performance
 - Pass through multiple layers that enhance context information in features
 - Use only dilated convolutions to maintain resolution
 - (W, H, C) output feature for (W, H, C) input, can plug into any layer

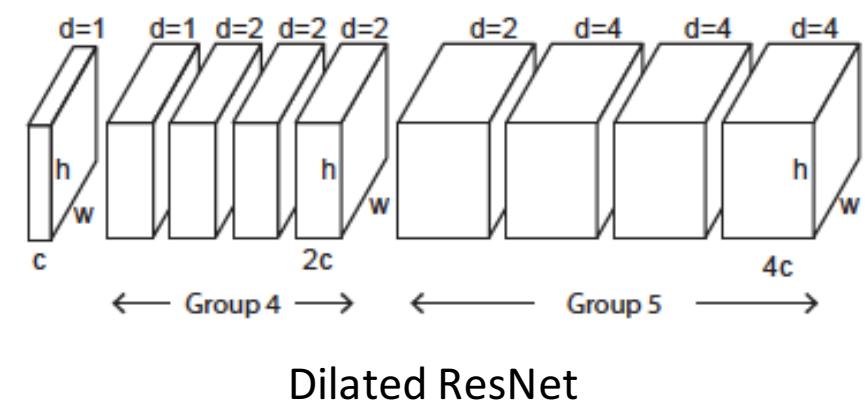
Layer	1	2	3	4	5	6	7	8
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67
Output channels								
Basic	C	C	C	C	C	C	C	C
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	C

Dilated Residual Network

- Goals: Exploit the power of residual networks with advantages of dilation
 - Once training signal lost by downsampling, hard to recover it
 - Preserve spatial resolution of feature maps
 - Provide training signals that densely cover the input field
 - Backpropagation can now learn to preserve small but salient details
 - Also beneficial when classification network is transferred to other tasks



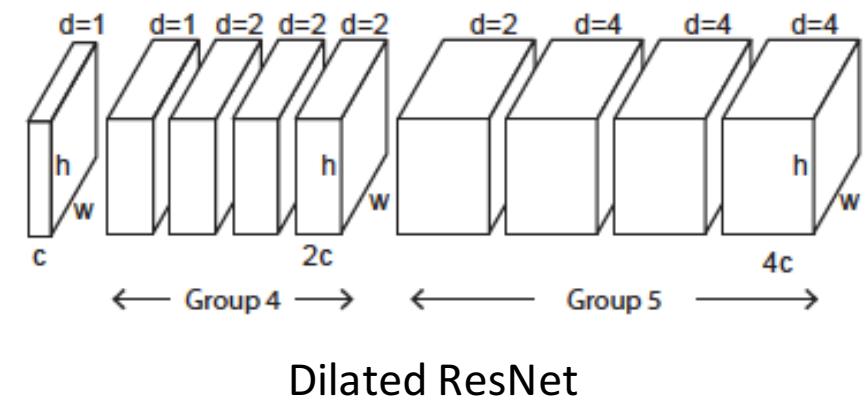
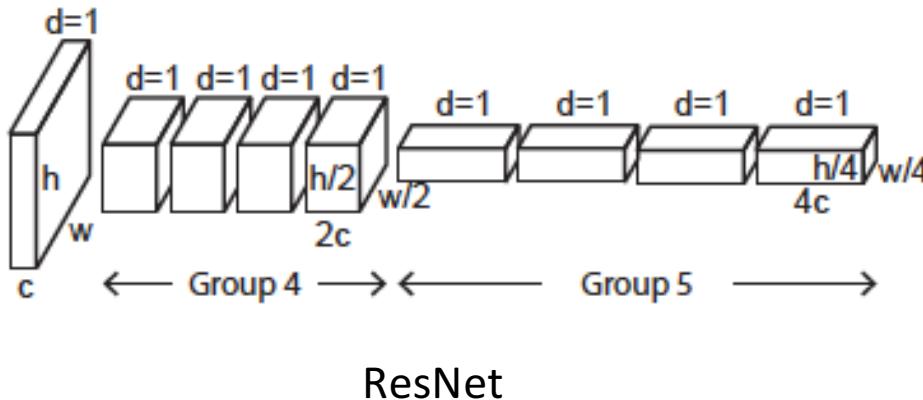
CSE 252D, SP21: Manmohan Chandraker



[Yu et al., Dilated Residual Networks]

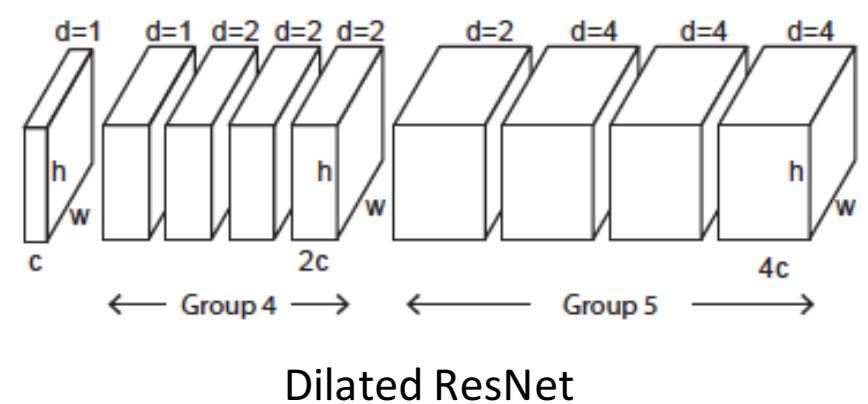
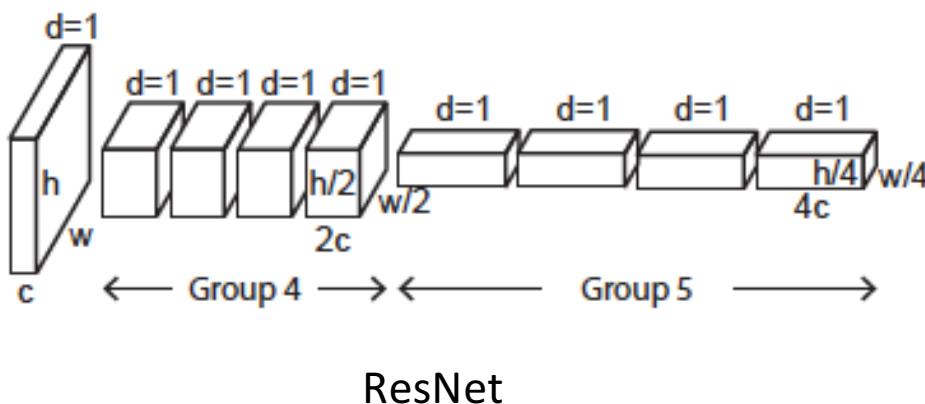
Dilated Residual Network

