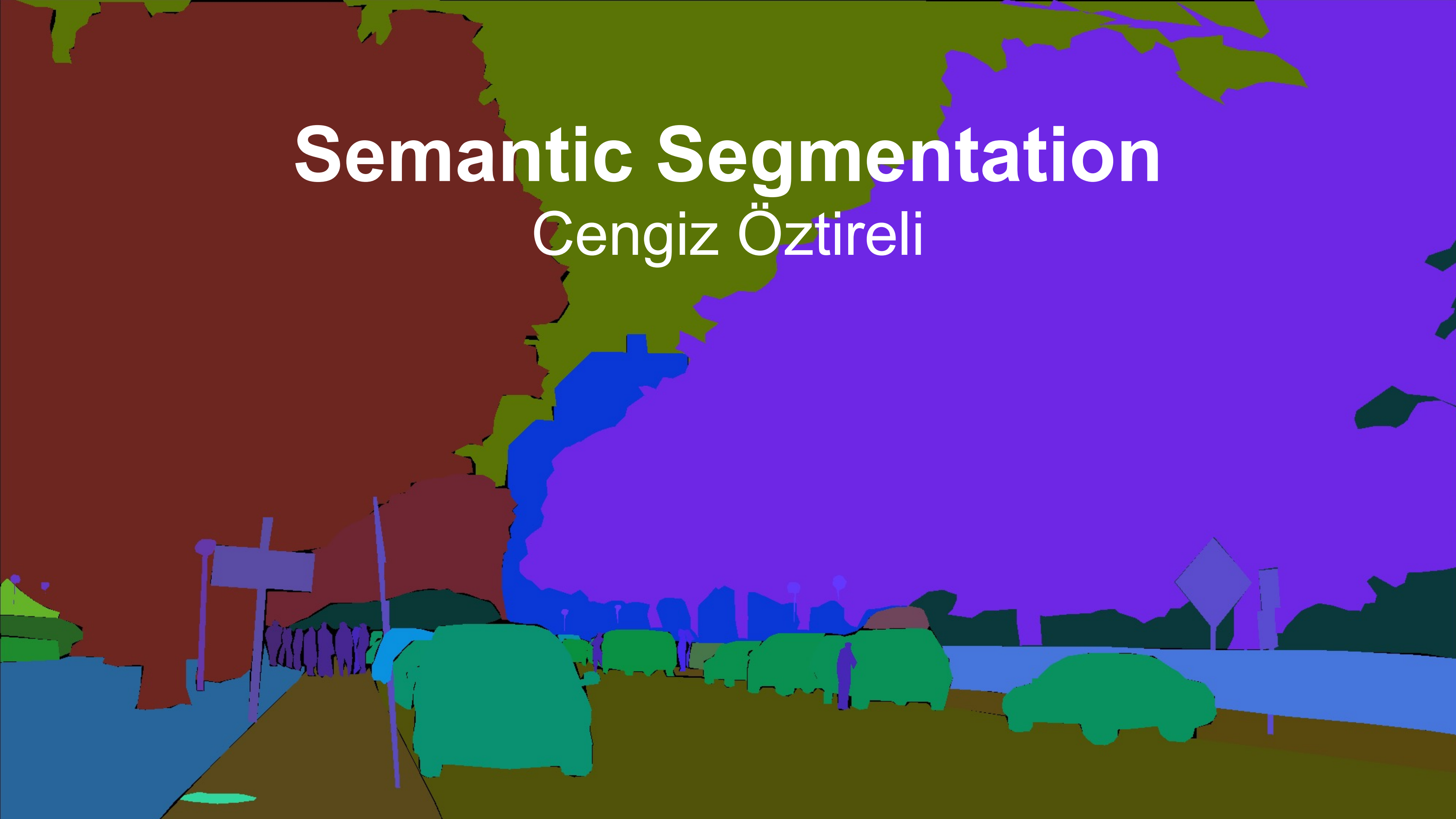
Semantic Segmentation

Cengiz Öztireli

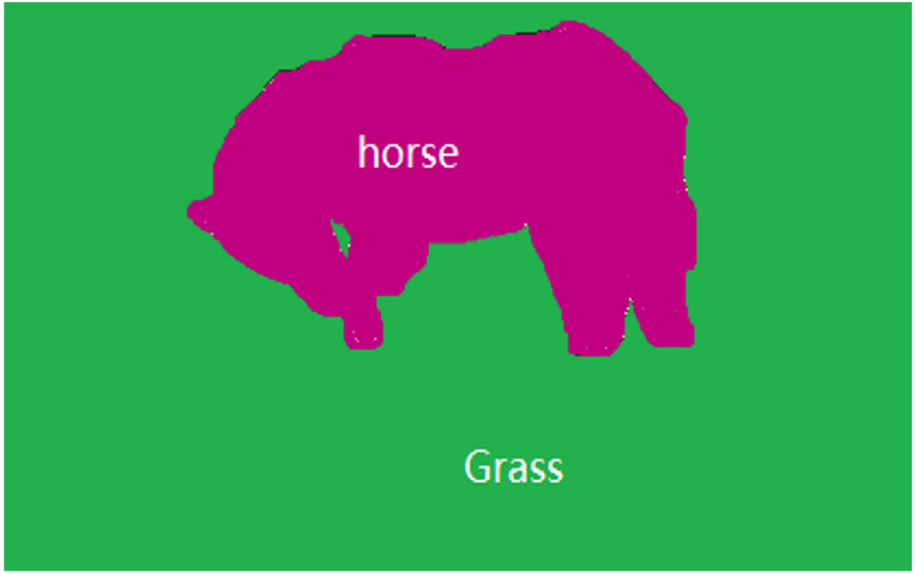
Semantic Segmentation

Cengiz Öztireli



Semantic Segmentation

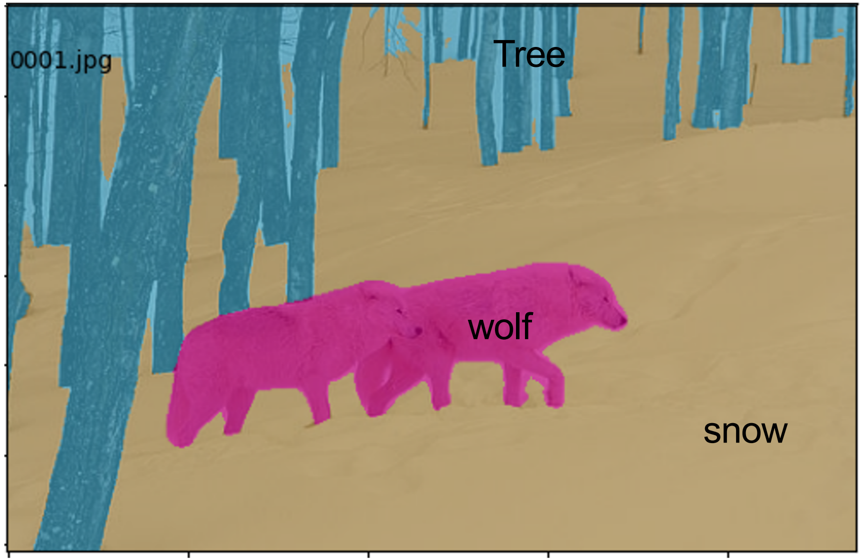
Pixel level classification Problem



Semantic Segmentation

