Deep Scene Understanding from Images for Monitoring Applications

Matteo Poggi, Fabio Tosi, **Pierluigi Zama Ramirez**
Computer Vision Lab (CVLab), University of Bologna

Computer Vision Laboratory

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

2 – Semantic Segmentation

Image Classification

Input

Output
Choose among
these categories

Dog
Cat
Bird
Frog
Person

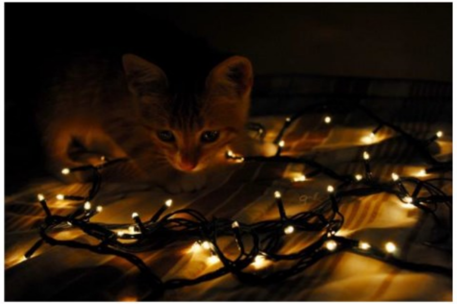
Some challenges

Intraclass variations

Background clutter

Occlusions

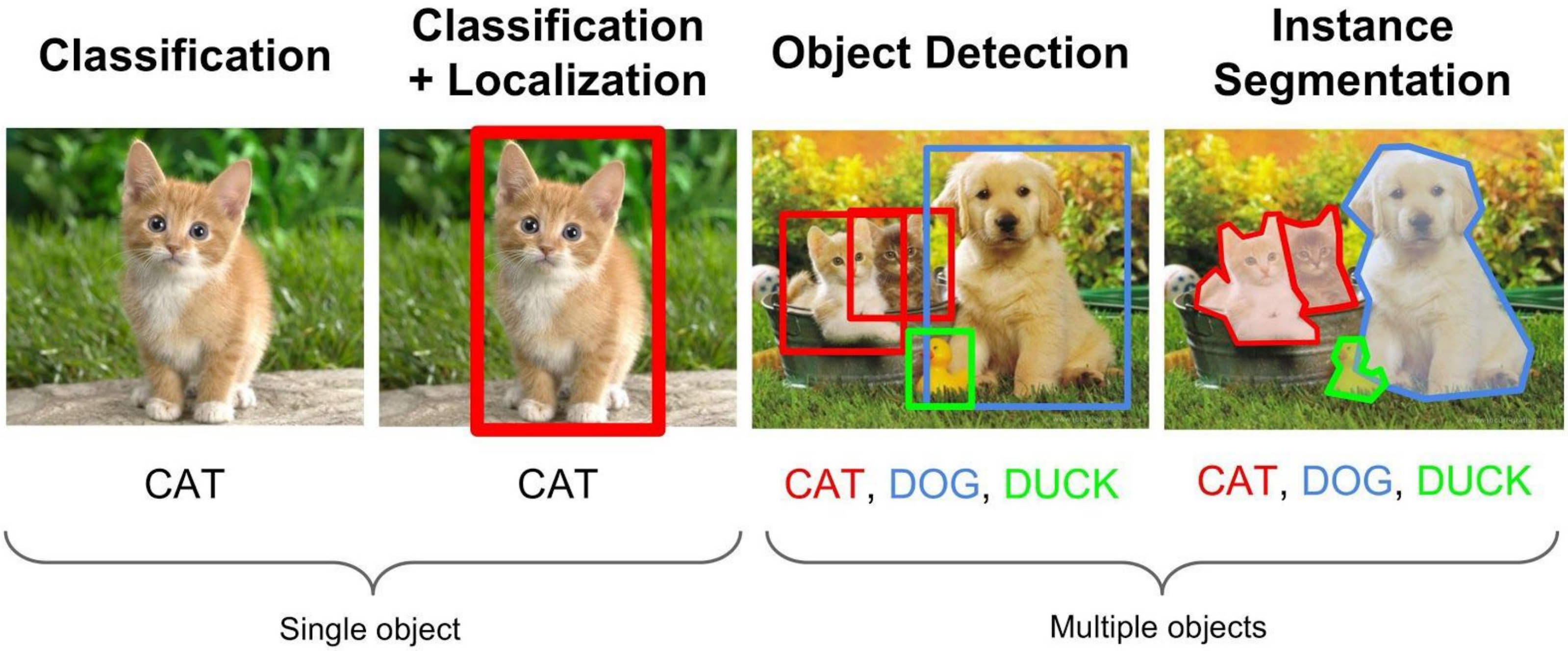
Viewpoint variations



Illumination changes

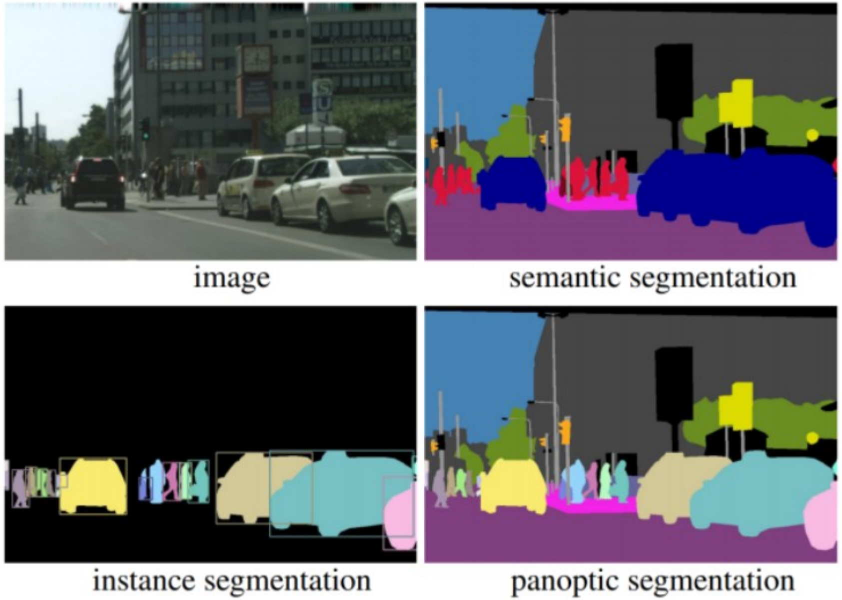
General weirdness of the world...

Semantic Scene Understanding



Source http://cs224d.stanford.edu/index.html

Semantic Image Segmentation



Semantic segmentation: classifying each pixel belonging to a particular label. It does not consider different instances of the same object.

Instance Segmentation: Assigns a unique label to every instance of a particular object in the image.

Panoptic Segmentation: Instance + Semantic Segmentation

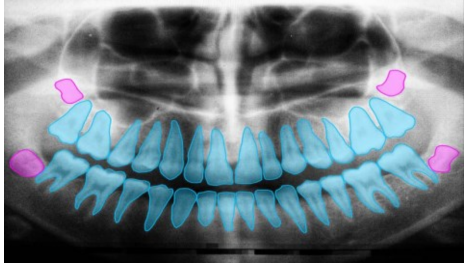
Source: Sultana, F., Sufian, A., & Dutta, P. (2020). Evolution of image segmentation using deep convolutional neural network: a survey. *Knowledge-Based Systems*, *201*, 106062.

Applications

Self-Driving Cars

Virtual Fitting Rooms

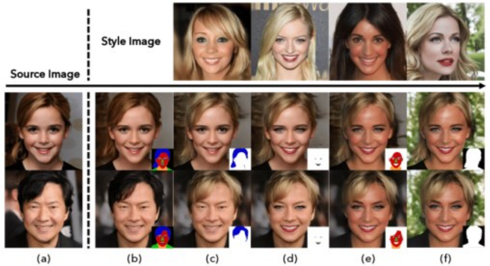
Medical Imaging and Diagnostics



GeoSensing

Precision Agriculture

Facial Segmentation for Face Editing

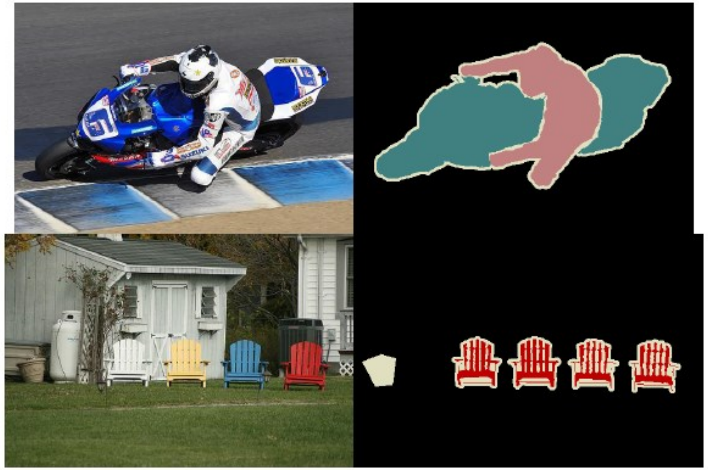


Dataset

train/val images: 118K/5K
>100 categories

Trainval images: 11540 (6,929 segmentation masks)
20 categories

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning



COCO
Common Objects in Context

Dataset

train/val images: 20K/2K
150 categories

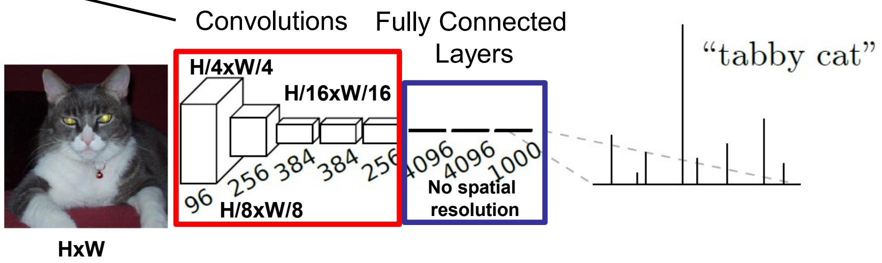
ADE20K

train/val images: 2750/500
30 categories, 19 used

CITYSCAPES DATASET

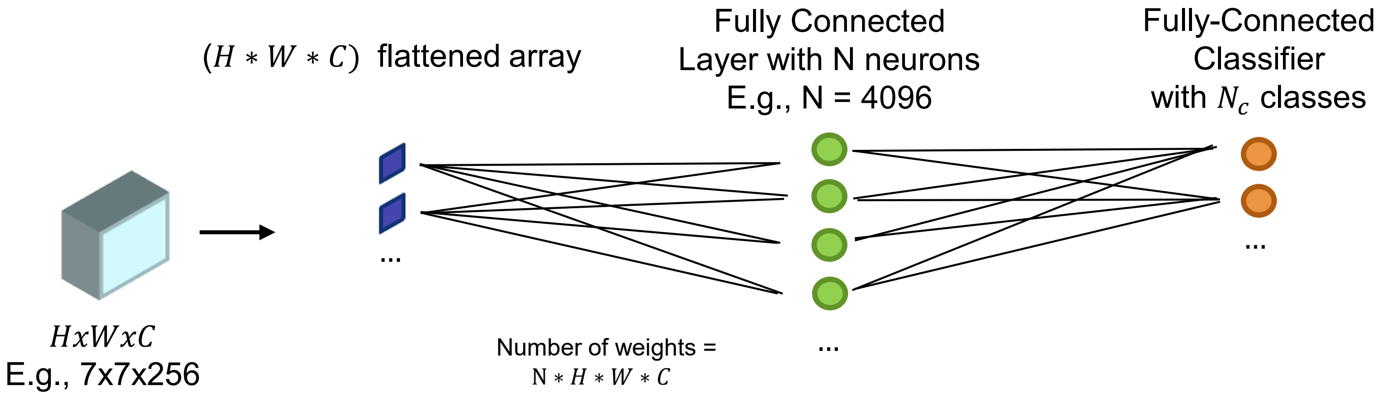
Classification Network

CNNs typically process images reducing size and increasing the number of channels



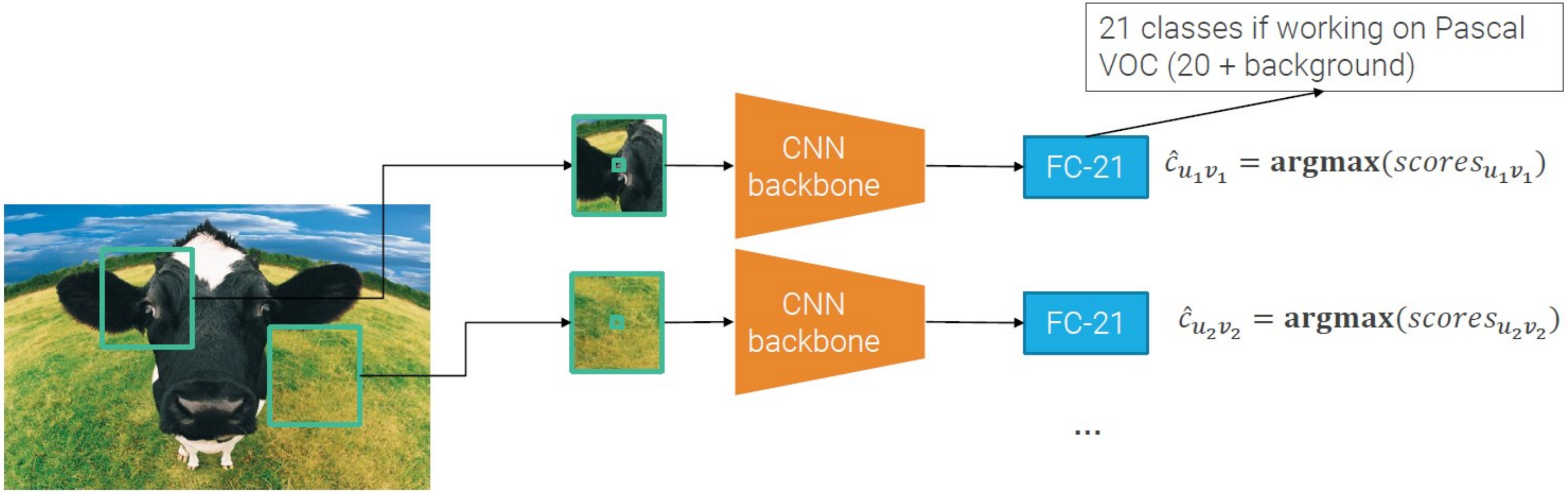
Typically uses Fully Connected Layers as final layers

Fixed Resolution



Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Slow R-CNN for segmentation

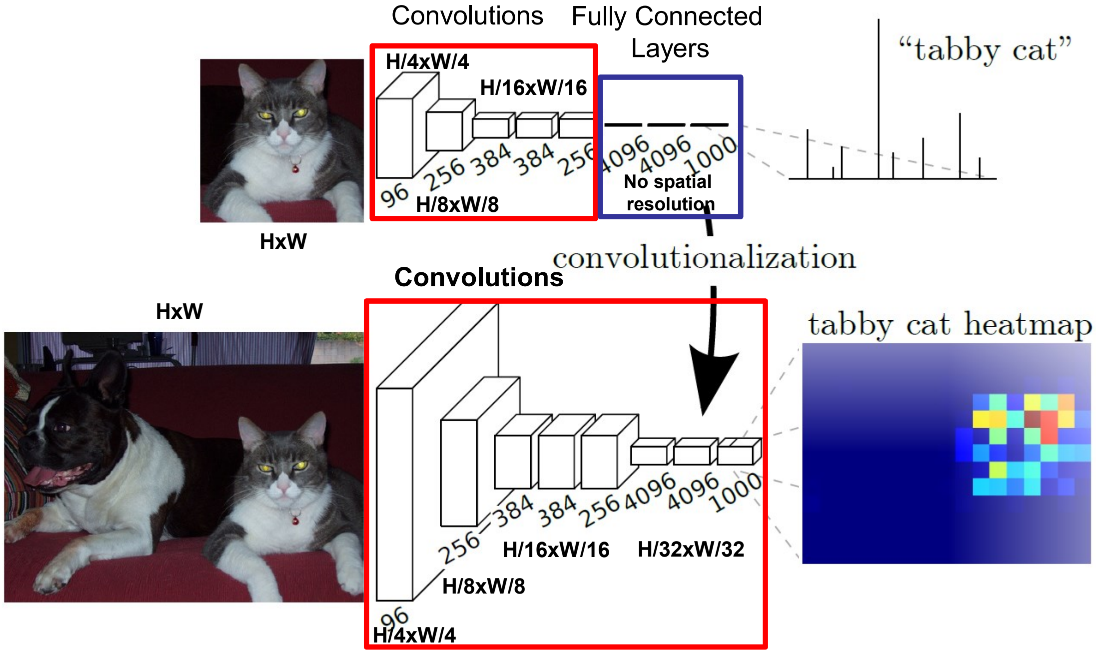


Slide window at all possible positions.
No proposals, must process each pixel.

Loss is the sum of the standard multi class loss over all pixels

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

Fully Convolutional Networks (FCN)



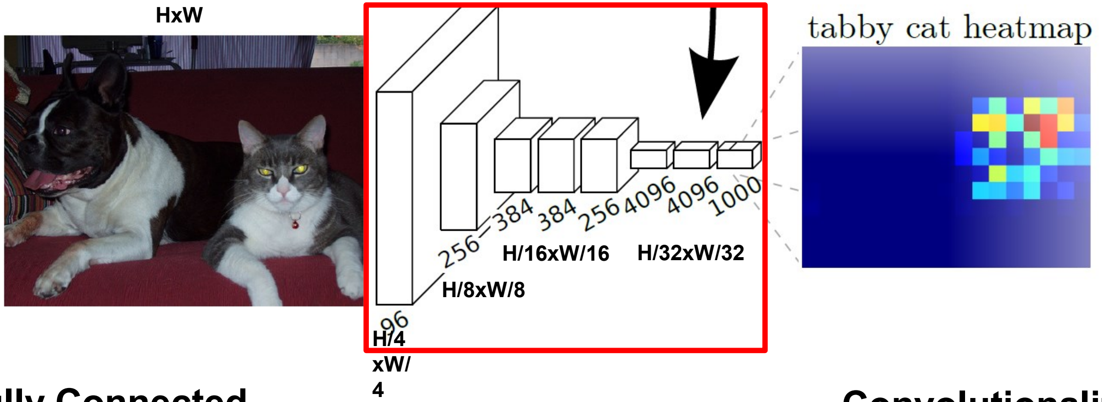
Classification Network:
Uses Fully Connected Layers in the final layers

Fully Convolutional Network:
Convert Fully Connected into Convolutional layers

Transforming fully connected layers into convolution layers enables a classification net to output a heatmap. Adding layers and a spatial loss produces an efficient machine for end-to-end dense learning.