- Goals: Exploit the power of residual networks with advantages of dilation
 - Once training signal lost by downsampling, hard to recover it
 - Preserve spatial resolution of feature maps
 - Provide training signals that densely cover the input field
 - Backpropagation can now learn to preserve small but salient details
 - Also beneficial when classification network is transferred to other tasks
- ResNet uses stride 2 to downsample from block G_k to G_{k+1}
- Do not use stride to downsample, maintain resolution
- Dilate subsequent layers by another factor of 2 to maintain receptive field



Dilated Residual Network

- State-of-the-art performance on classification, segmentation
- High resolution features, classification network easily used for segmentation
- But two problems with DRN:
 - Gridding artifacts
 - Memory consumption



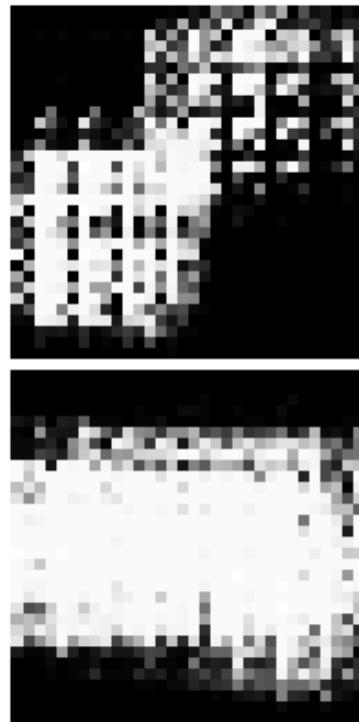
Dilated Residual Network: Degridding

- Gridding artifacts sometimes observed with DRN



(a) Input

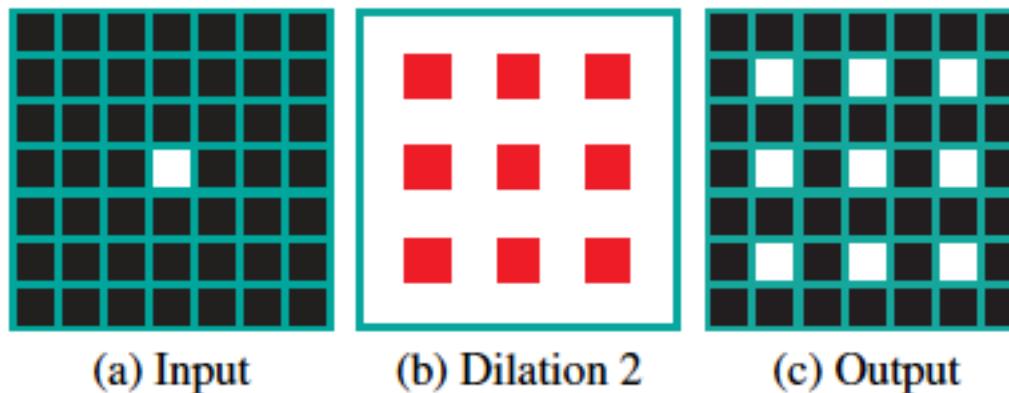
(b) ResNet-18



(c) DRN-A-18

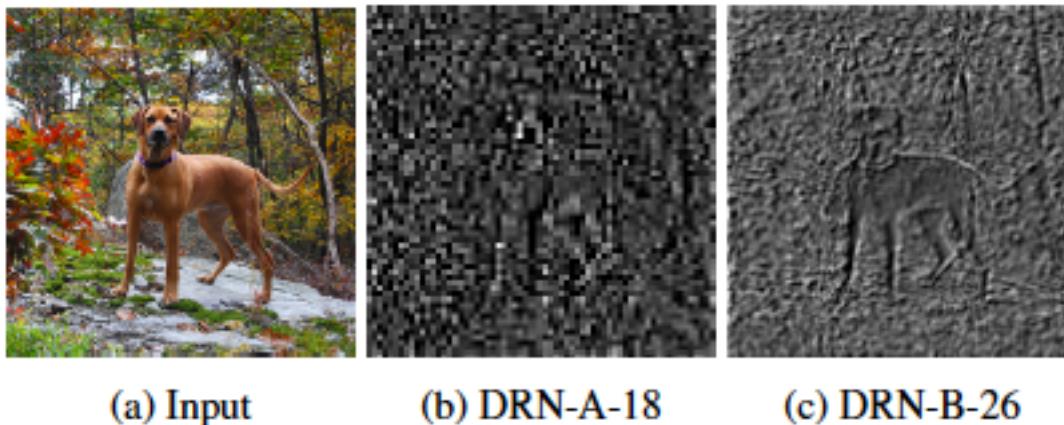
Dilated Residual Network: Degridding

- Gridding artifacts sometimes observed with DRN
 - Frequency in features maps exceeds sampling rate of filter
 - Filters with higher dilation are at “lower frequency”



Dilated Residual Network: Degridding

- Degridding solutions
 - Replace max-pooling in preceding layer with convolutional layers



(a) Input

(b) DRN-A-18

(c) DRN-B-26



Dilated Residual Network: Degridding

- Degridding solutions
 - Replace max-pooling in preceding layer with convolutional layers
 - Add convolution layers at end of network, with progressively lower dilation