- Label each pixel with a pre-defined class
- Dense prediction problem
- Does not differentiate instances

Semantic Segmentation

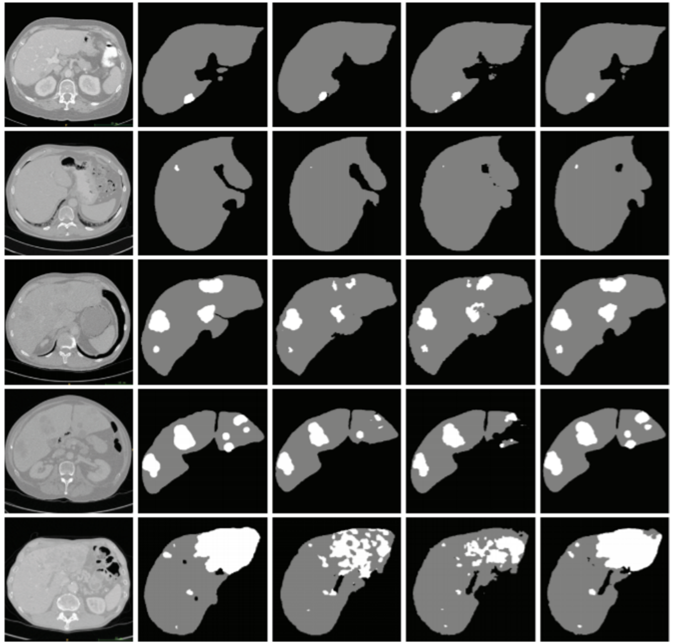


- Label each pixel with a pre-defined class
- Dense prediction problem
- Do not differentiate instances