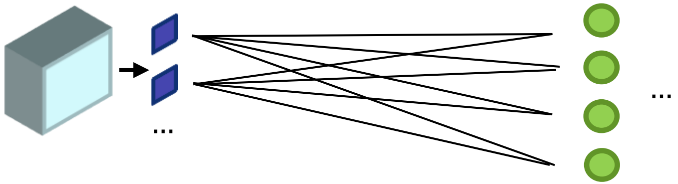
Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Fully Convolutional Networks (FCN)



Fully Convolutional Network: Convert Fully Connected into Convolutional layers

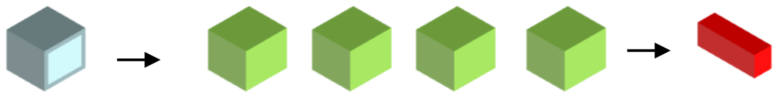
Fully Connected



Number of weights = $N * H * W * C$

Convolutionalization

Number of weights = $N * H * W * C$



Feature map
$HxWxC$
E.g., 7x7x256

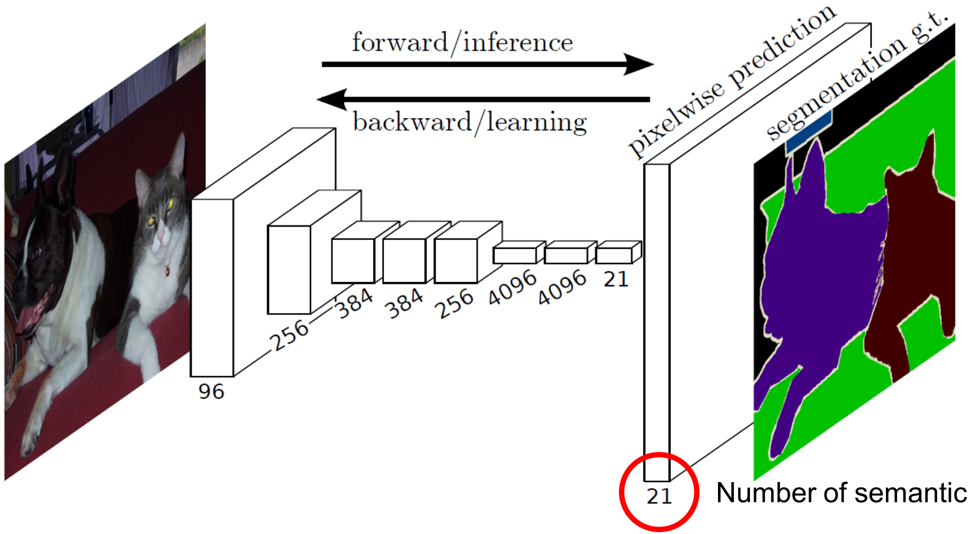
N convolutional filters
$HxWxC$

1x1xN

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Fully Convolutional Networks (FCN)

Upsampling the output to original resolution (e.g., bilinear interpolation)



Number of semantic classes

Pixel-wise loss (e.g., Cross Entropy)

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

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Fully Convolutional Networks
Upsampling

Input

1	2
3	4

Cx2x2

Nearest Neighbor

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

Cx4x4

Bilinear interpolation

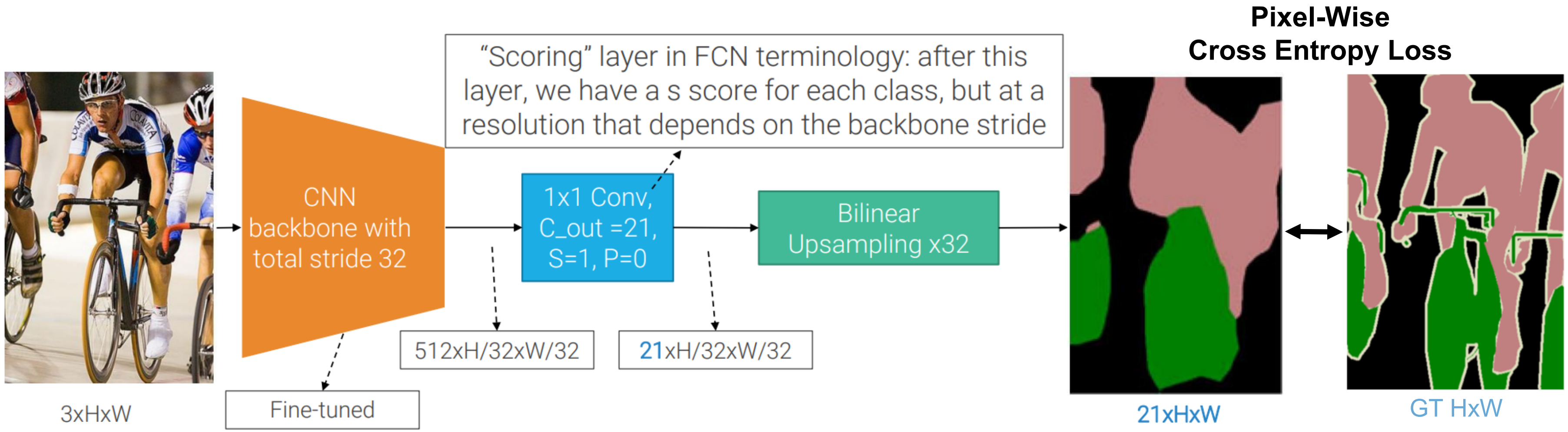
1	1.25	1.75	2
1.50	1.75	2.25	2.5
2.5	2.75	3.25	3.5
3	3.25	3.75	4

Cx4x4

One way to perform upsampling can be to use standard, not-learned image processing operators

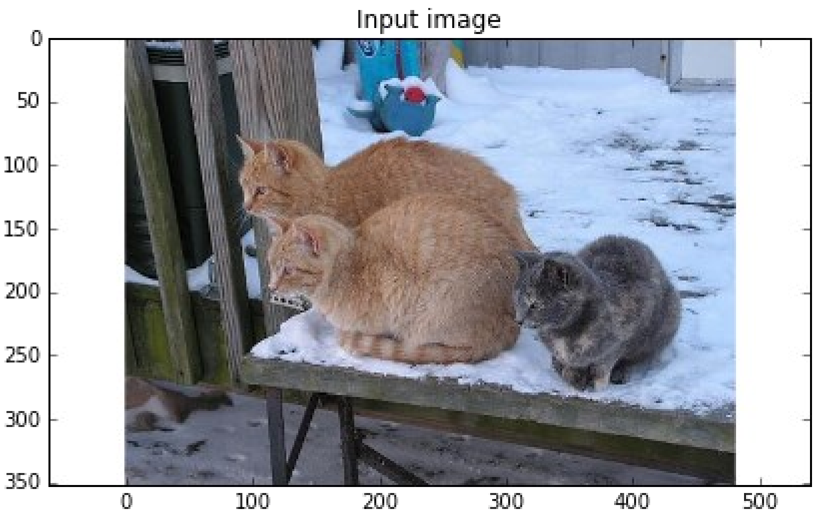
Source: Samuele Salti, Machine Learning for Computer Vision, University of Bologna

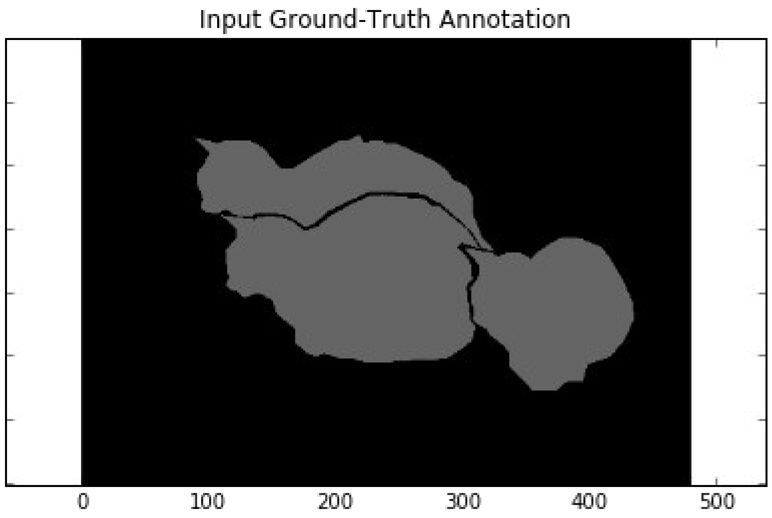
FCN-32s



Source: Samuele Salti, Machine Learning for Computer Vision, University of Bologna

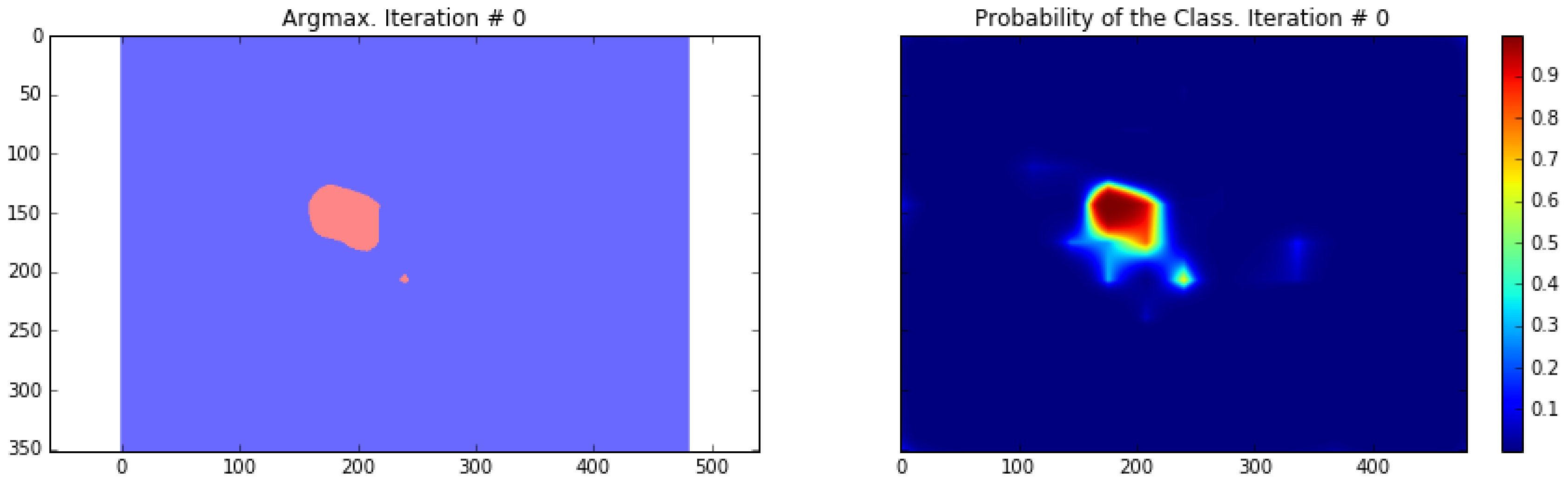
Overfitting example





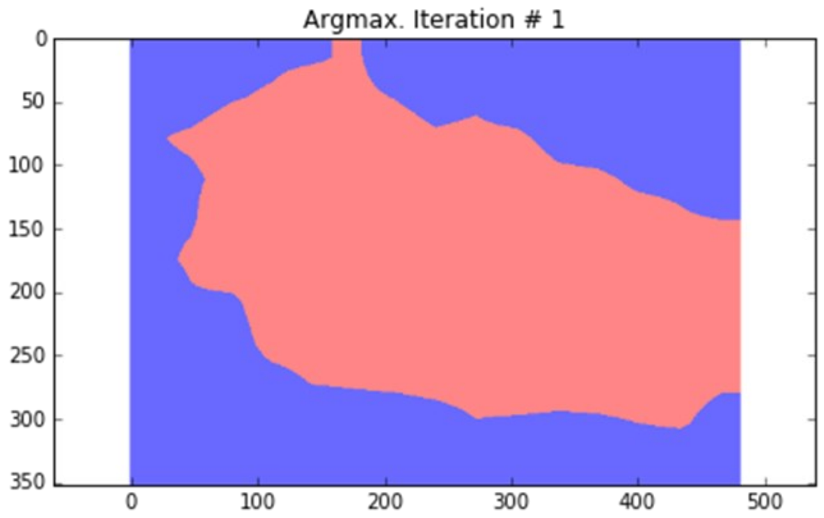
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

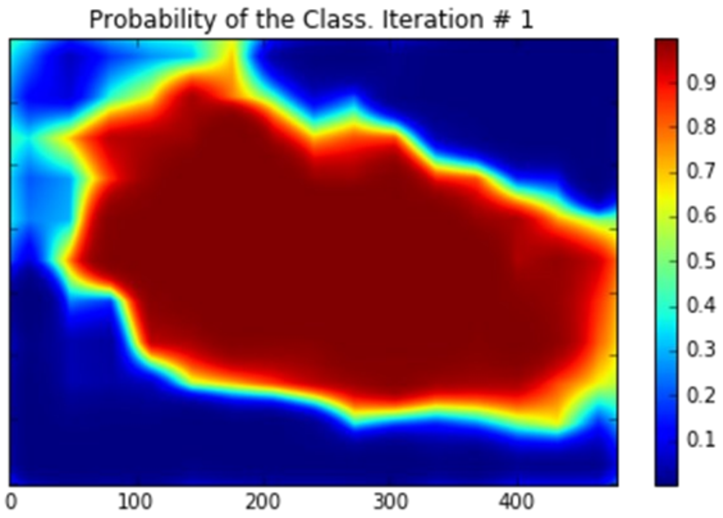
Overfitting example



Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

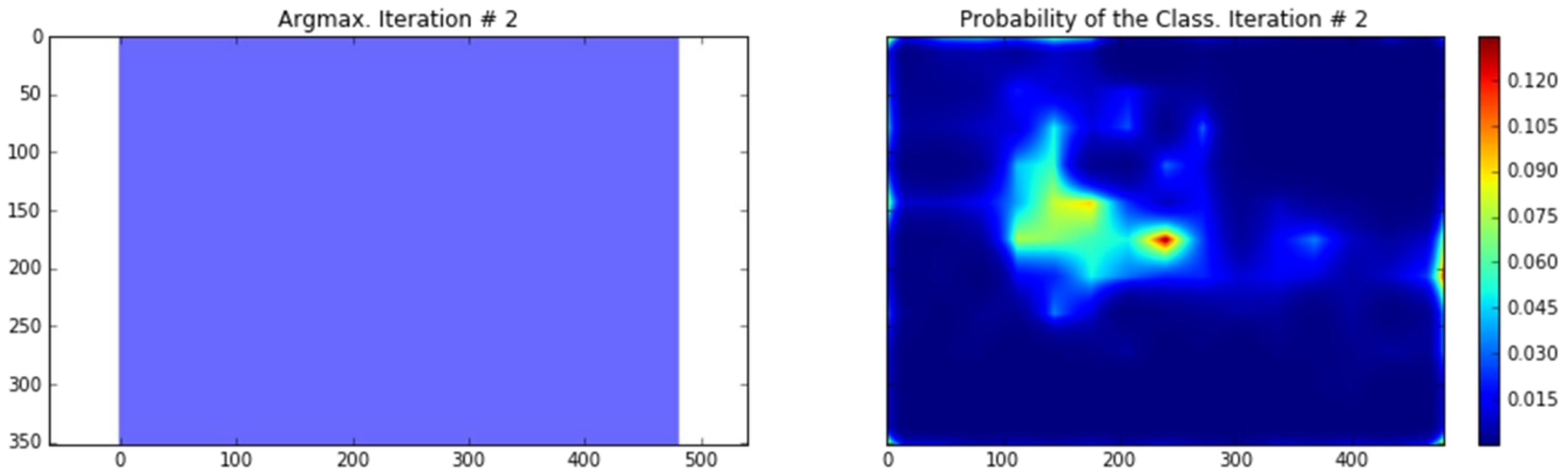
Overfitting example





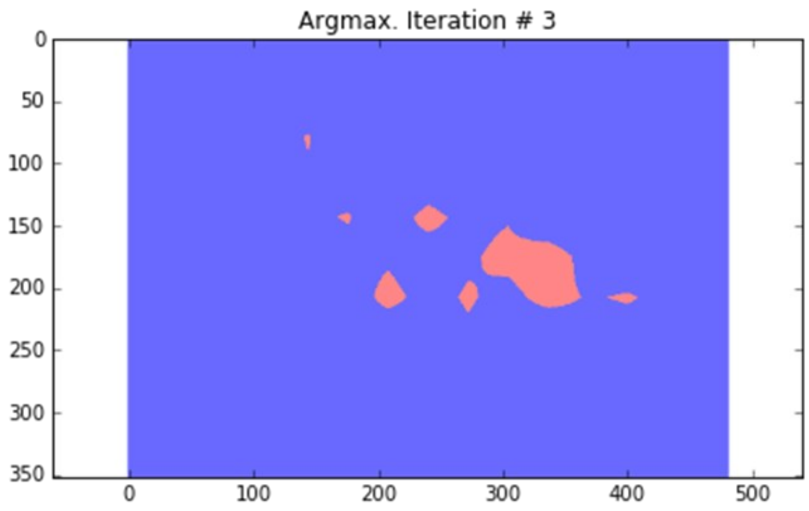
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