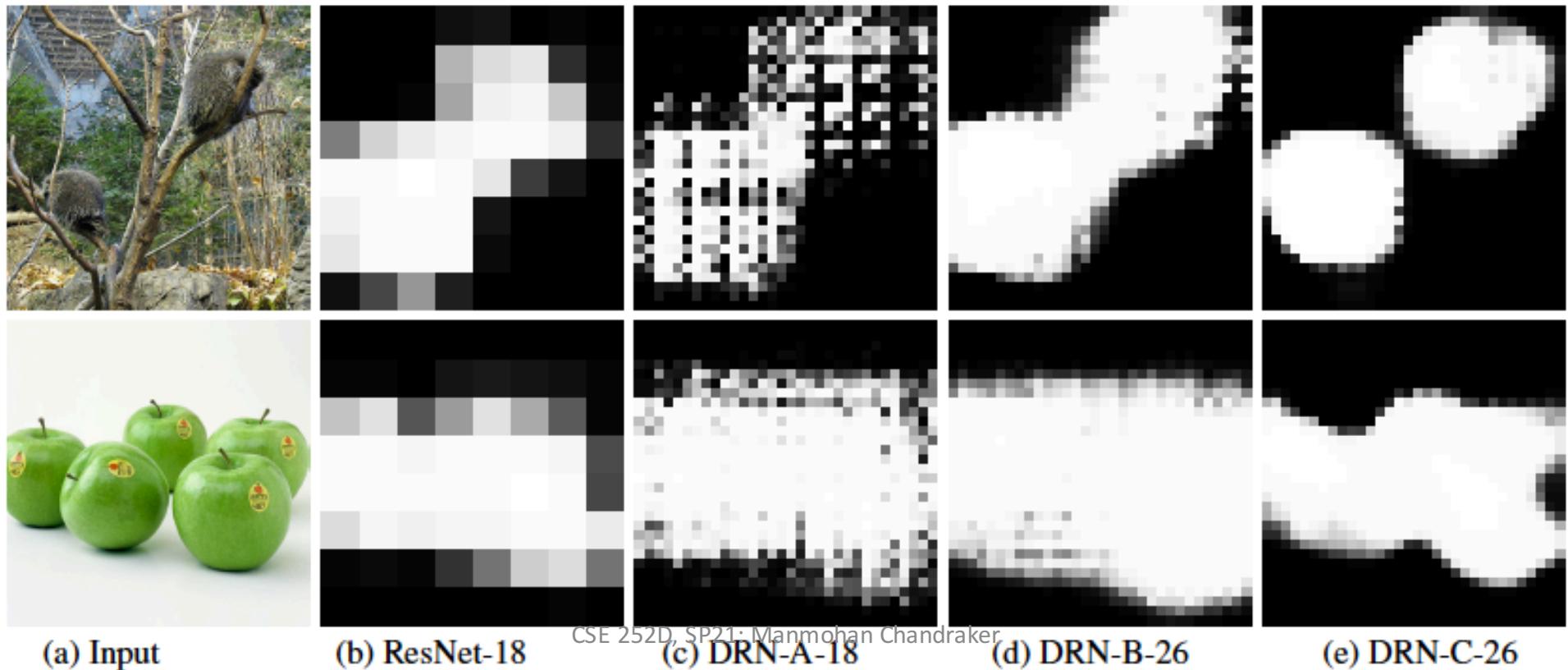
Dilated Residual Network: Degridding

- Degridding solutions
 - Replace max-pooling in preceding layer with convolutional layers
 - Add convolution layers at end of network, with progressively lower dilation
 - Remove skip connections in newly added layers
 - Skip connections can propagate gridding artifacts from previous layers



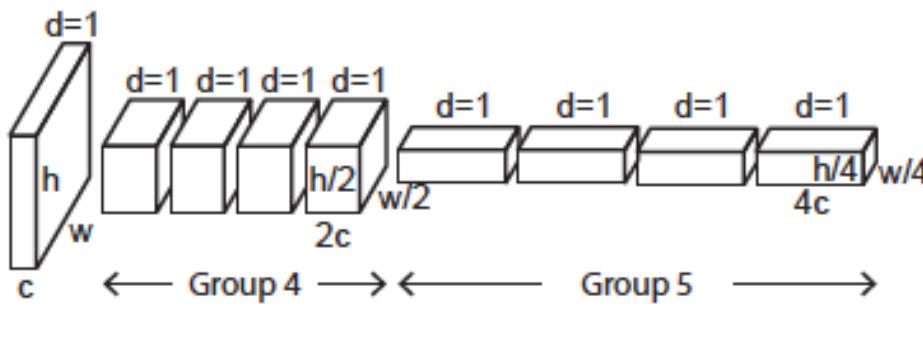
Dilated Residual Network: Degridding

- Degridding solutions
 - Replace max-pooling in preceding layer with convolutional layers
 - Add convolution layers at end of network, with progressively lower dilation
 - Remove skip connections in newly added layers
 - Skip connections can propagate gridding artifacts from previous layers



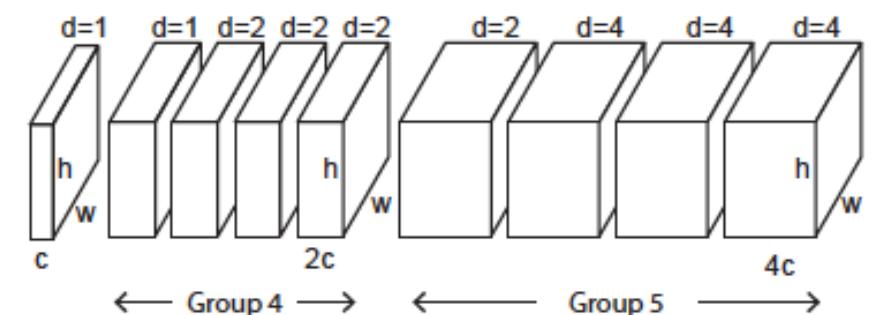
Dilated Residual Network : Memory

- State-of-the-art performance on classification, segmentation
- High resolution features, classification network easily used for segmentation
- But two problems with DRN:
 - Gridding artifacts
 - Memory consumption
- Dilation preserves number of parameters, but feature maps are bigger
- In practice, only G_4 and G_5 use dilation
- Output produced at 8-times smaller resolution