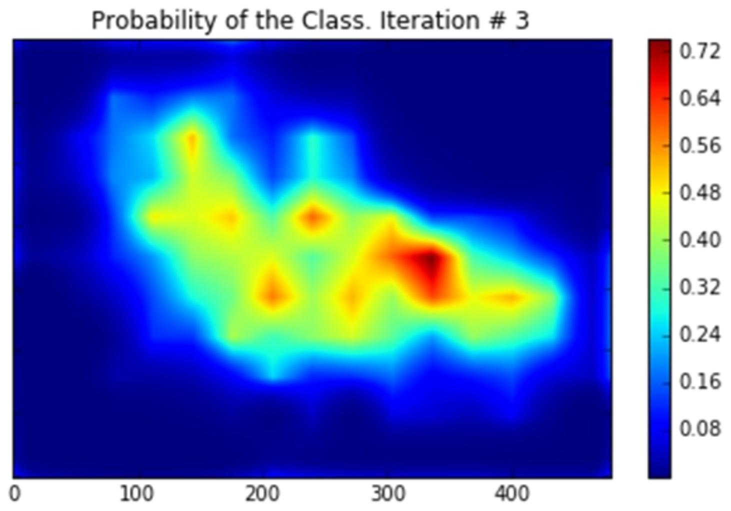
Overfitting example



Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

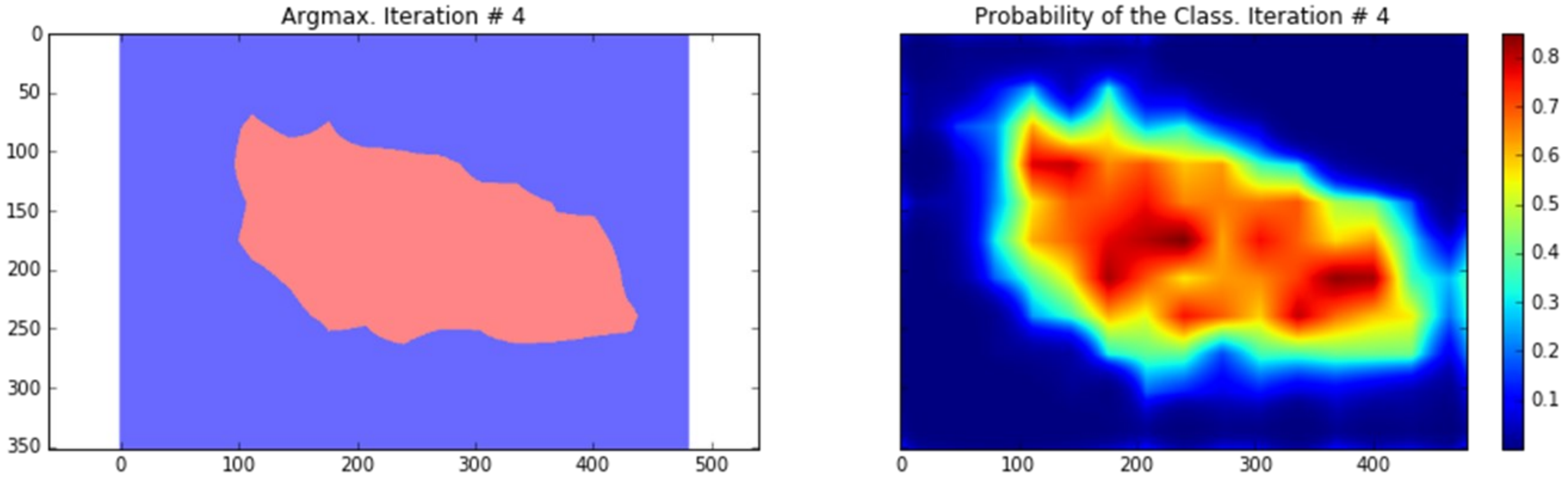
Overfitting example





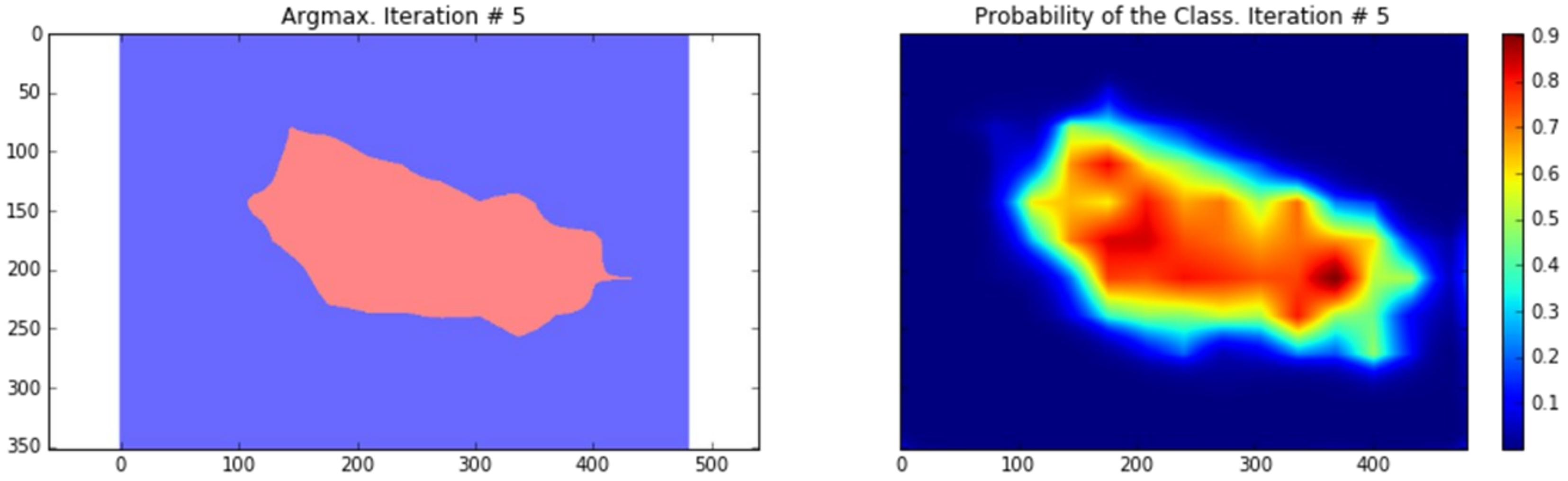
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Overfitting example



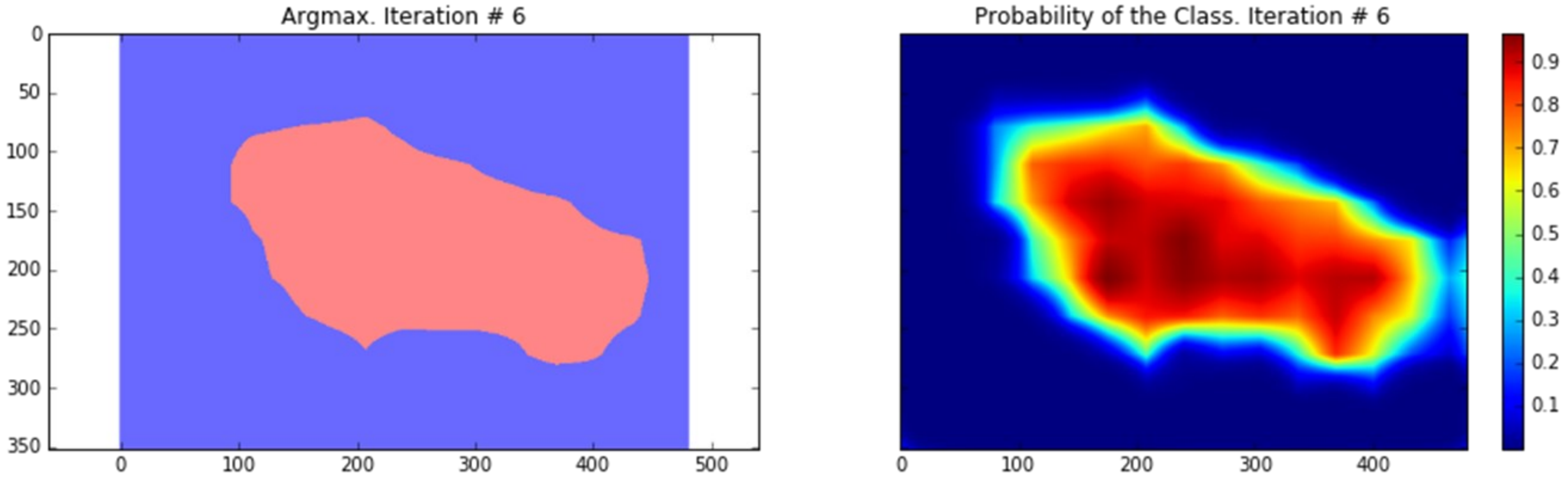
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Overfitting example



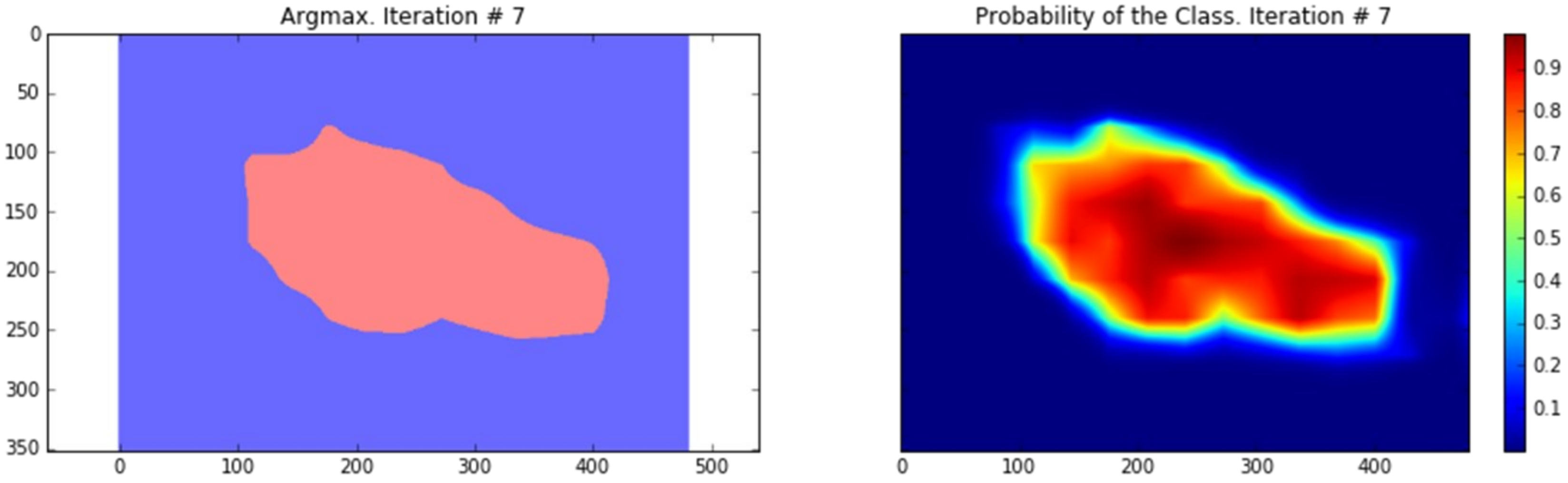
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Overfitting example



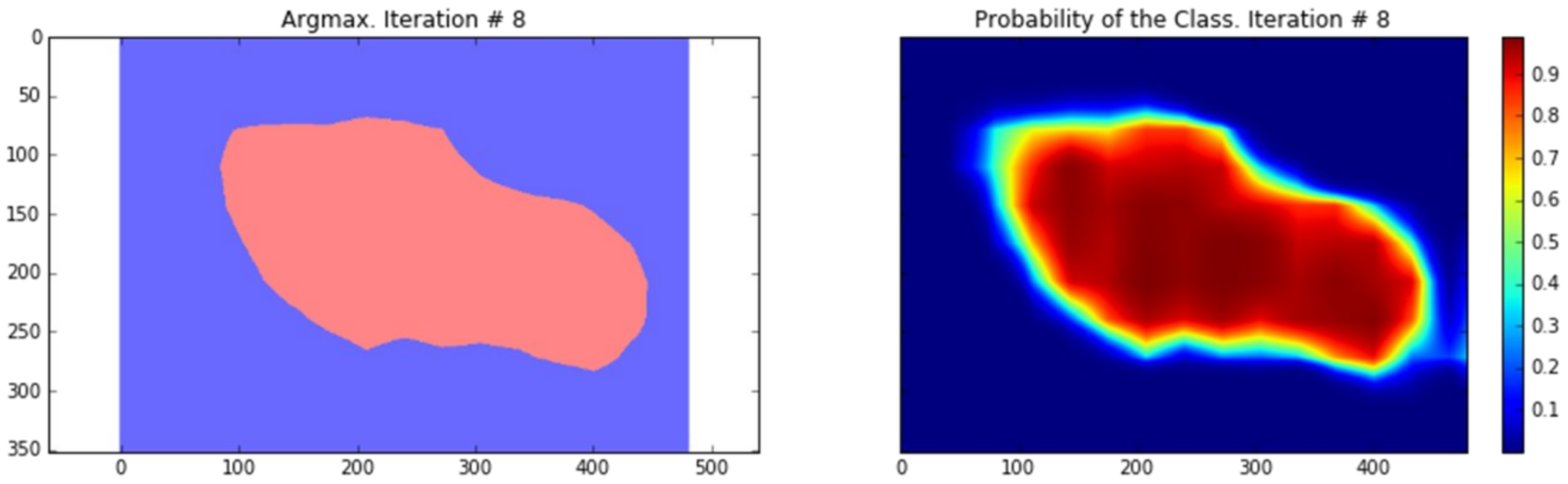
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Overfitting example



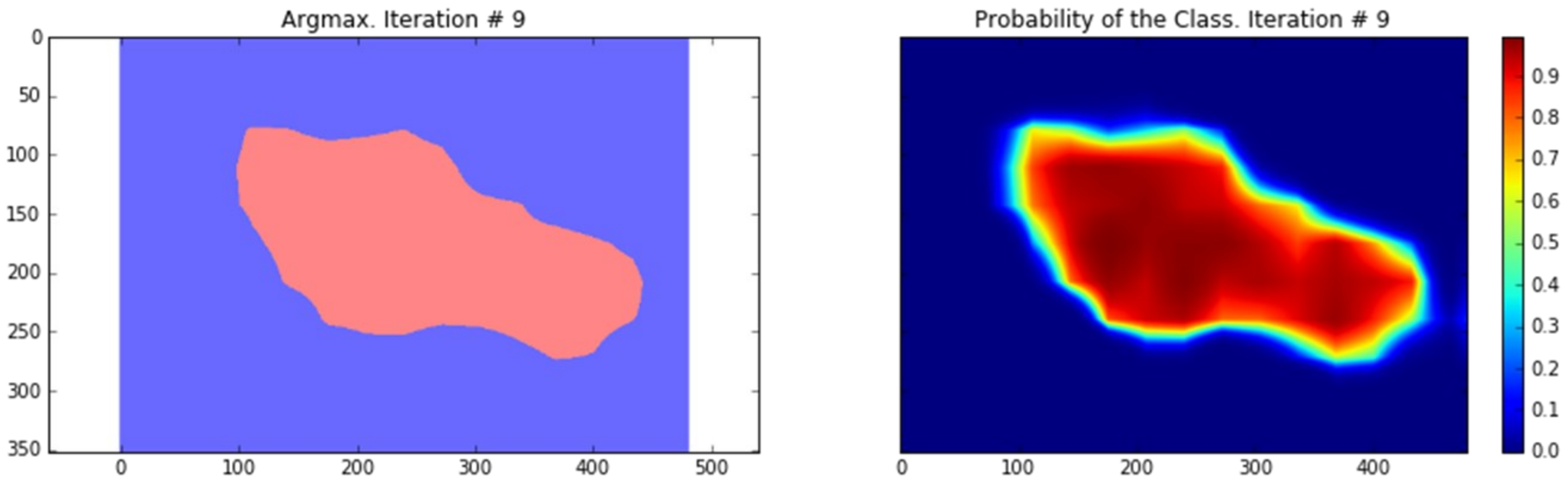
Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Overfitting example



Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

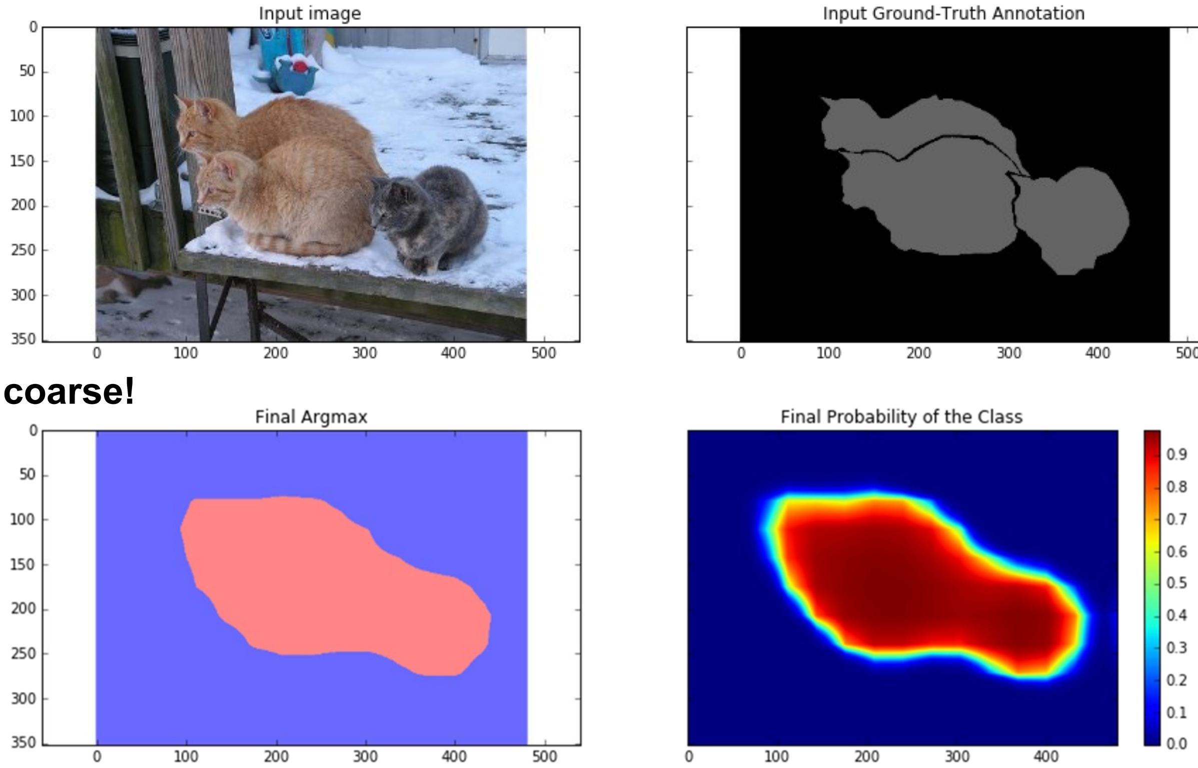
Overfitting example



Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Overfitting example

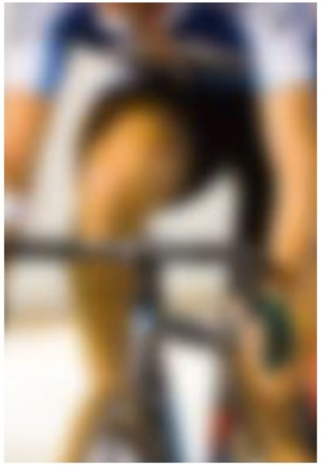
Prediction is too coarse!