ResNet

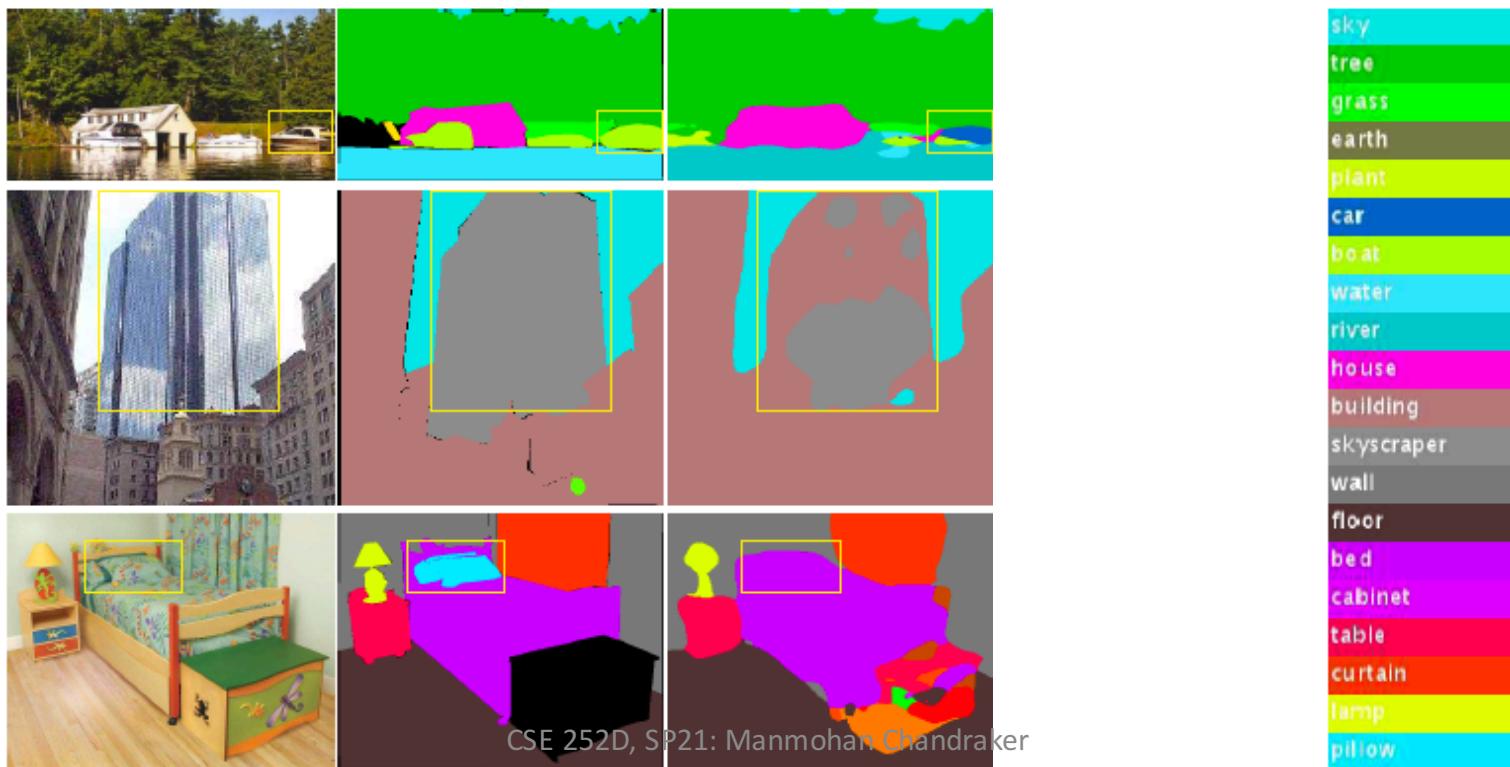
CSE 252D, SP21: Manmohan Chandraker



Dilated ResNet

Importance of Context: Issue with FCN

- Use of co-occurring visual statistics is crucial for semantic segmentation
 - Mismatched classes: predict car on water
 - Confusion classes: predict same object as skyscraper and building
 - Inconspicuous classes: miss the pillow since cannot correlate with bed
- It seems not enough context information is being learned

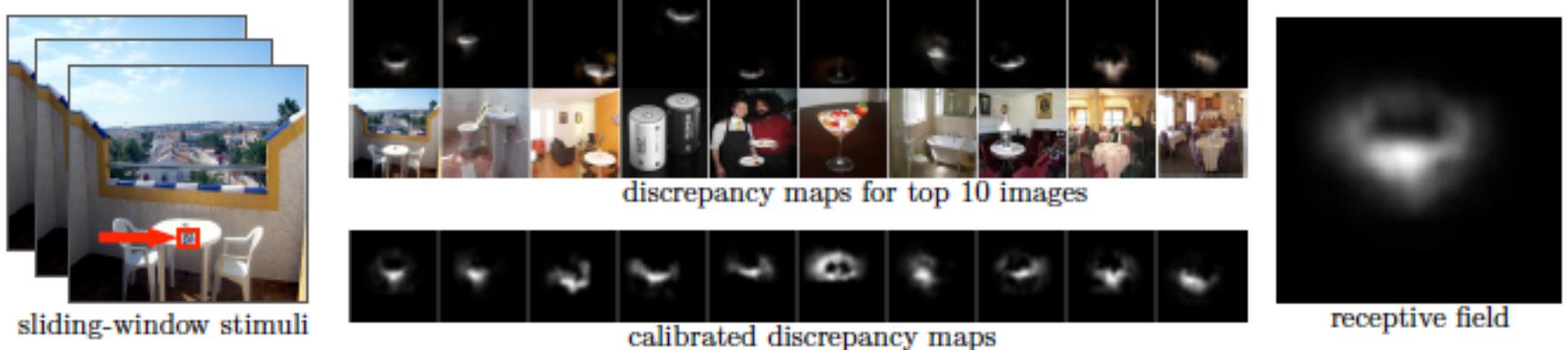


More Context Information

- But a deep ResNet should encode sufficient context
 - Theoretically, ResNet has receptive field that covers entire image
 - Empirical size of receptive field can be much smaller

More Context Information

- But a deep ResNet should encode sufficient context
 - Theoretically, ResNet has receptive field that covers entire image
 - Empirical size of receptive field can be much smaller
- A data-driven way to determine empirical receptive field for a unit
 - Consider top K images that cause maximum activation for a unit
 - Slide a small occluder on each of the K images
 - A region is important if there is significant change in the activation
 - Empirical receptive field: average of discrepancy maps for the K images

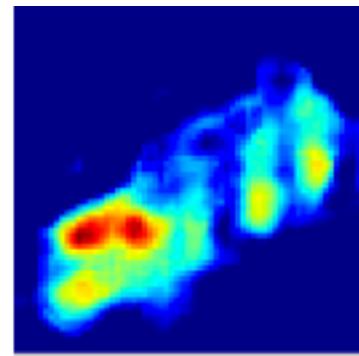


More Context Information

- But a deep ResNet should encode sufficient context
 - Theoretically, ResNet has receptive field that covers entire image
 - Empirical size of receptive field can be much smaller
- A data-driven way to determine empirical receptive field for a unit
 - Consider top K images that cause maximum activation for a unit
 - Slide a small occluder on each of the K images
 - A region is important if there is significant change in the activation
 - Empirical receptive field: average of discrepancy maps for the K images



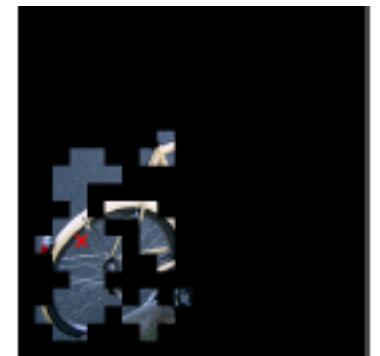
(a) Original Image



(b) Activation map



(c) Theoretical RF



(d) Empirical RF

[Zhou et al., ICLR 2015]

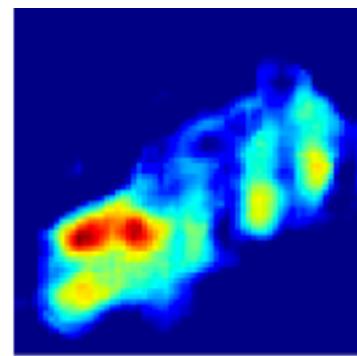
More Context Information

- But a deep ResNet should encode sufficient context
 - Theoretically, ResNet has receptive field that covers entire image
 - Empirical size of receptive field can be much smaller
- A data-driven way to determine empirical receptive field for a unit
 - Consider top K images that cause maximum activation for a unit
 - Slide a small occluder on each of the K images
 - A region is important if there is significant change in the activation
 - Empirical receptive field: average of discrepancy maps for the K images

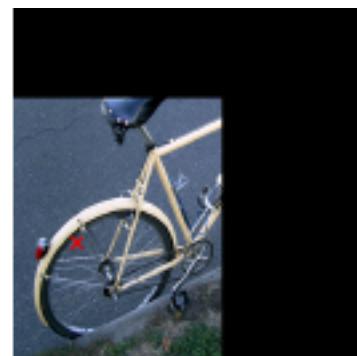
	pool1	pool2	conv3	conv4	pool5
Theoretic size	19	67	99	131	195
ImageNet-CNN actual size	17.9 ± 1.6	36.7 ± 5.4	51.1 ± 9.9	60.4 ± 16.0	70.3 ± 21.6



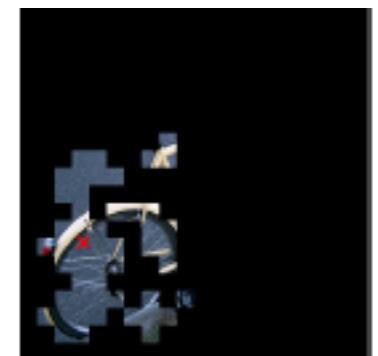
(a) Original Image



(b) Activation map



(c) Theoretical RF



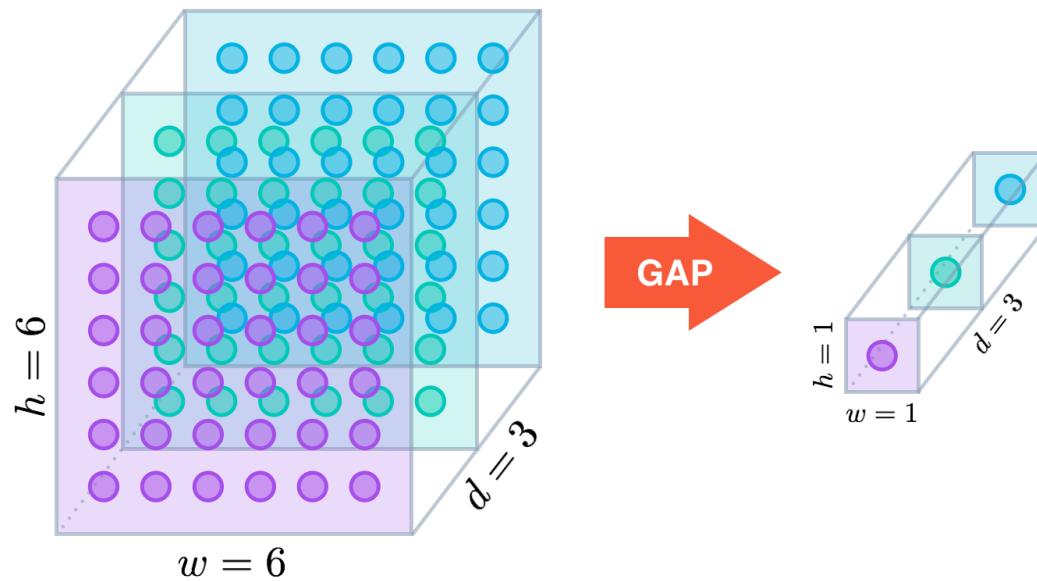
(d) Empirical RF

CSE 252D, SP21: Manmohan Chandraker

[Zhou et al., ICLR 2015]

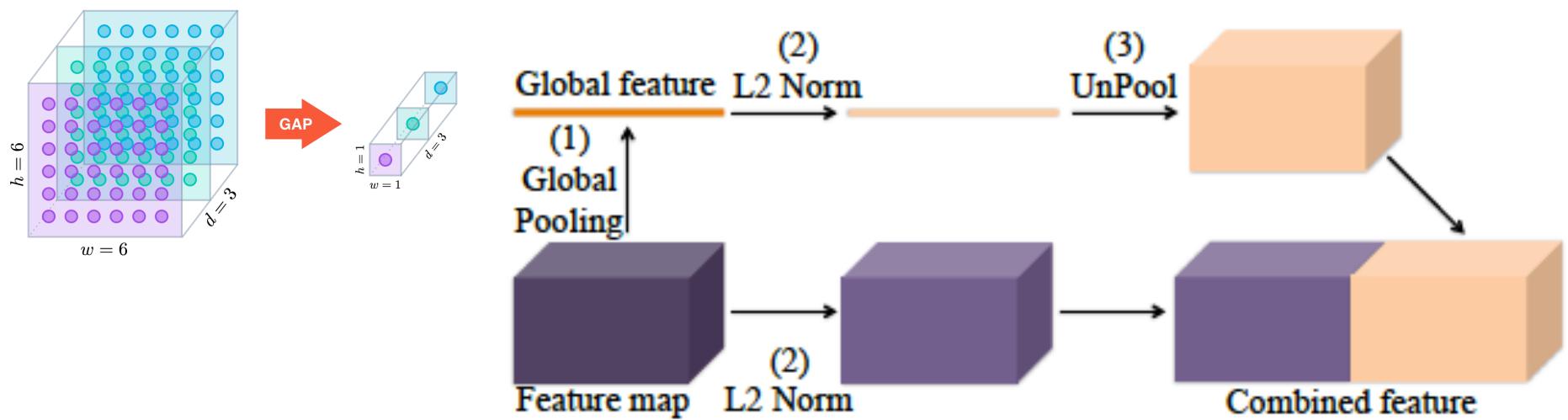
More Context Information : Global Pooling

- Need a mechanism to explicitly encode context information
- Option 1: Global average pooling



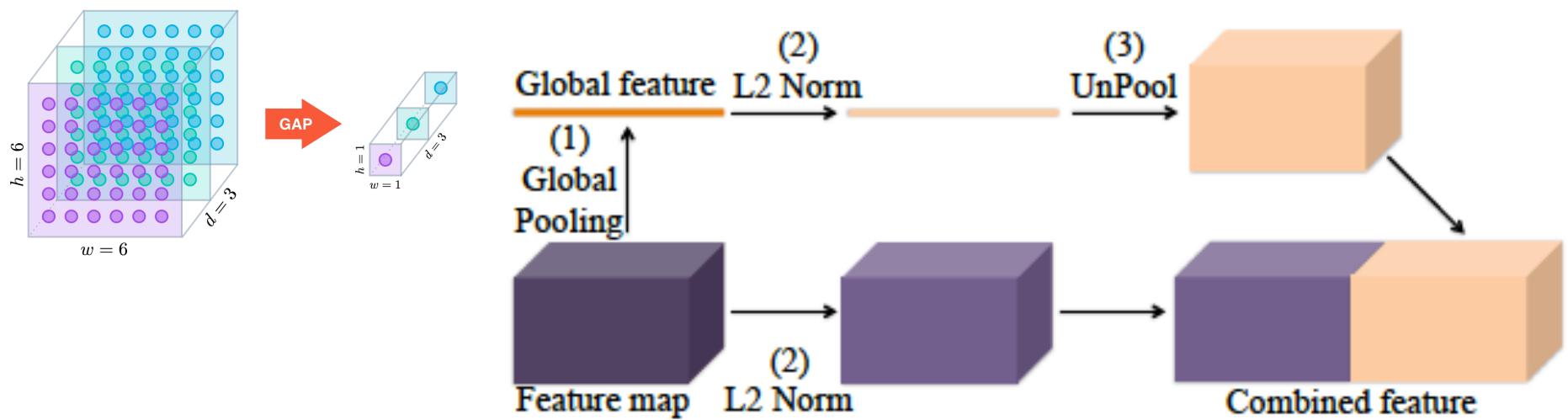
More Context Information: Global Pooling