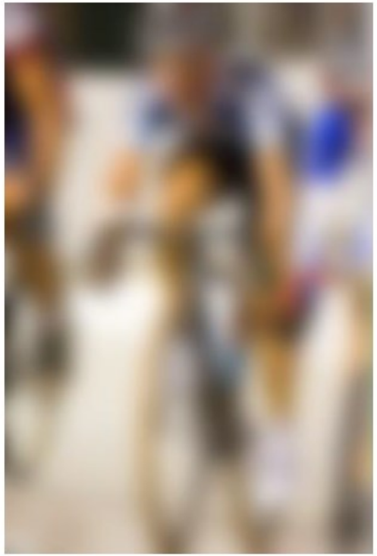


Source: Daniil Pakhomov & Vittal Premachandran. Dense-ai: Image Segmentation and Object Detection library

Example of 1/32 image

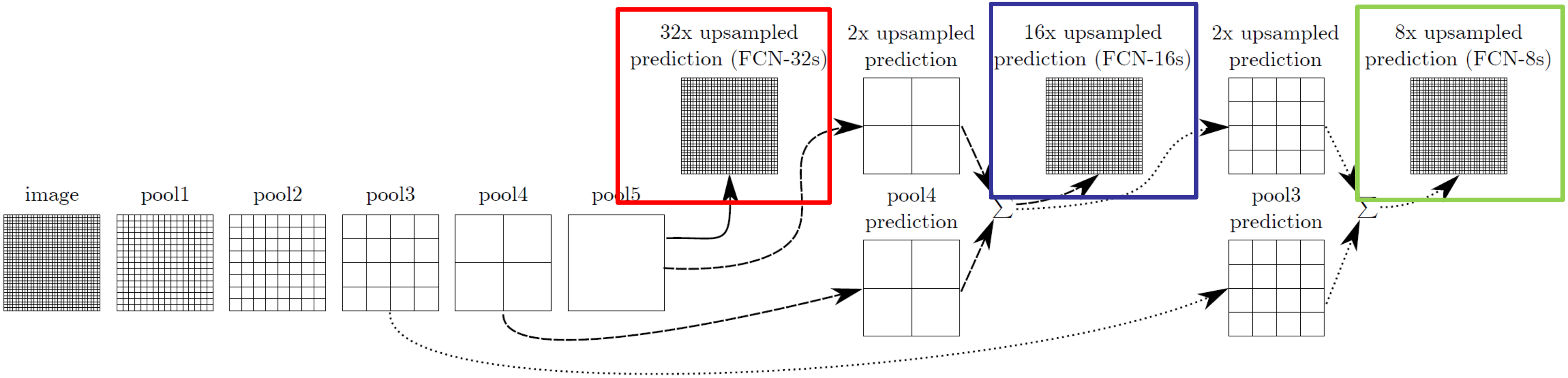
What and Where





Global Information resolves what.
Local information resolves where.

High-Frequency Details: Skip Connections



Single-stream FCN (FCN-32s)
upsamples stride 32 predictions to full resolution

FCN-16s
Final + pool4 layer, at stride 16. Predict finer details, while retaining high-level semantic information.

FCN-8s
Additional predictions from pool3, at stride 8, provide further precision.

Evaluation Metrics: mIoU

Recall IoU for Overlapping Boxes

To check if a prediction and a GT box overlap, we measure the **Intersection over Union (IoU)** score (aka Jaccard index or similarity)

$$IoU(BB_i, BB_j) = \frac{area\ of\ intersection}{area\ of\ union}$$

$$= \frac{|BB_i \cap BB_j|}{|BBi| + |BBj| - |BBi \cap BBj|}$$

Source: Samuele Salti, Machine Learning for Computer Vision, University of Bologna

Evaluation Metrics: mIoU

$$TP_c = \sum_{images} \#\, pixels\ where\ y_{uv} = c\ and\ \hat{y}_{uv} = c$$

$$IoU_c = \frac{area\ of\ intersection}{area\ of\ union}$$

$$\sum_{images} (\#\, pixels\ where\ y_{uv} = c + \#\, pixels\ where\ \hat{y}_{uv} = c) - TP_c$$

To compute **mIoU** score for a dataset, we average IoU_c over classes:

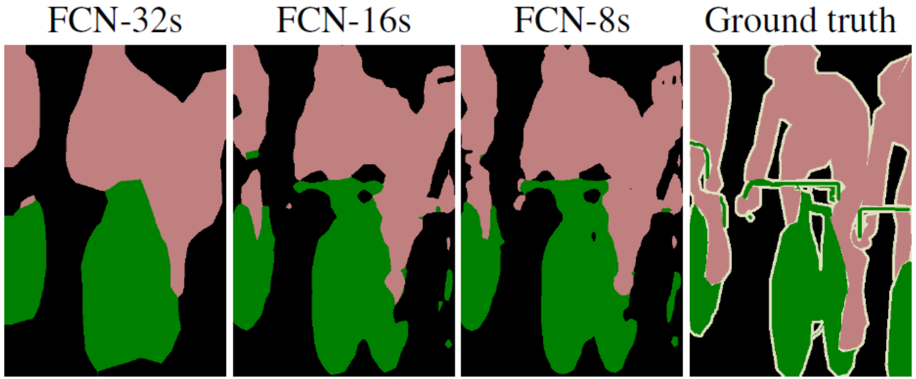
$$mIoU = \frac{1}{C} \sum_{c=1}^{C} IoU_c$$

$\hat{y}$

y

Source: Samuele Salti, Machine Learning for Computer Vision, University of Bologna

FCN Ablation on Skip Connections



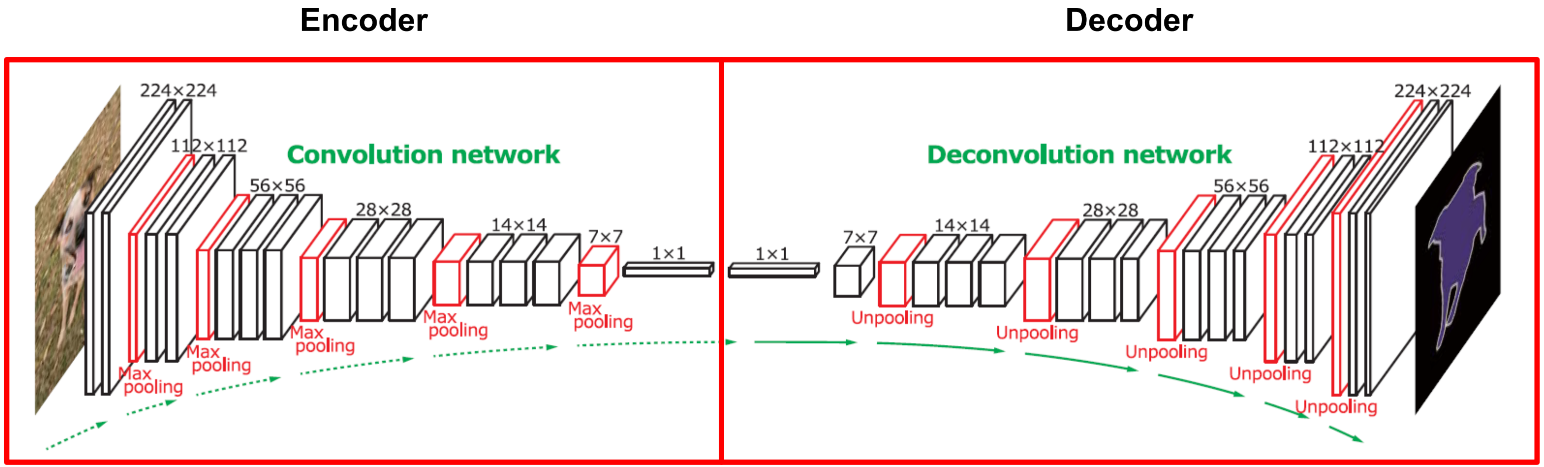
	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s	90.5	76.5	63.6	83.5
FCN-16s	91.0	78.1	65.0	84.3
FCN-8s at-once	91.1	**78.5**	65.4	84.4
FCN-8s staged	**91.2**	77.6	**65.5**	**84.5**
FCN-32s fixed	82.9	64.6	46.6	72.3
FCN-pool5	87.4	60.5	50.0	78.5
FCN-pool4	78.7	31.7	22.4	67.0
FCN-pool3	70.9	13.7	9.2	57.6

- No difference between training end-to-end (“at-once”) or coarse-to-fine (i.e., first train FCN-32s than add skips and fine-tune, “staged” in the table).
- Fine-tuning the backbone is very important (fixed row is without fine-tuning)
- Merging different resolutions with skips is also very important: for instance, FCN-pool4 reports the (poor) quality of the predictions we get from strided16 activations, if we do not merge them with coarser data

Refining fully convolutional nets by fusing information from layers with different strides improves segmentation detail. The first three images show the output from our 32, 16, and 8 pixel stride nets.

With VGG backbone, found basically no improvements after predicting from stride 8 activations (2 skips).

Learned Upsampling: Encoder-Decoder

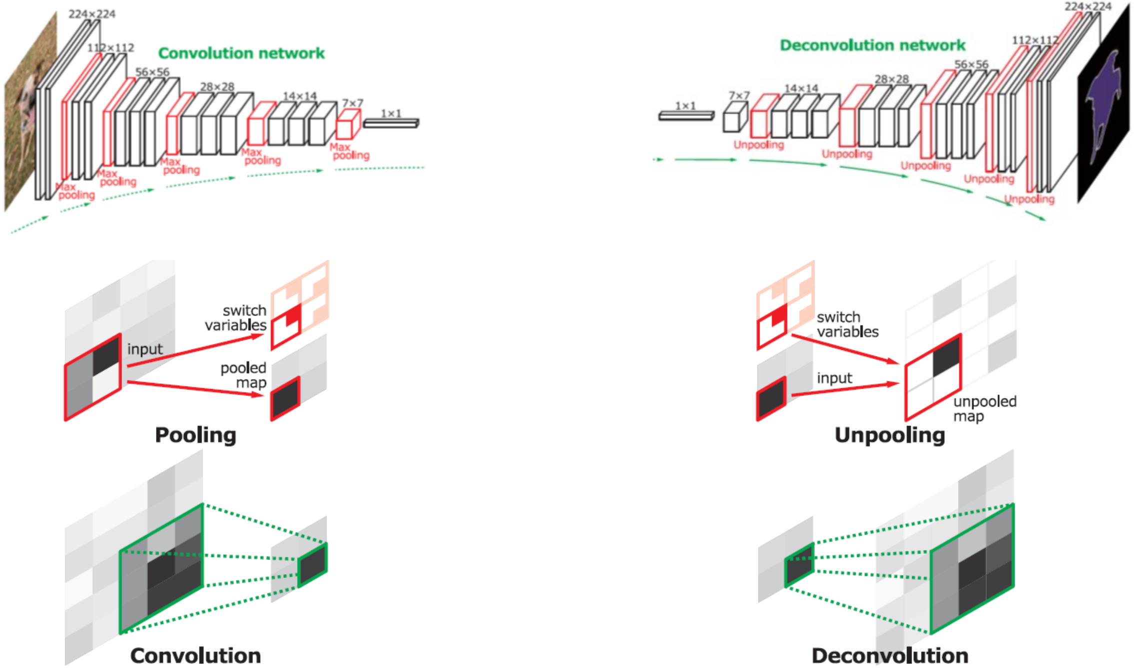


DeconvNet:

- Encoder-Decoder
- Unpooling Layers
- Deconvolutions (or Transposed Convolutions)
- Last layers are **fully connected** implemented as convolutional (based on training dataset resolution)
- Instance-wise Segmentation

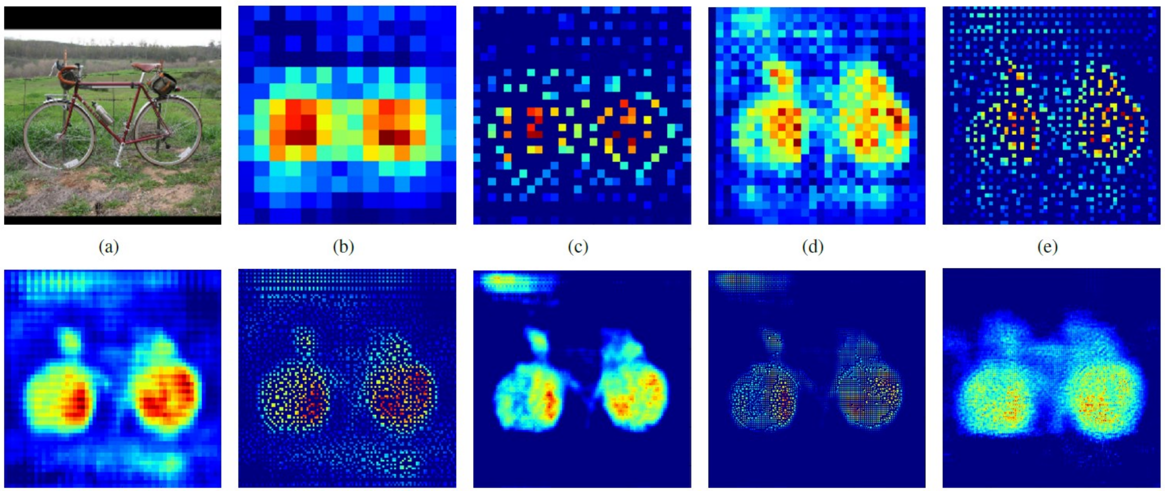
Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520-1528).

Upsampling the predictions in the decoder



Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520-1528).

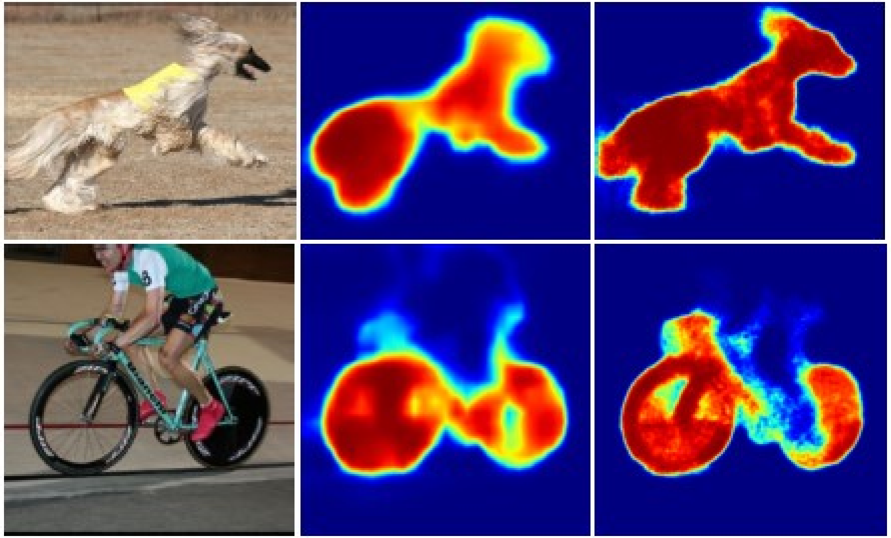
Learned Upsampling: Encoder-Decoder



Visualization of activations in the decoder of DeconvNet.
Finer details of the object are revealed, as the features are forward-propagated through the layers in the decoder

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520-1528).

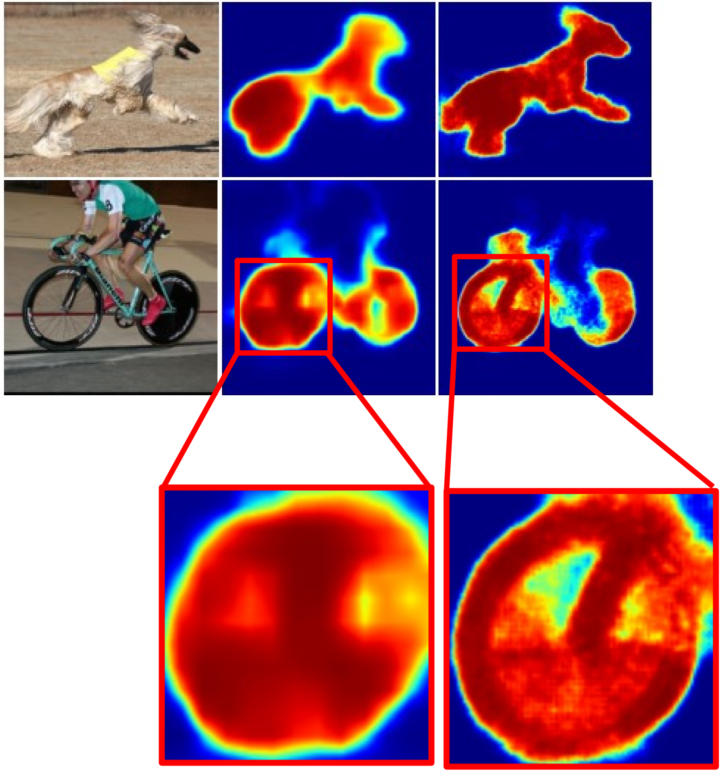
FCN vs DeconvNet
Class-Conditional probability maps visualization



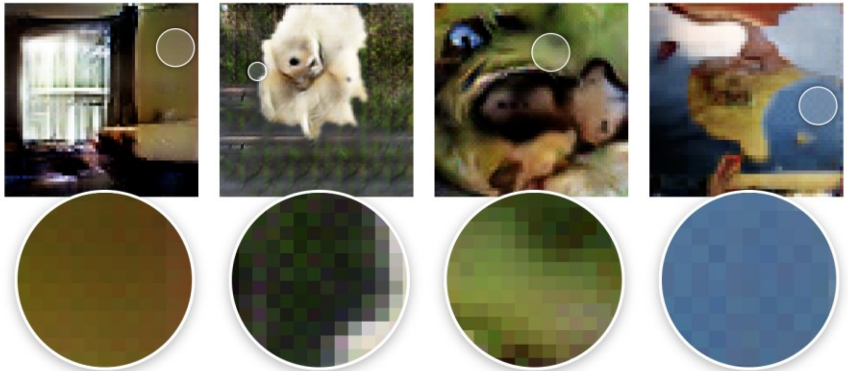
Input image FCN DeconvNet

Deconvolutions: Grid Artifacts

Deconvolutions in Semantic Segmentation

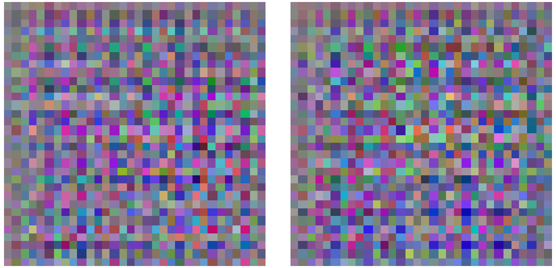


Deconvolutions in Image Generation



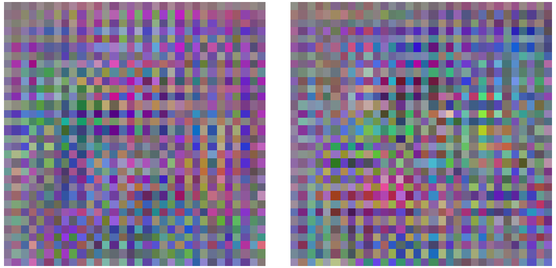
Partial Improvement:
Resize-Convolutions
(aka Up-Convolutions)
Upsampling NN + Convolution

Deconvolutions: Grid Artifacts



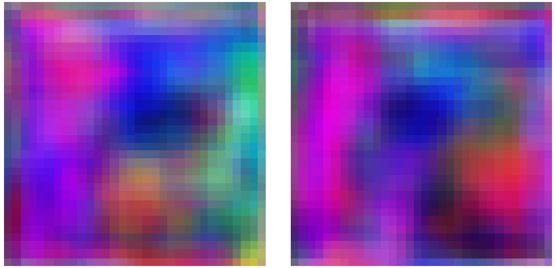
Deconvolution in last two layers.