- Need a mechanism to explicitly encode context information
- Option 1: Global average pooling
 - Obtain context vector from feature maps in a layer
 - Unpool (replicate) to feature dimensions and concatenate
 - L2-normalization before concatenation for stable training



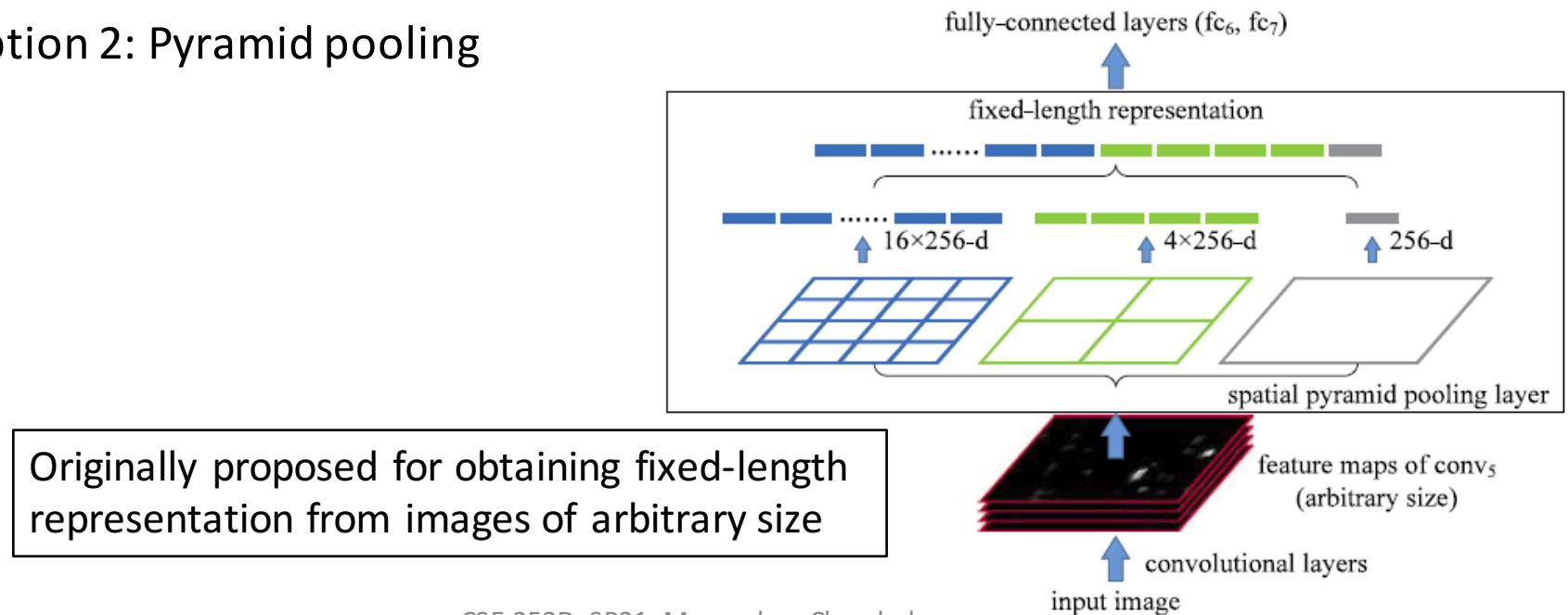
More Context Information: Global Pooling

- Need a mechanism to explicitly encode context information
- Option 1: Global average pooling
 - Obtain context vector from feature maps in a layer
 - Unpool (replicate) to feature dimensions and concatenate
 - L2-normalization before concatenation for stable training
 - Issue: spatial relationships may be lost in global pooling



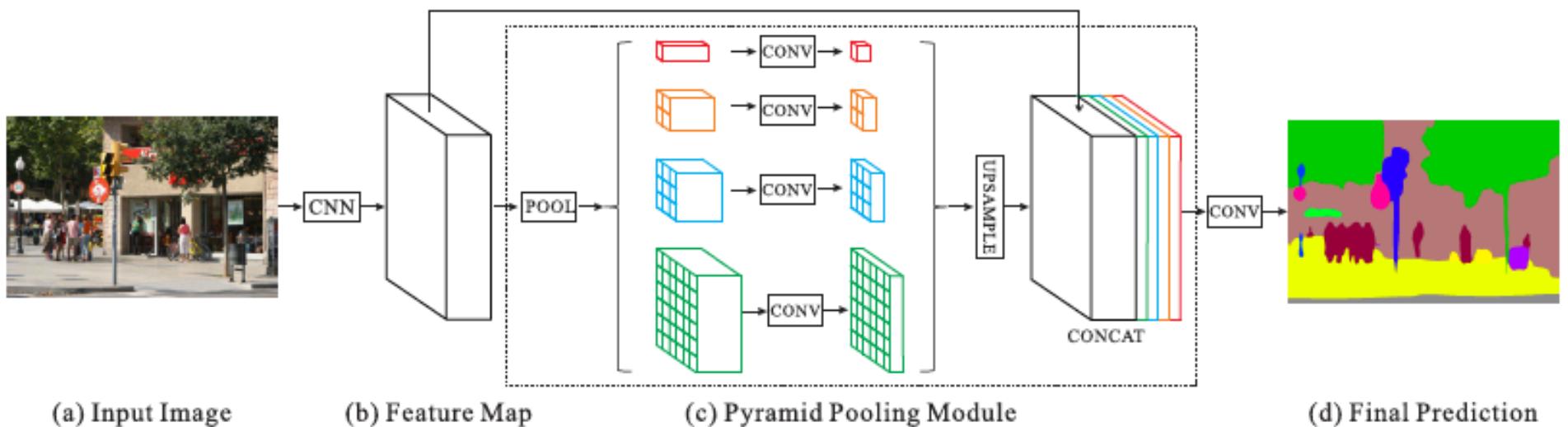
More Context Information: Pyramid Pooling