Artifacts prior to any training.



Deconvolution only in last layer.

Artifacts prior to any training.

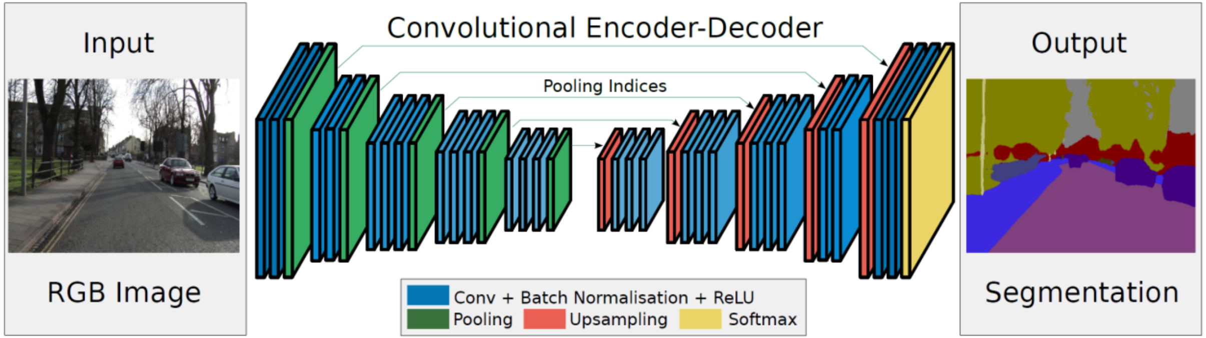


All layers use resize-convolution.

No artifacts before or after training.

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10), e3.

SegNet

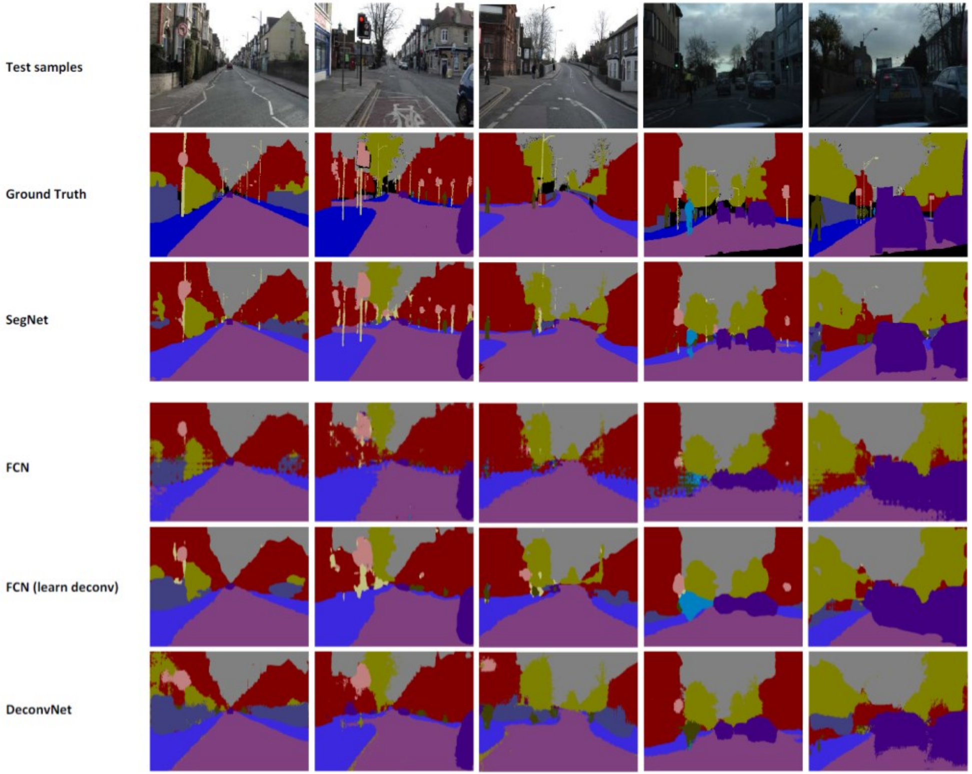


SegNet:

- Fully-convolutional (no fully-connected)
- Encoder-Decoder
- Decoder: Unpooling Layers (as DeconvNet) + Convolution

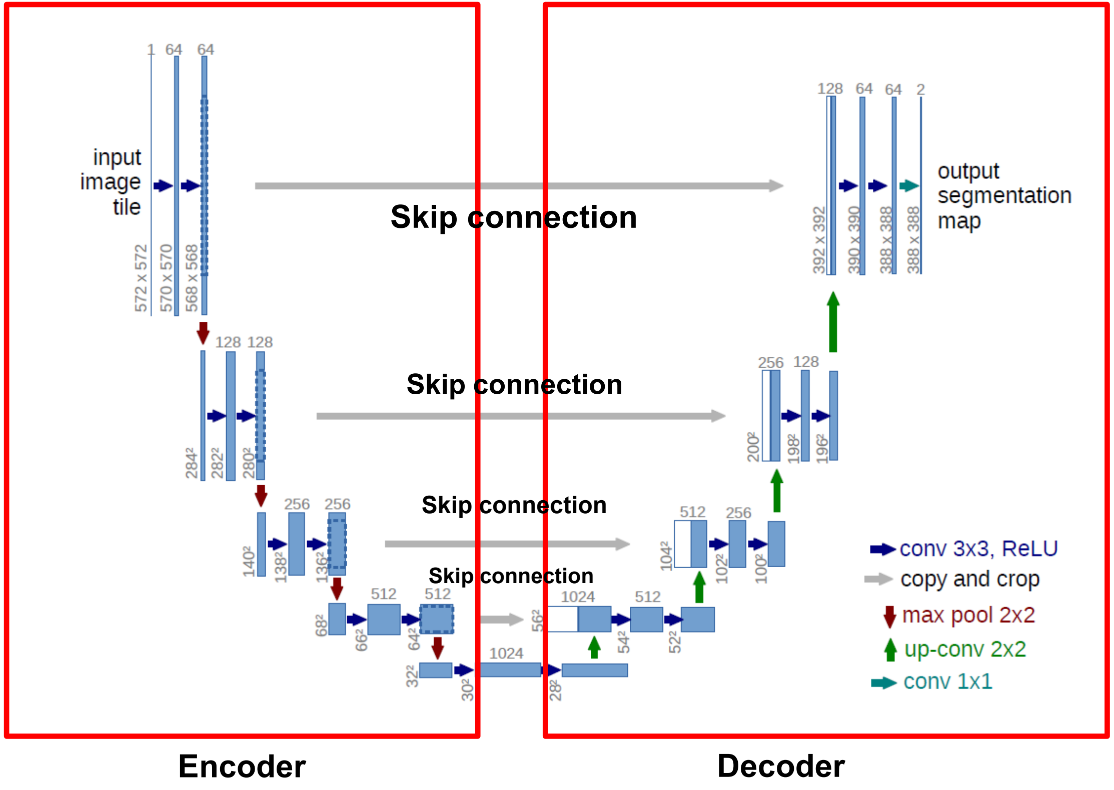
Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481-2495.

Qualitative Comparison



Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481-2495.

Recovering low-level information: U-Net



U-Net

- Fully-convolutional
- Encoder-Decoder
- Decoder: Up-Convolutions
- More Skip-connections

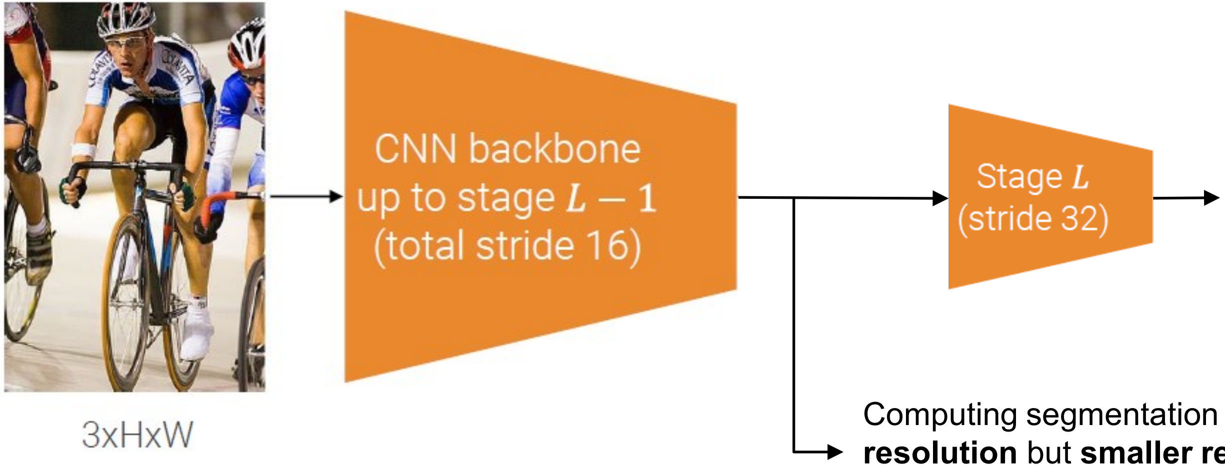
Skip connections use **concatenation** instead of summation as in FCN.

2x2 stride 2 transposed convolutions ("up convolutions") are used to upsample the activations in the decoder, while halving the number of channels.

Normal 3x3 convolutions are used in the decoder as well: with further processing, even **initial layers** of the backbone can effectively **contribute** to the final segmentation mask, as opposed to what happened in FCN.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

Dilated Convolutions



Predicting from this activation has rich semantic info due to depth of processing and large receptive field but is **spatially coarse**

Computing segmentation mask from this activation has **detailed resolution** but **smaller receptive field** and **not enough semantic info**

Is the architecture inherited by classification backbones the best one for semantic segmentation (or dense tasks in general)?

We would like to have rich features with large spatial resolutions, large receptive fields and constant cost.

Atrous Convolution

Dilated (or «atrous ») convolutions expose an additional parameters, the dilation rate $\boldsymbol{r}$. Equivalent to inserting holes (‘ trous ’ in French) between filter weights. r=1 gives the usual, dense, convolution.

$$[K * I](i,j) = \sum_{n=1}^{3}\sum_{m}\sum_{l} K_n(m,l)I_n(j-\boldsymbol{r}\,m, i-\boldsymbol{r}\,l) + b$$

K_{11}	K_{12}	K_{13}
K_{21}	K_{22}	K_{23}
K_{31}	K_{32}	K_{33}

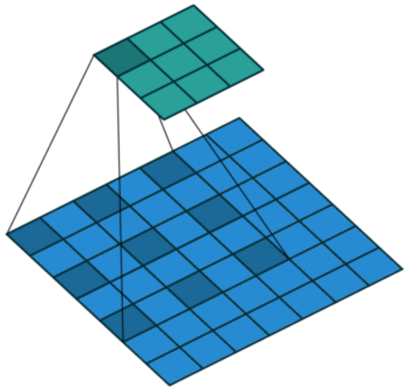
3x3 kernel
$r = 1$

K_{11}	0	K_{12}	0	K_{13}
0	0	0	0	0
K_{21}	0	K_{22}	0	K_{23}
0	0	0	0	0
K_{31}	0	K_{32}	0	K_{33}

3x3 kernel
$r = 2$

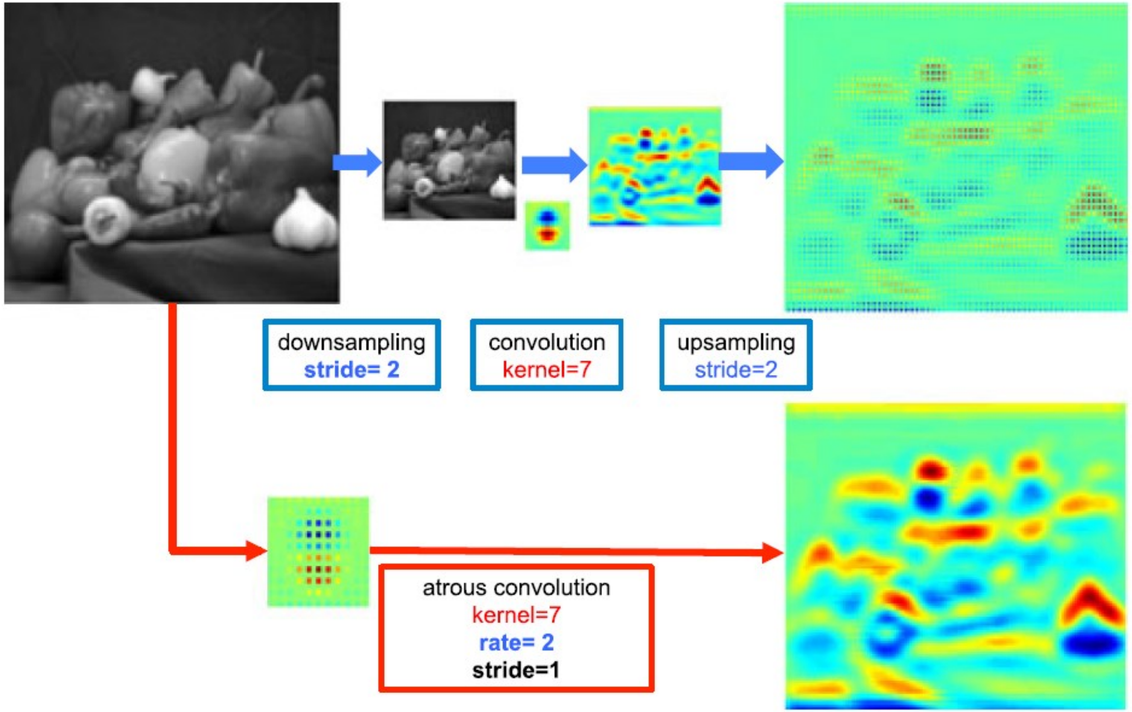
K_{11}	0	0	0	K_{12}	0	0	0	K_{13}
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
K_{21}	0	0	0	K_{22}	0	0	0	K_{23}
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
K_{31}	0	0	0	K_{32}	0	0	0	K_{33}

3x3 kernel
$r = 4$



Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834-848.

Atrous Convolution (2D)



Sparse feature extraction with standard convolution on a low-resolution input feature map.

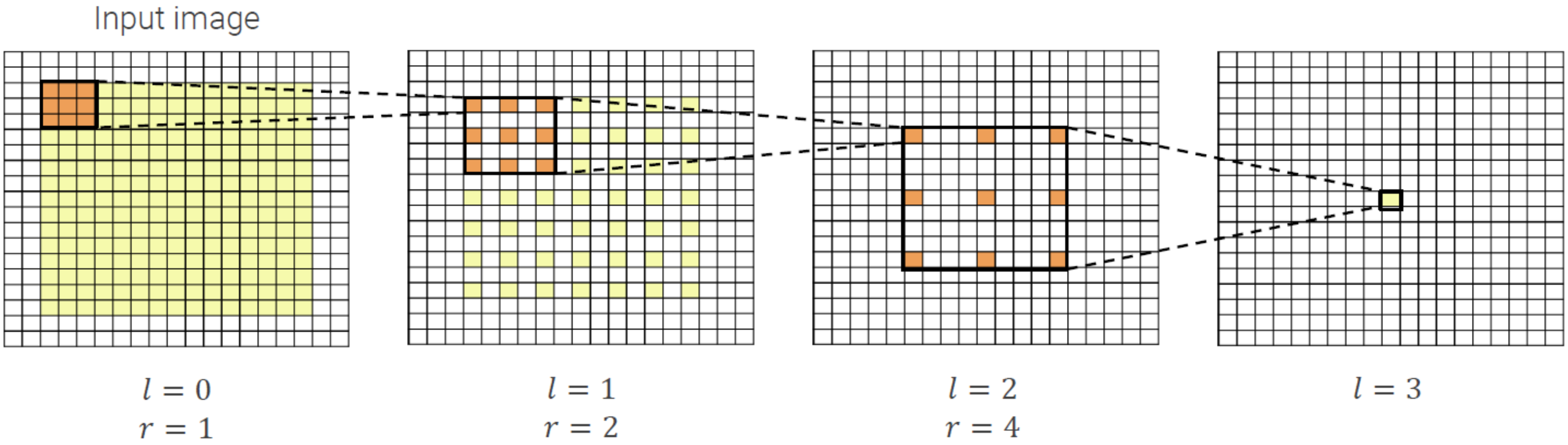
Dense feature extraction with atrous convolution with rate r=2, applied on a high-resolution input feature map.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834-848.

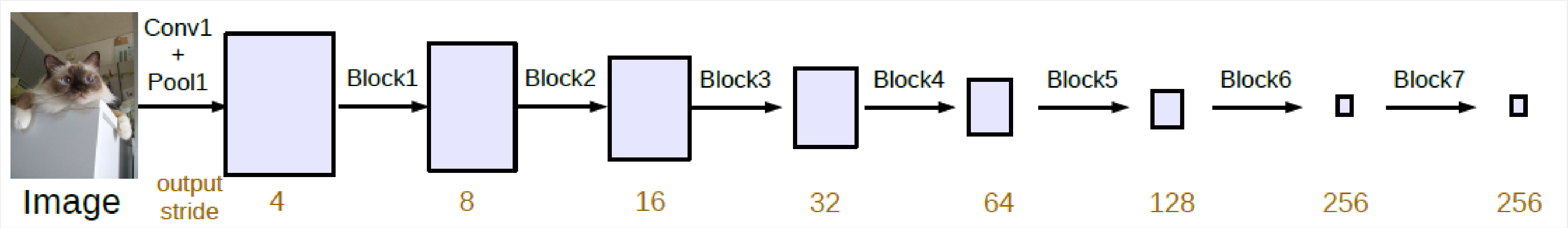
Atrous Convolution

If we stack dilated convolutions **with exponentially increasing dilation rate $r_l = 2^l$, the effective receptive field grows exponentially with the number of layers, while the number of parameters grows linearly, and resolution is not reduced.**

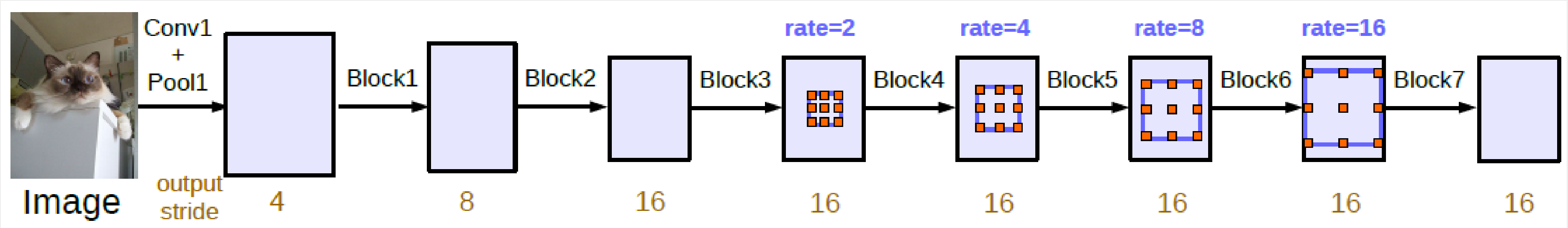
In general, at level l the receptive field of an activation entry will be $(2^{l+1} - 1) * (2^{l+1} - 1)$. For instance, at level 3 below, the receptive field is 15x15, while with dense 3x3 convolutions it would have been 7x7.



Going Deeper: Convolutions vs Atrous Convolutions



(a) Going deeper without atrous convolution.

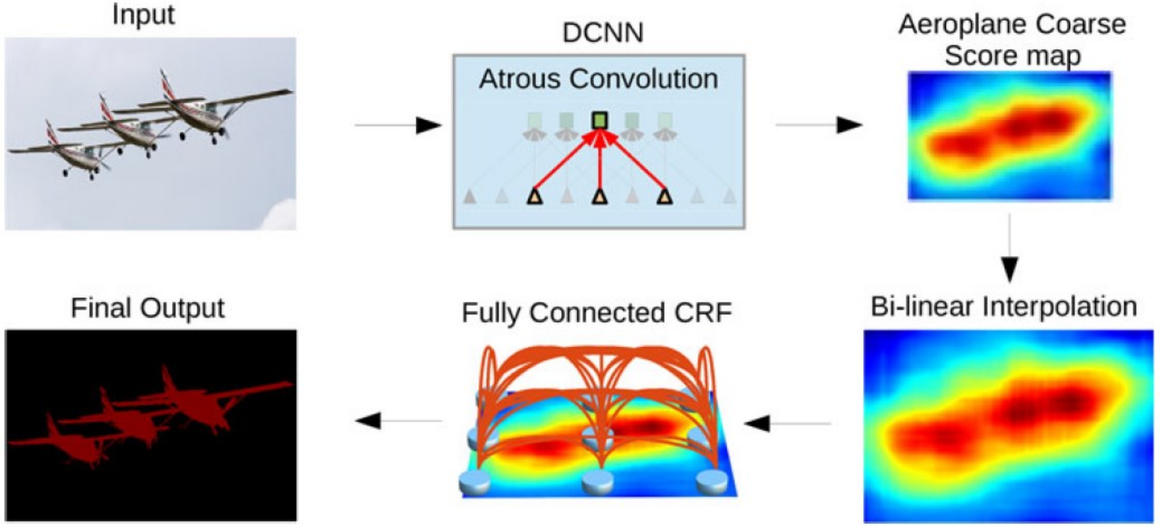


(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

Convolutions and downsampling increase receptive field.
However, several spatial information and high-frequency details are lost.

Deeplabv1

Challenge 1: reduced feature resolution -> Atrous Convolutions



Deeplab v1:

- Full Convolutional Atrous Convolution
- Fully Connected Conditional Random Field (CRF)

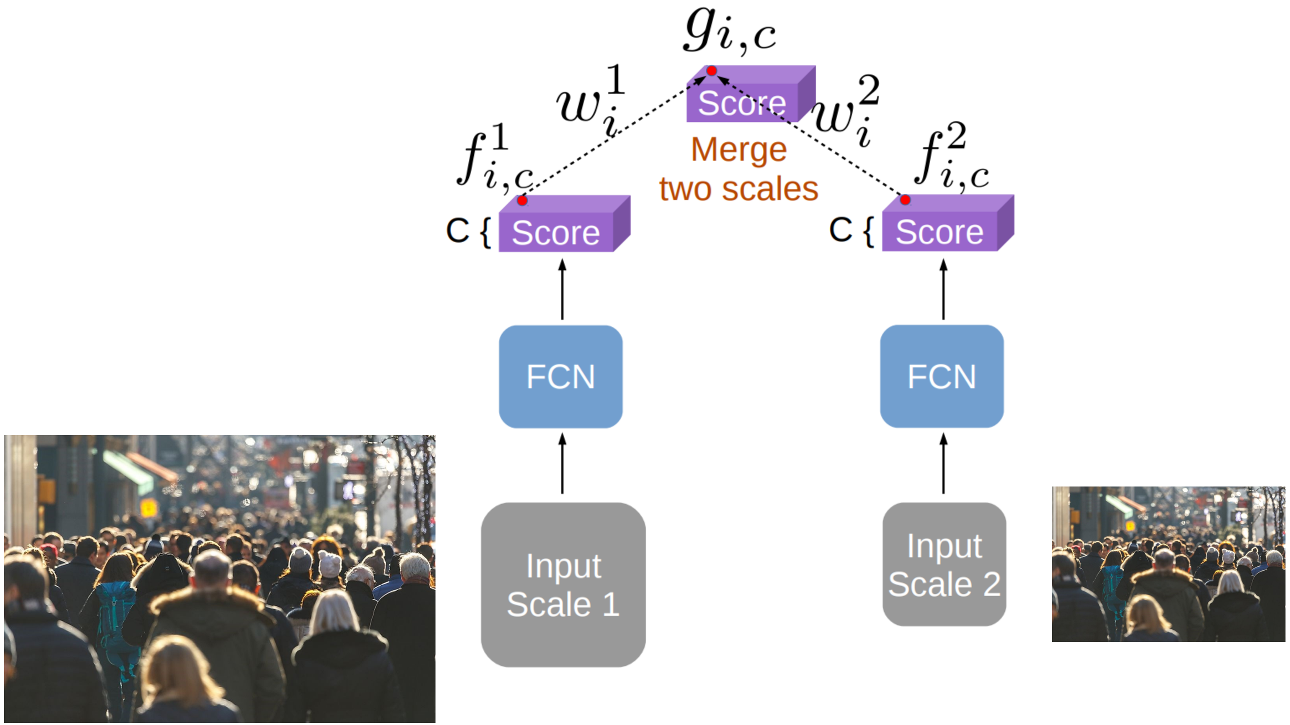
Challenge 2: reduced localization accuracy due to DCNN invariance -> CRF

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.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834-848.

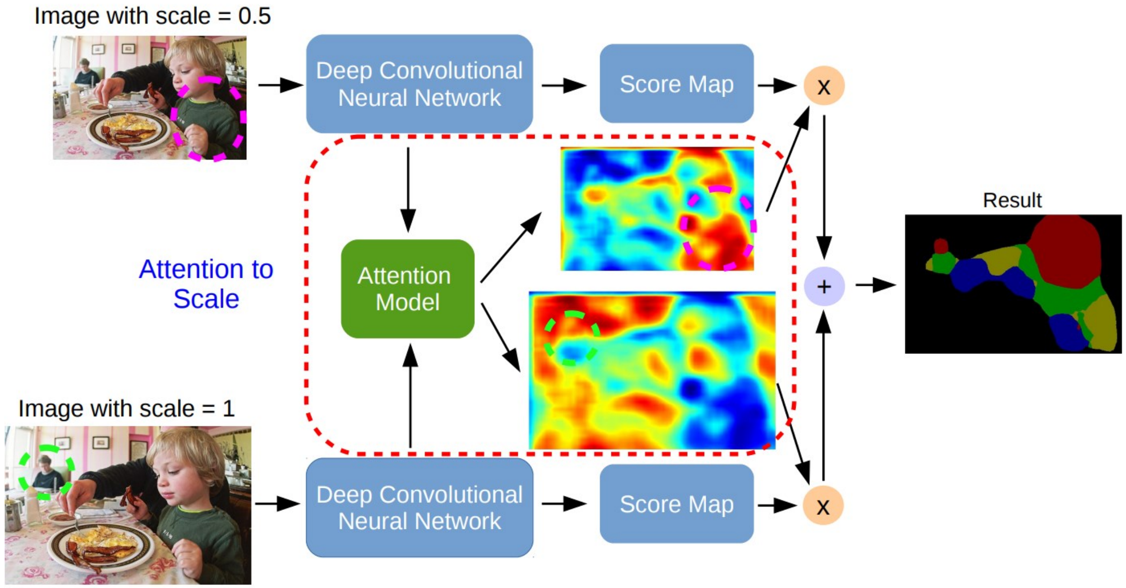
Objects Appears at Different Scales and Resolutions

Image Pyramid



Merging score maps (i.e., last layer output before SoftMax) for two scales.

Attention to Scale



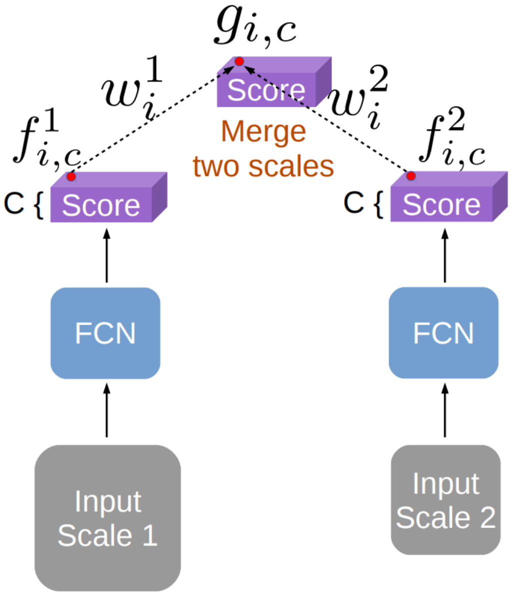
The attention model learns to put different weights on objects of different scales.

For example, the model learns to put large weights on the small-scale person (green dashed circle) for features from scale = 1, and large weights on the large-scale child (magenta dashed circle) for features from scale = 0.5.

The network component and the attention model are trained jointly.

Chen, L. C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3640-3649).

Attention to Scale



Naive Image Pyramid

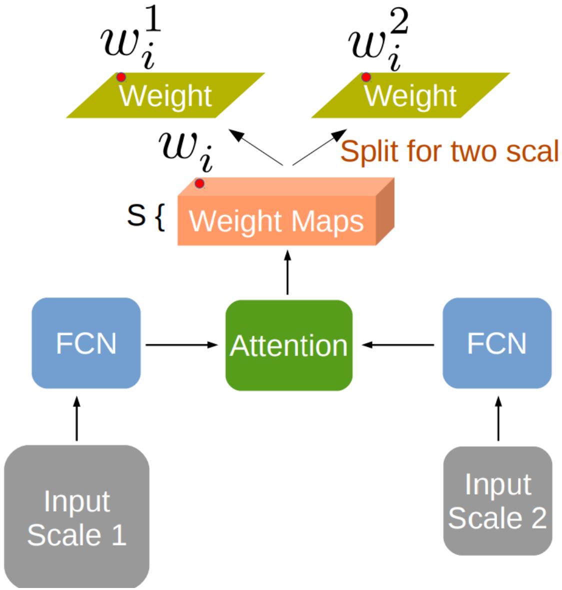
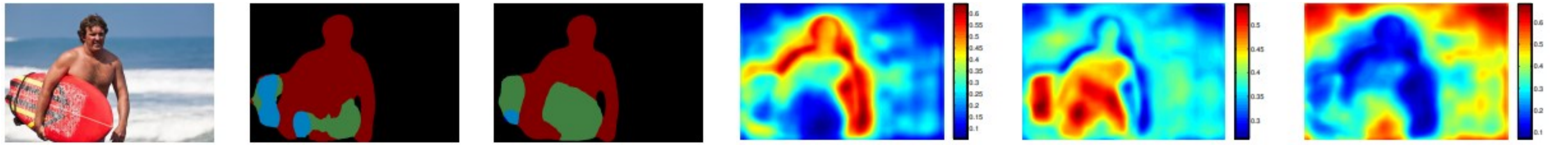
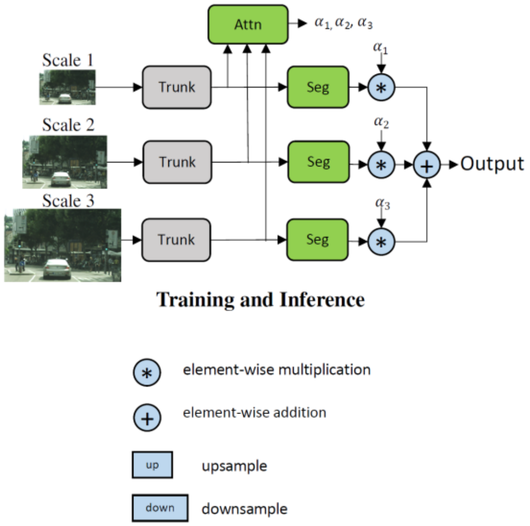


Image Pyramid with Attention

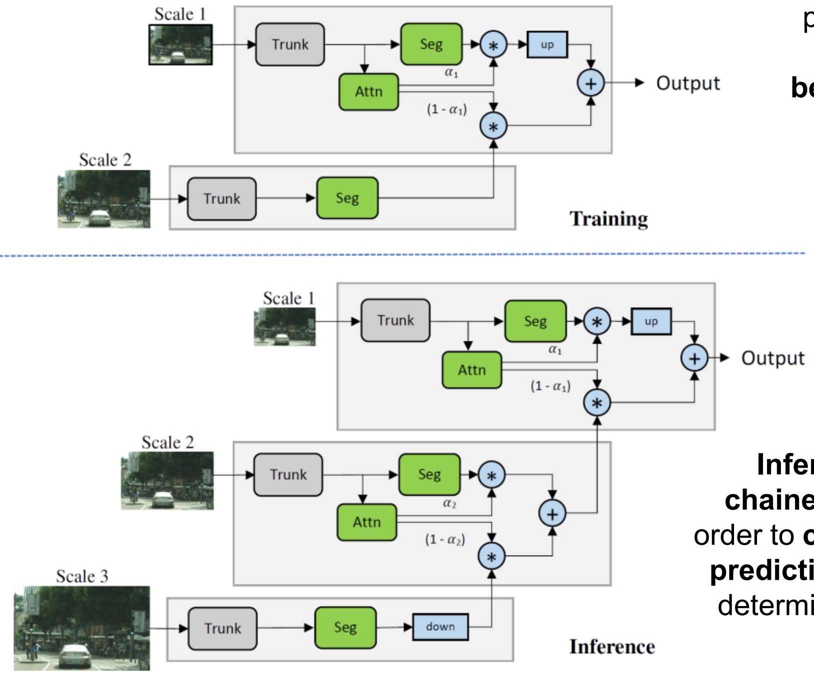
Input | FCN | Attention to Scale | Scale 1 Attention | Scale 0.75 Attention | Scale 0.5 Attention



Hierarchical Multiscale Attention



Architecture from *Attention to Scale,* where the attention for each scale is learned explicitly.



Hierarchical attention architecture.

An illustration of the training pipeline, whereby the network learns to predict **attention between adjacent scale pairs**.

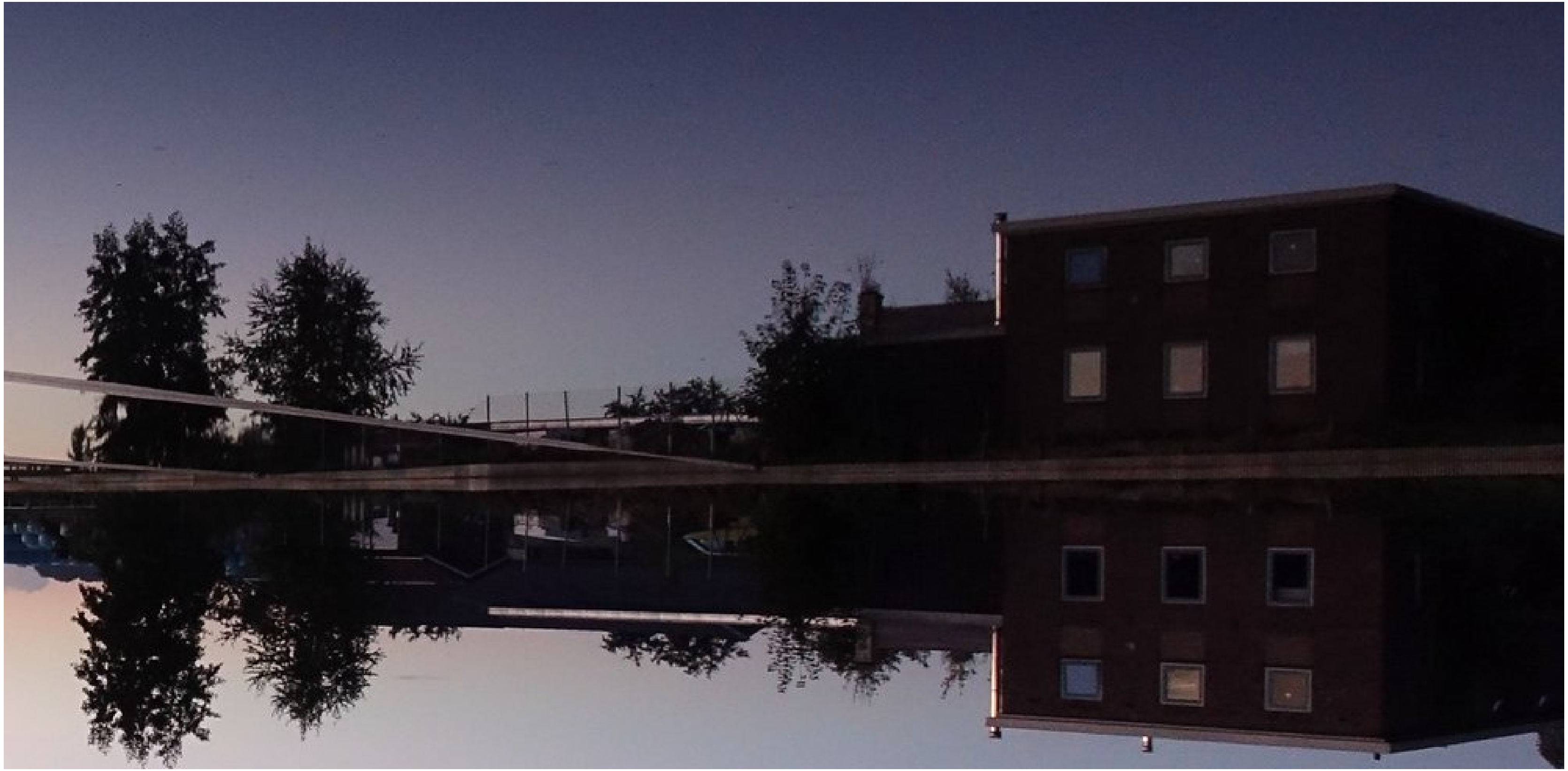
Training dataset is already **online augmented** thus the network sees a lot of scales.