- Need a mechanism to explicitly encode context information
- Option 1: Global average pooling
 - Obtain context vector from feature maps in a layer
 - Unpool (replicate) to feature dimensions and concatenate
 - L2-normalization before concatenation for stable training
 - Issue: spatial relationships may be lost in global pooling
- Option 2: Pyramid pooling



More Context Information: Pyramid Pooling

- Role in segmentation: Extract both global and regional context
 - A hierarchical global prior, with information across scales and regions
- Fuse features in several different pyramid scales
 - Pool input feature map into hierarchy of context features
 - Do 1x1 convolution to ensure each pyramid level gets equal weight
 - Upsample to original resolution and concatenate



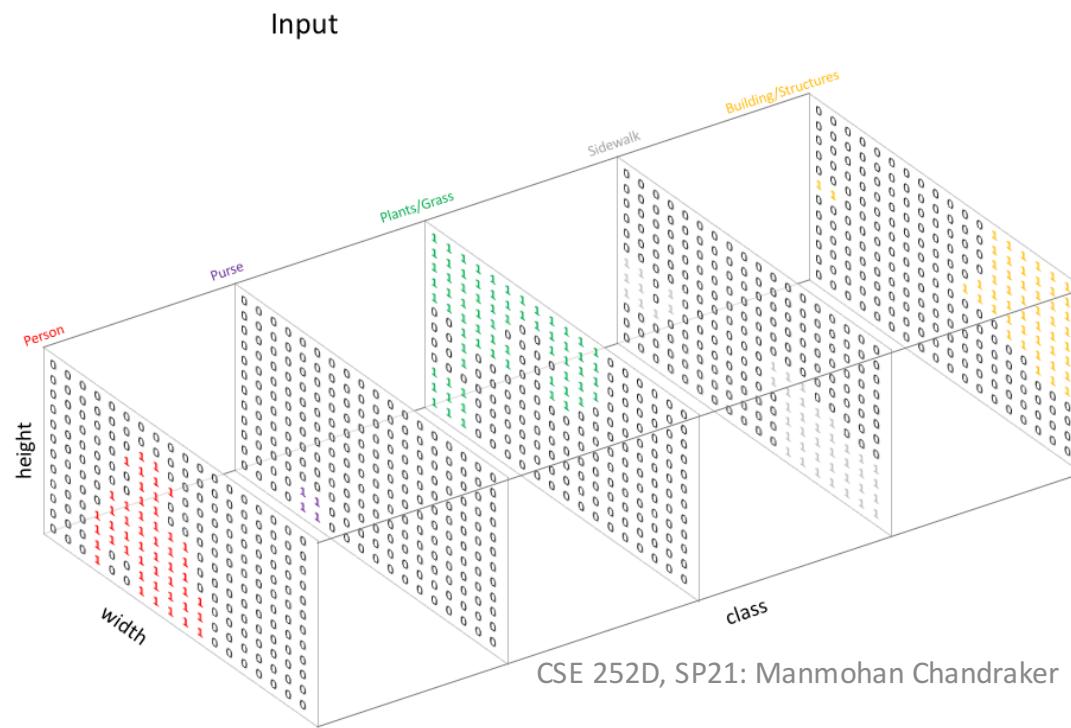
Dense Multiclass Prediction



segmented →

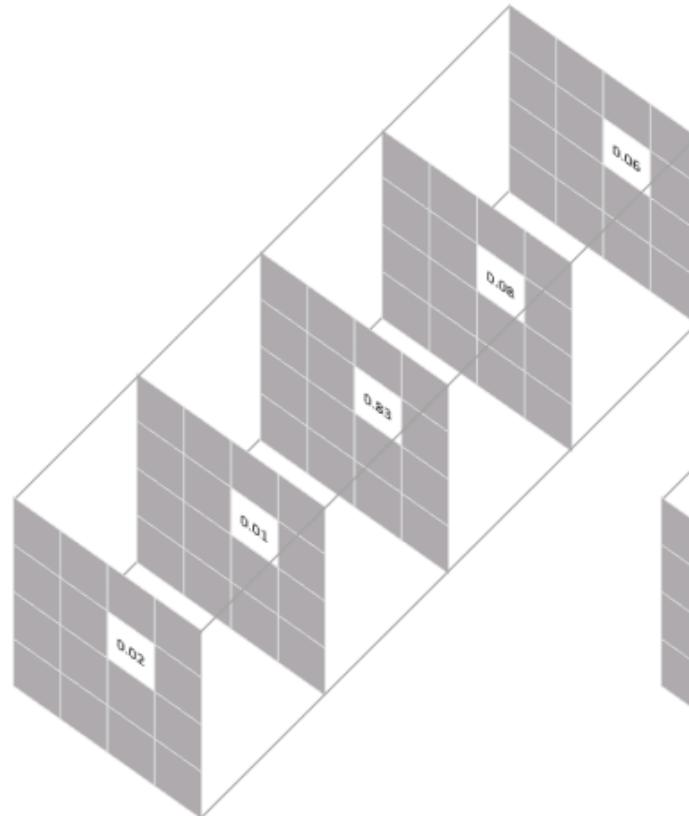
- 1: Person
- 2: Purse
- 3: Plants/Grass
- 4: Sidewalk
- 5: Building/Structures

3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5
3	3	3	3	3	3	3	3	3	3	1	1	3	3	3	3	5	5	5	5	5
3	3	3	3	3	3	3	3	3	1	1	1	1	3	3	3	5	5	5	5	5
3	3	3	3	3	3	3	3	3	1	1	1	3	3	3	3	5	5	5	5	5
5	5	3	3	3	3	3	3	1	1	3	3	3	5	5	5	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	1	1	1	4	4	4	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	1	1	1	4	4	4	4	5	5	5	5
4	4	4	1	1	1	1	1	1	1	1	1	1	4	4	4	4	4	4	4	4
3	3	3	1	1	1	1	1	1	1	1	1	1	4	4	4	4	4	4	4	4
3	3	3	1	2	2	1	1	1	1	1	1	1	4	4	4	4	4	4	4	4
3	3	3	1	2	2	1	1	1	1	1	1	1	4	4	4	4	4	4	4	4

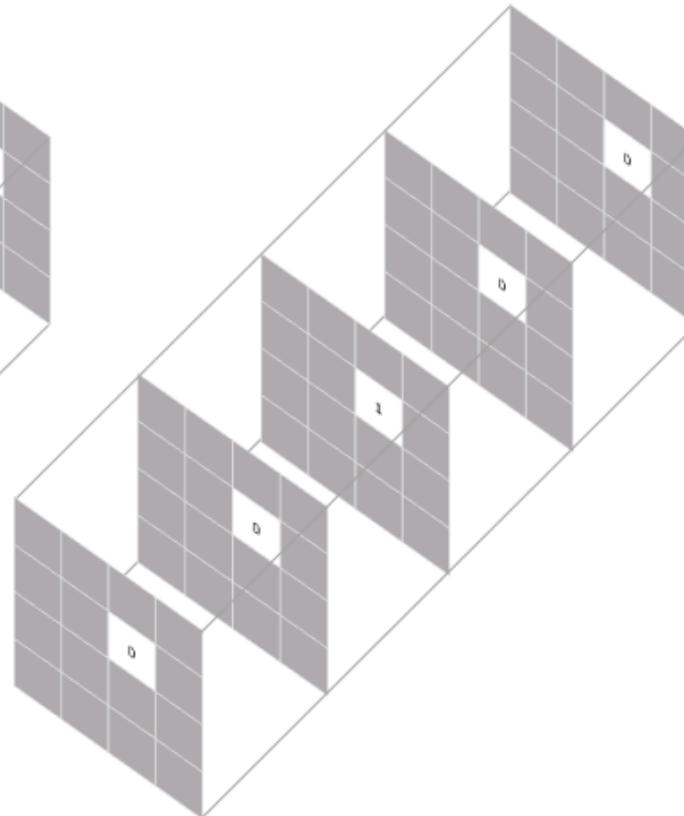


[Images from: Jeremy Jordan]

Pixel-wise Cross-Entropy Loss



Prediction for a selected pixel



Target for the corresponding pixel

Pixel-wise loss is calculated as the log loss, summed over all possible classes

$$-\sum_{\text{classes}} y_{true} \log(y_{pred})$$

This scoring is repeated over all pixels and averaged

- Can weight each output channel for class imbalance in training set (FCN)
- Can assign higher weight to pixels near the boundary (U-Net)

Semantic Segmentation Metrics

Confusion matrix: $\mathbf{C}_{ij} = \sum_{I \in \mathcal{D}} |\{z \in I \text{ such that } S_{gt}^I(z) = i \text{ and } S_{ps}^I(z) = j\}|$

Number of pixels with ground truth label i : $\mathbf{G}_i = \sum_{j=1}^L \mathbf{C}_{ij}$

Number of pixels with prediction j : $\mathbf{P}_j = \sum_i \mathbf{C}_{ij}$

Overall pixel accuracy: $OP = \frac{\sum_{i=1}^L \mathbf{C}_{ii}}{\sum_{i=1}^L \mathbf{G}_i}$

Per-class accuracy: $PC = \frac{1}{L} \sum_{i=1}^L \frac{\mathbf{C}_{ii}}{\mathbf{G}_i}$

Intersection-over-union: $JU = \frac{1}{L} \sum_{i=1}^L \frac{\mathbf{C}_{ii}}{\mathbf{G}_i + \mathbf{P}_i - \mathbf{C}_{ii}}$

Semantic Segmentation Metrics

Confusion matrix: $\mathbf{C}_{ij} = \sum_{I \in \mathcal{D}} |\{z \in I \text{ such that } S_{gt}^I(z) = i \text{ and } S_{ps}^I(z) = j\}|$ $\begin{bmatrix} 6 & 4 \\ 2 & 18 \end{bmatrix}$

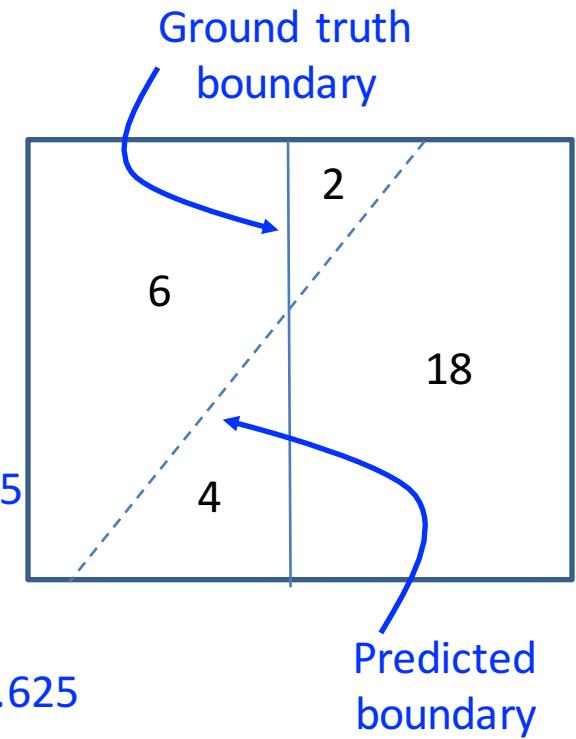
Number of pixels with ground truth label i : $\mathbf{G}_i = \sum_{j=1}^L \mathbf{C}_{ij}$
[10, 20]

Number of pixels with prediction j : $\mathbf{P}_j = \sum_i \mathbf{C}_{ij}$
[8, 22]

Overall pixel accuracy: $OP = \frac{\sum_{i=1}^L \mathbf{C}_{ii}}{\sum_{i=1}^L \mathbf{G}_i} \frac{6 + 18}{10 + 20} = 0.8$

Per-class accuracy: $PC = \frac{1}{L} \sum_{i=1}^L \frac{\mathbf{C}_{ii}}{\mathbf{G}_i} \frac{1}{2} \left(\frac{6}{10} + \frac{18}{20} \right) = 0.75$

Intersection-over-union: $J_I = \frac{1}{L} \sum_{i=1}^L \frac{\mathbf{C}_{ii}}{\mathbf{G}_i + \mathbf{P}_i - \mathbf{C}_{ii}} = 0.625$



Semantic Segmentation Metrics

Confusion matrix: $\mathbf{C}_{ij} = \sum_{I \in \mathcal{D}} |\{z \in I \text{ such that } S_{gt}^I(z) = i \text{ and } S_{ps}^I(z) = j\}|$

Number of pixels with ground truth label i : $\mathbf{G}_i = \sum_{j=1}^L \mathbf{C}_{ij}$

Number of pixels with prediction j : $\mathbf{P}_j = \sum_i \mathbf{C}_{ij}$

Overall pixel accuracy: $OP = \frac{\sum_{i=1}^L \mathbf{C}_{ii}}{\sum_{i=1}^L \mathbf{G}_i}$

Biased in favor of large classes

Per-class accuracy: $PC = \frac{1}{L} \sum_{i=1}^L \frac{\mathbf{C}_{ii}}{\mathbf{G}_i}$

Biased against background classes,
boost all foreground classes

Intersection-over-union: $JU = \frac{1}{L} \sum_{i=1}^L \frac{\mathbf{C}_{ii}}{\mathbf{G}_i + \mathbf{P}_i - \mathbf{C}_{ii}}$

Balances false positives
and false negatives

FCN Semantic Segmentation Results

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [16]	52.6	51.6	~ 50 s
FCN-8s	62.7	62.2	~ 175 ms

Relative to prior state-of-the-art SDS:

- 30% relative improvement for mean IoU
- 286 times faster