Inference is performed in a **chained/hierarchical** manner in order to **combine multiple scales of predictions**. Lower scale attention determines the contribution of the next higher scale.

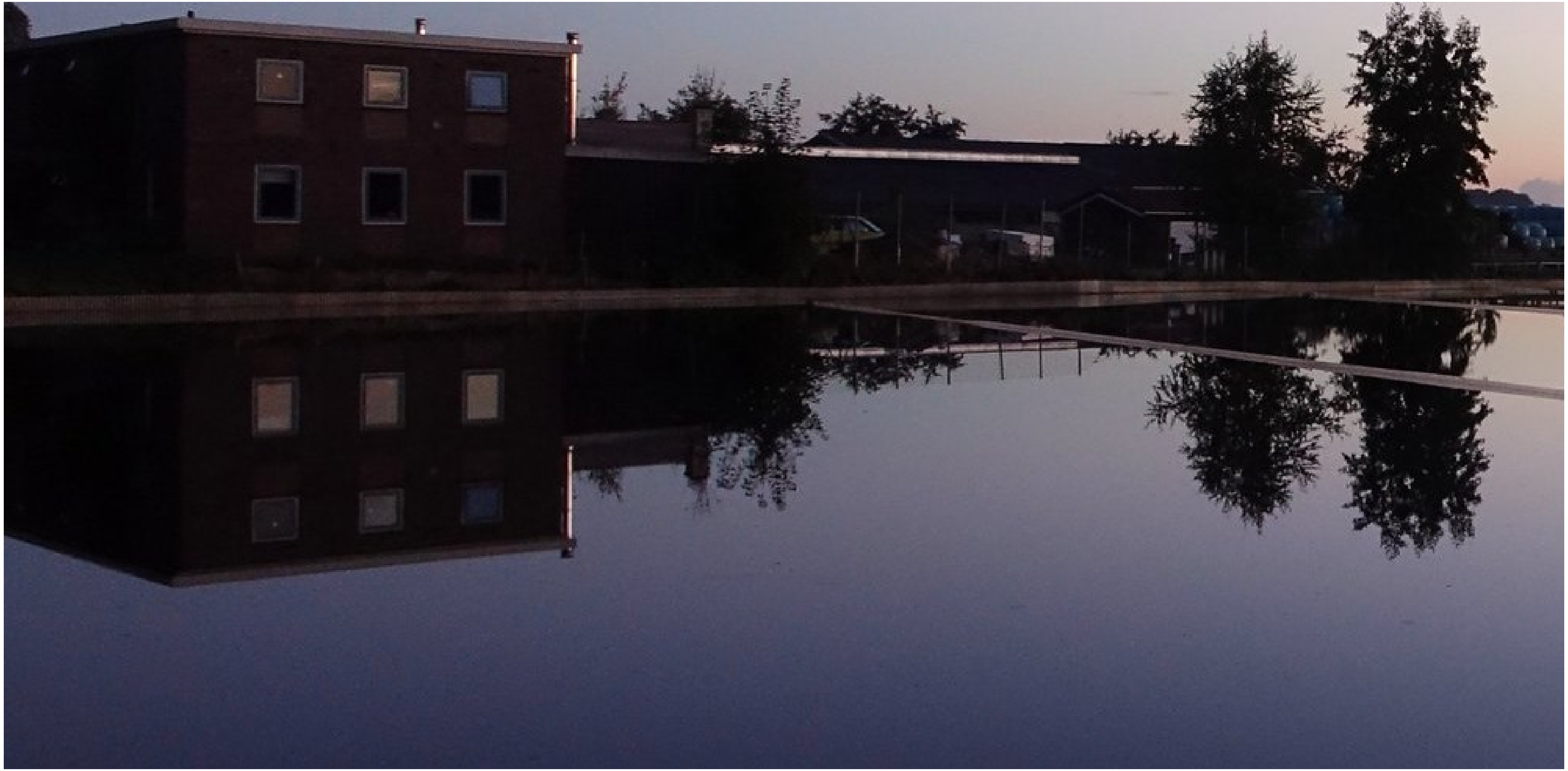
Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*.

Complex Scene Example

Complex Scene Example



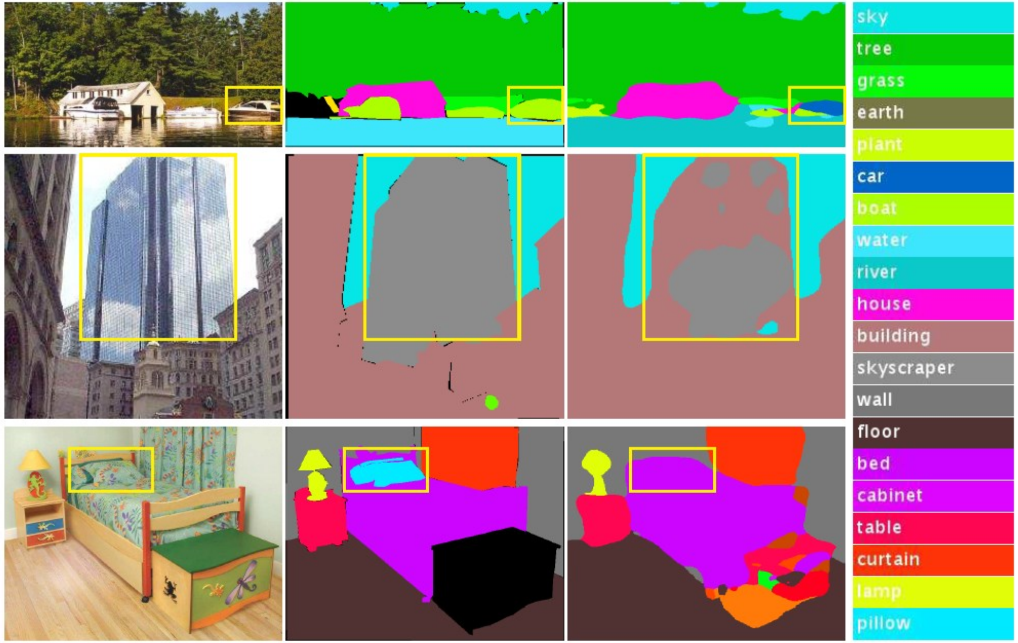
Complex Scene Example



Complex Scene Example

Context can be crucial to correctly classify an object.

Complex Scene Example



Input GT FCN

Mismatched Relationship Car or Boat? Objects are similar. However, context may guide us (there is water..)

Confusion Categories Building or Skyscraper? With a large enough receptive field we would understand that is the same object.

Inconspicuous Classes Pillow or Sheet? Objects may be really small, overlooking at the global scene category may fail to parse such objects

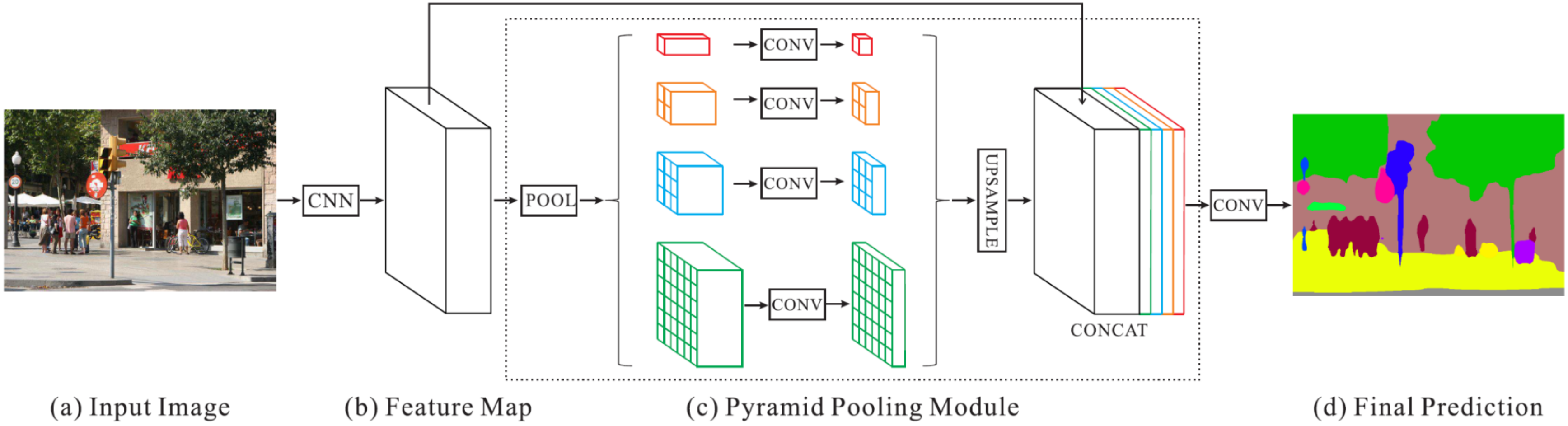
Many errors are partially or completely related to contextual relationship and global information for different receptive fields.

Complex Scene Examples

Images may contain objects at very different scales and resolutions.

We can extract different context information depending on image resolution and network receptive field. Context can be crucial to correctly classify an object (e.g., water not trees)

Encoding multi-scale semantics: Pyramidal Networks

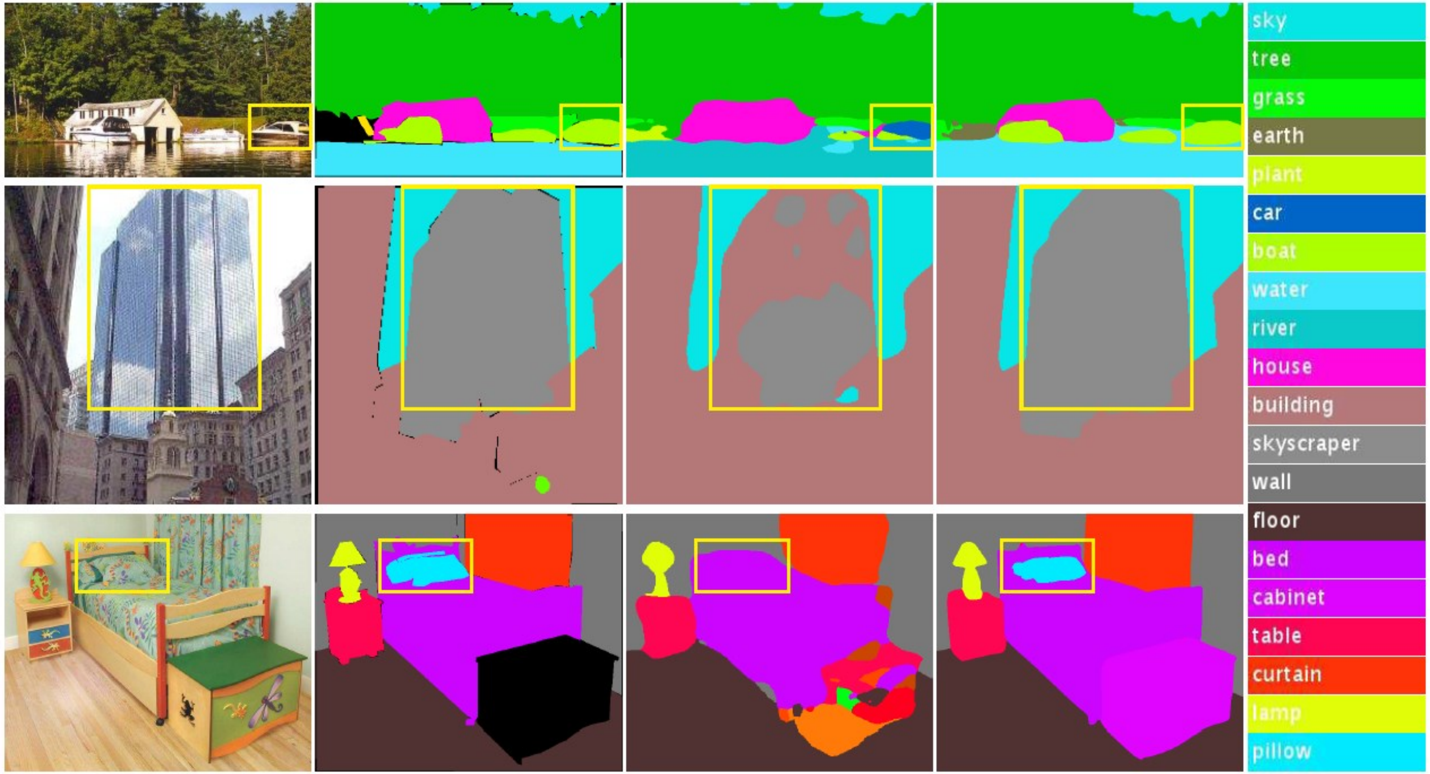


PSPNet

- Full Convolutional
- SPP

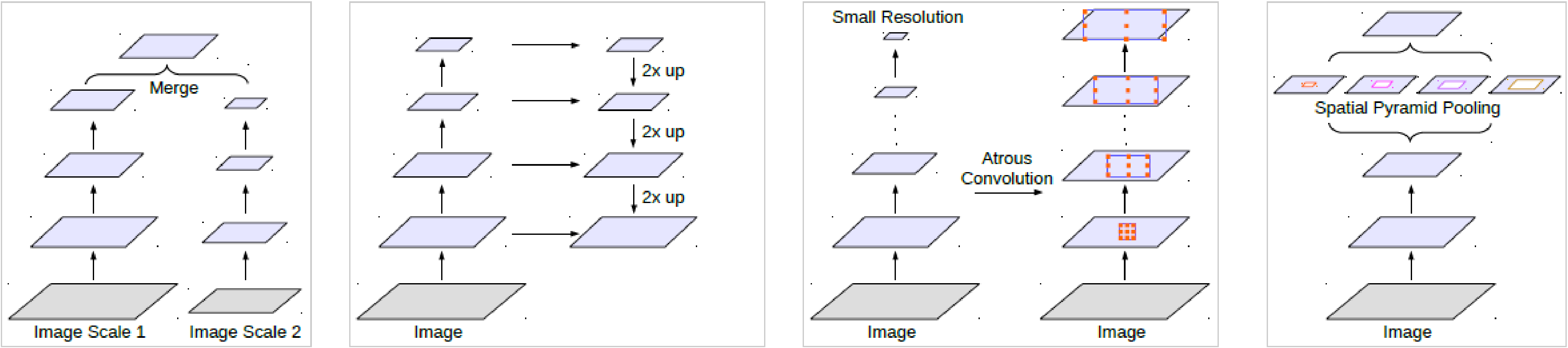
Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).

PSPNet vs FCN



Input GT FCN PSPNet

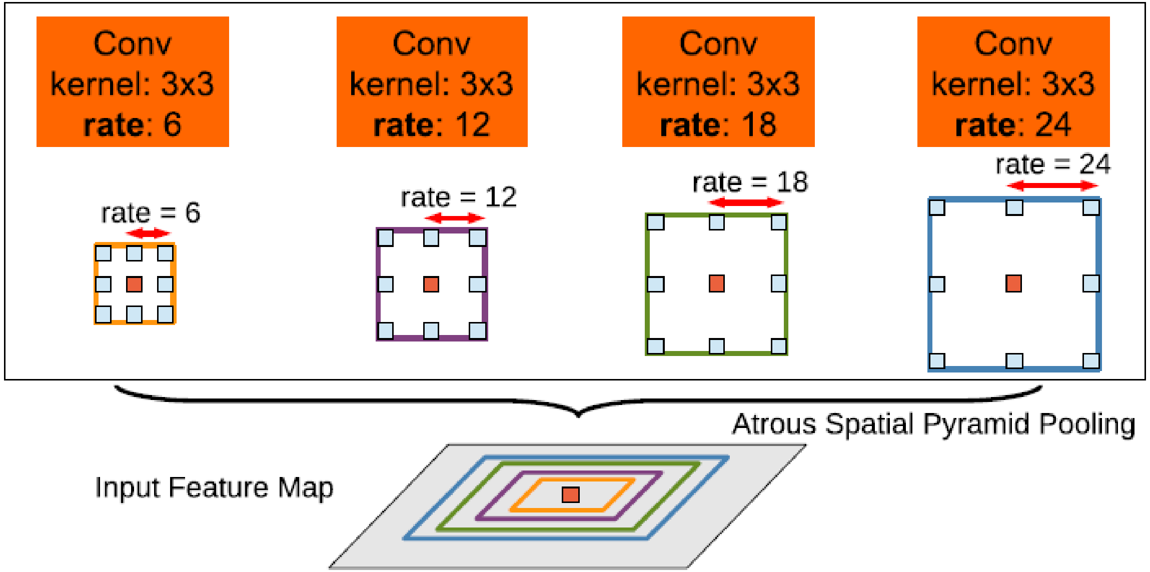
Recap: Alternative Architectures To Capture Multi-scale Context



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

Deeplab v2

Atrous Convolution + Spatial Pyramidal Pooling (SPP): ASPP



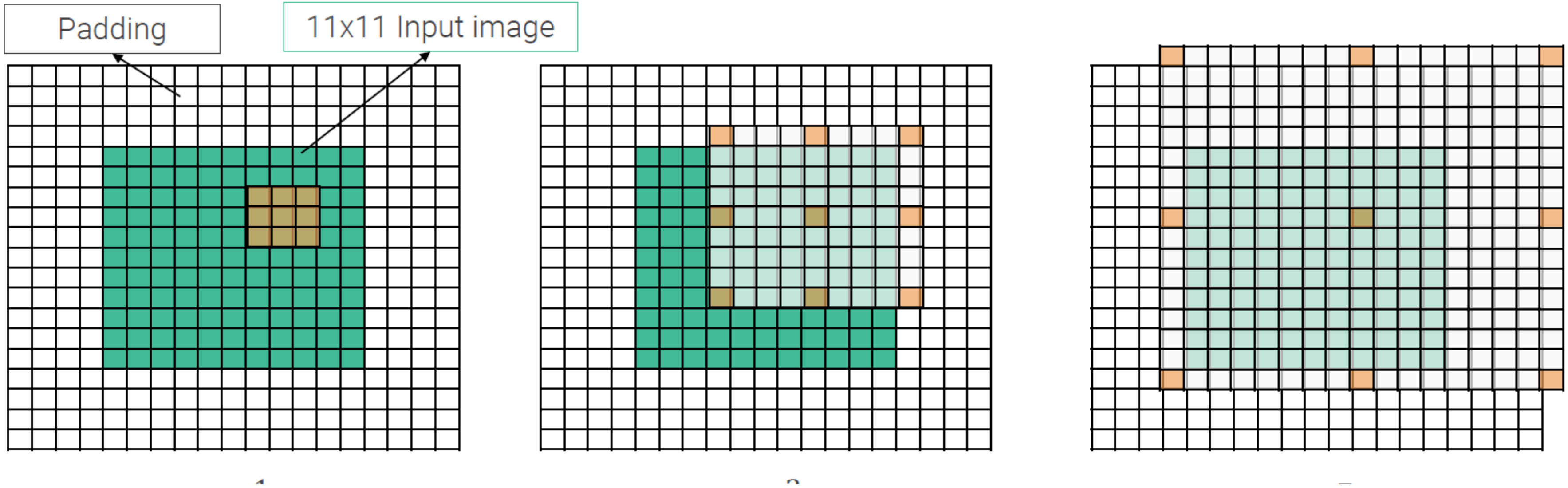
Use of **Atrous Spatial Pyramid Pooling (ASPP)**. The idea is to apply multiple atrous convolution with different sampling rates to the input feature map, and fuse together. As objects of the same class can have different scales in the image, ASPP helps to account for different object scales which can improve the accuracy.

Deeplabv2

- Deeplabv1 + ASPP

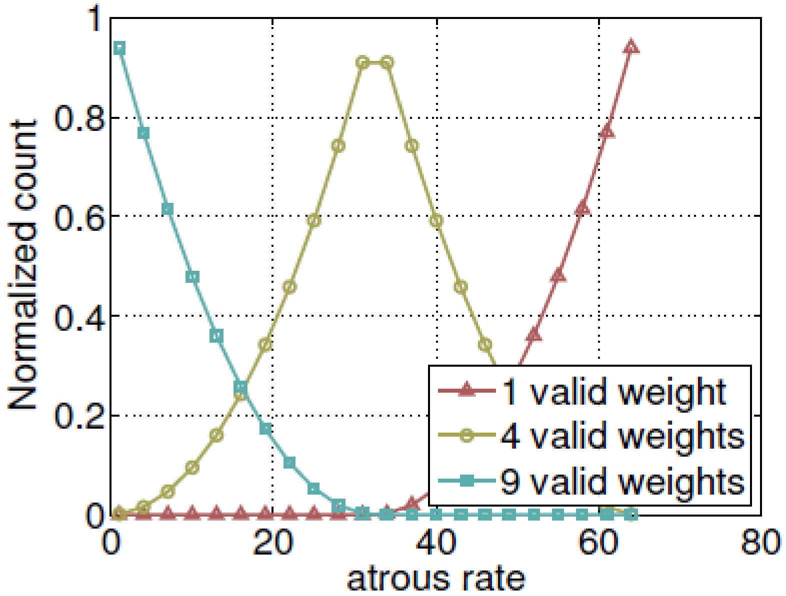
Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834-848.

Problem of atrous convolution with large rates



When the dilation rate grows, the number of positions where all the 9 weights of the kernel are used (i.e. are not multiplied by zero padding) shrinks.

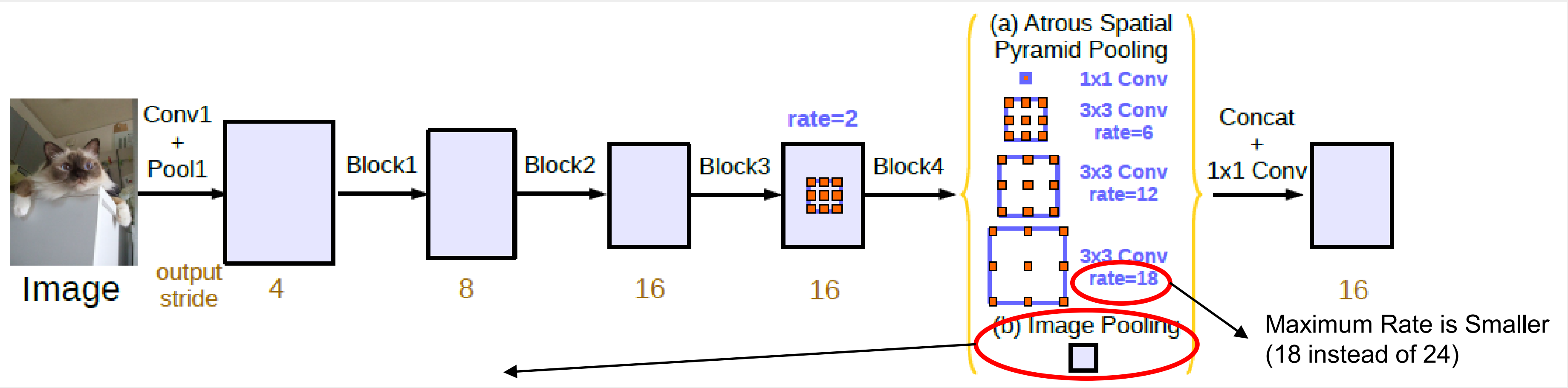
Problem of atrous convolution with large rates



Normalized counts of valid weights with a 3x3 filter on a 65x65 feature map as atrous rate varies. When atrous rate is small, all the 9 filter weights are applied to most of the valid region on feature map, while atrous rate gets larger, the 3x3 filter degenerates to a 1x1 filter since only the center weight is effective.

Deeplab v3
Encoder Modifications and Training Improvements

Encoder



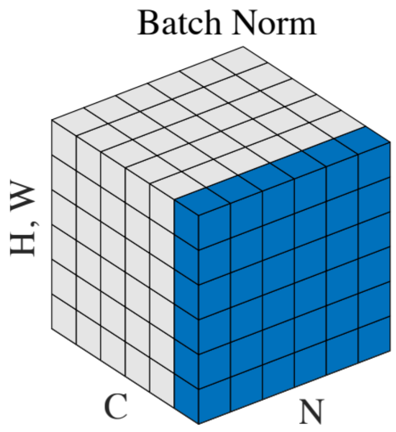
Instead of increasing atrous convolutions rate, they do **image pooling**.

Deeplab v3

Encoder Modifications and Training Improvements

Training Improvements

Added Batch Normalization. Regularize training and better convergence. Pretrain with larger batch (16) and larger stride (16) to learn better statistics.



Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

Deeplab v3

Encoder Modifications and Training Improvements

Training Improvements

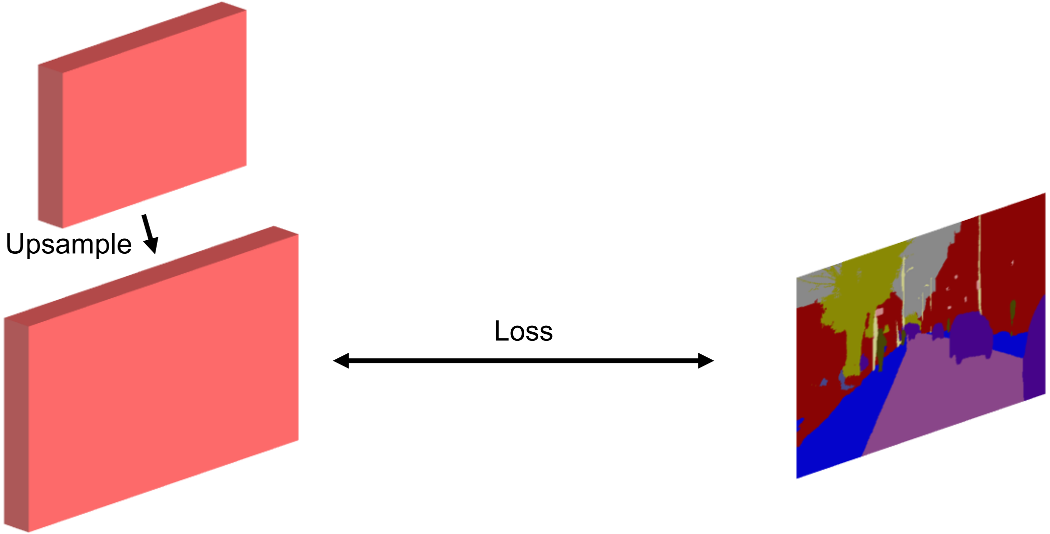
- *Added Batch Normalization.* Regularize training and better convergence. Pretrain with larger batch (16) and larger stride (16) to learn better statistics.
- *Large Crop Size.* To capture enough context information, we need a large crop size

Deeplab v3

Encoder Modifications and Training Improvements

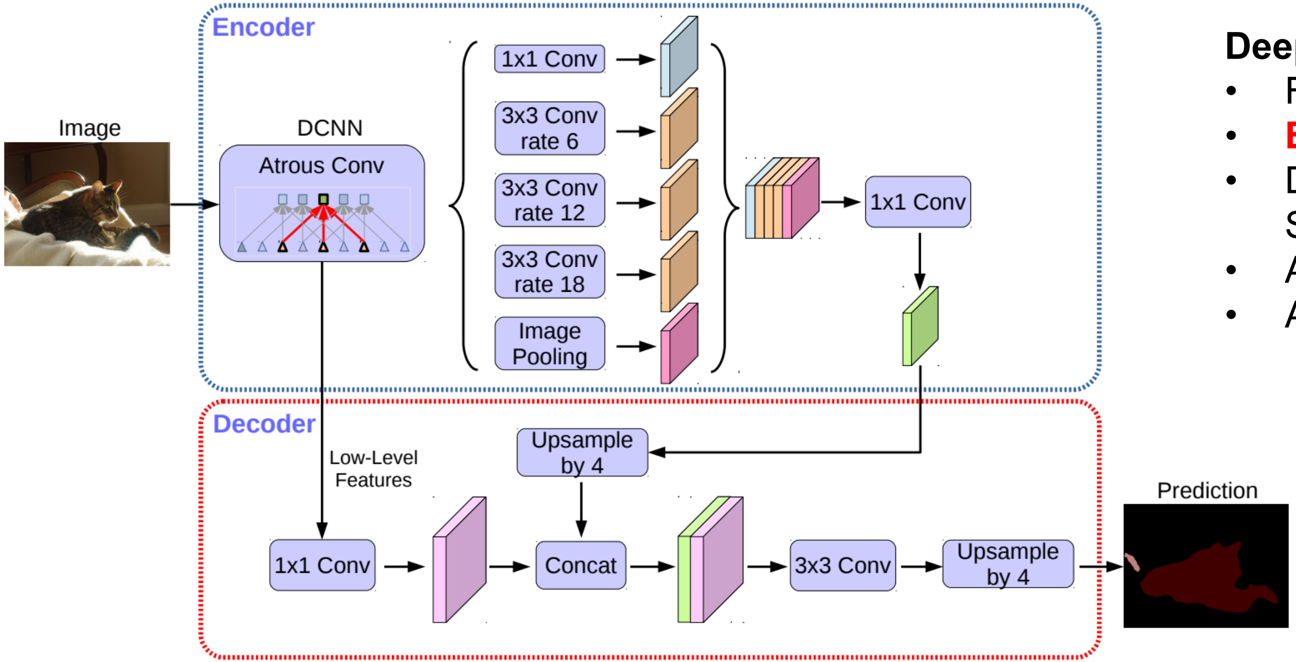
Training Improvements

- *Added Batch Normalization.* Regularize training and better convergence. Pretrain with larger batch (16) and larger stride (16) to learn better statistics.
- *Large Crop Size.* To capture enough context information, we need a large crop size
- *Upsampling Logits.* Previous Deeplabs downsample the GT. From Deeplabv3 upsample the logits to match GT resolution (more high-resolution information in the gradient).




DeepLab v3

Deeplab v3+
Adding Decoder



Deeplab v3+
- Fully-convolutional
- **Encoder-Decoder**
- Decoder: Up-Convolutions + Skip Connections
- Atrous Convolutions
- ASPP (with Image Pooling)

Still one of the most popular semantic segmentation network!

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 801-818).

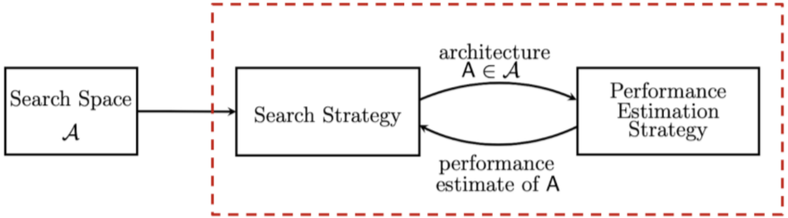
Deeplab v3+
Adding Decoder

Image | w/ BU | w/ Decoder

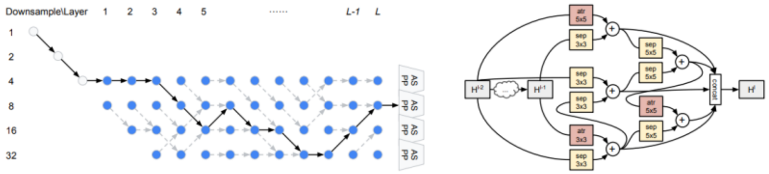
Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 801-818).

Recent Advances: NAS and Transformers

Neural Architecture Search (NAS)

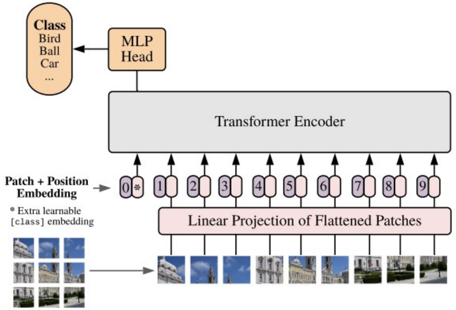


Auto-DeepLab

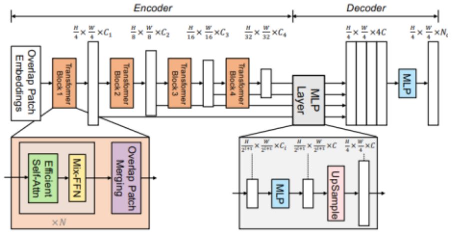


Liu, C., Chen, L. C., Schroff, F., Adam, H., Hua, W., Yuille, A. L., & Fei-Fei, L. (2019). Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 82-92).

Vision Transformer



SegFormer



Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*.

Quantitative Results

End 2014

	Method	Encoder	Params ↓	ADE20K Flops ↓	ADE20K FPS ↑	ADE20K mIoU ↑	Cityscapes Flops ↓	Cityscapes FPS ↑	Cityscapes mIoU ↑
Real-Time	FCN [1]	MobileNetV2	9.8	39.6	64.4	19.7	317.1	14.2	61.5
	ICNet [11]	-	-	-	-	-	-	30.3	67.7
	PSPNet [17]	MobileNetV2	13.7	52.9	57.7	29.6	423.4	11.2	70.2
	DeepLabV3+ [20]	MobileNetV2	15.4	69.4	43.1	34.0	555.4	8.4	75.2
	SegFormer (Ours)	MiT-B0	**3.8**	**8.4**	**50.5**	**37.4**	125.5	15.2	**76.2**
				-	-	-	51.7	26.3	75.3
				-	-	-	31.5	37.1	73.7
				-	-	-	**17.7**	**47.6**	71.9
Non Real-Time	FCN [1]	ResNet-101	68.6	275.7	14.8	41.4	2203.3	1.2	76.6
	EncNet [24]	ResNet-101	**55.1**	218.8	14.9	44.7	1748.0	1.3	76.9
	PSPNet [17]	ResNet-101	68.1	256.4	15.3	44.4	2048.9	1.2	78.5
	CCNet [41]	ResNet-101	68.9	278.4	14.1	45.2	2224.8	1.0	80.2
	DeeplabV3+ [20]	ResNet-101	62.7	255.1	14.1	44.1	2032.3	1.2	80.9
	OCRNet [23]	HRNet-W48	70.5	164.8	**17.0**	45.6	1296.8	**4.2**	81.1
	GSCNN [35]	WideResNet38	-	-	-	-	-	-	80.8
	Axial-DeepLab [74]	AxialResNet-XL	-	-	-	-	2446.8	-	81.1
	Dynamic Routing [75]	Dynamic-L33-PSP	-	-	-	-	**270.0**	-	80.7
	Auto-Deeplab [50]	NAS-F48-ASPP	-	-	-	44.0	695.0	-	80.3
	SETR [7]	ViT-Large	318.3	-	5.4	50.2	-	0.5	82.2
	SegFormer (Ours)	MiT-B4	64.1	**95.7**	15.4	51.1	1240.6	3.0	83.8
	SegFormer (Ours)	MiT-B5	84.7	183.3	9.8	**51.8**	1447.6	2.5	**84.0**

End 2021

SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo
The University of Hong Kong   Nanjing University   NVIDIA   Caltech

Thank you for your attention!