UNIVERSITY OF CAMBRIDGE

Applications

Autonomous Driving



Liver Tumor Segmentation

Representation

Input

segmented

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

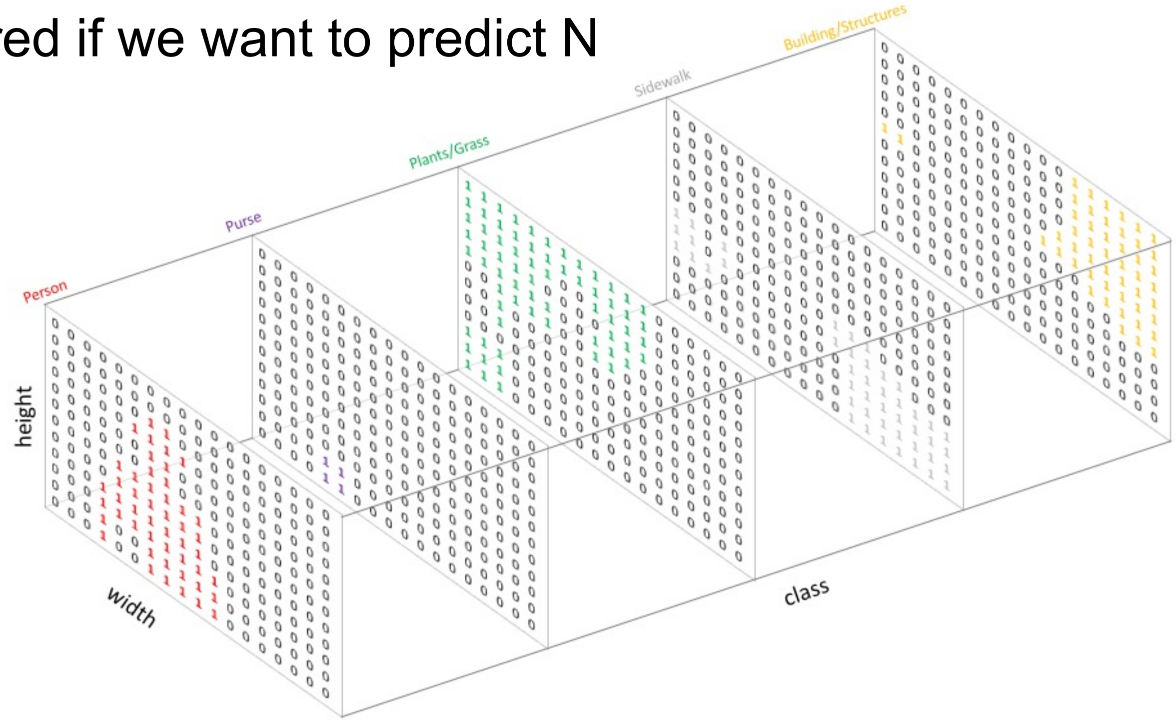
3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	1	1	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	1	1	1	1	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	1	1	3	3	3	5	5	5	5	5	5	5
5	5	3	3	3	3	1	1	3	3	5	5	5	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	4	4	4	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	4	4	4	4	4	5	5	5
4	4	4	1	1	1	1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

Semantic labels

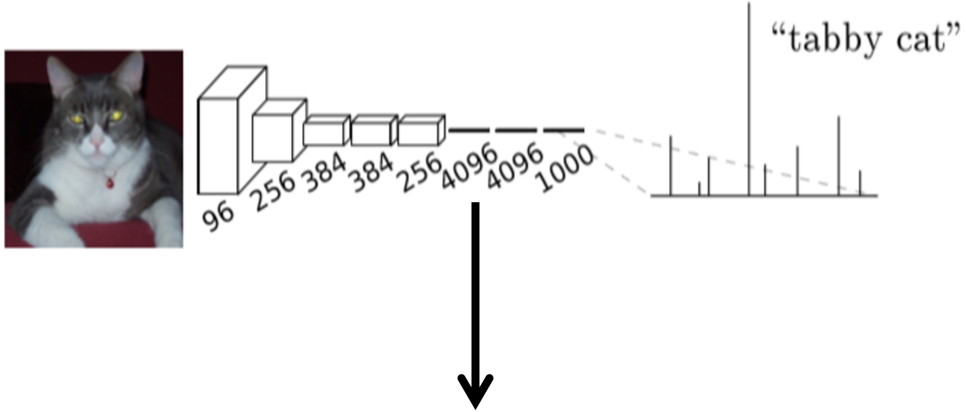
Representation

The target will be transferred to one-hot labels.

N channels will be required if we want to predict N classes.

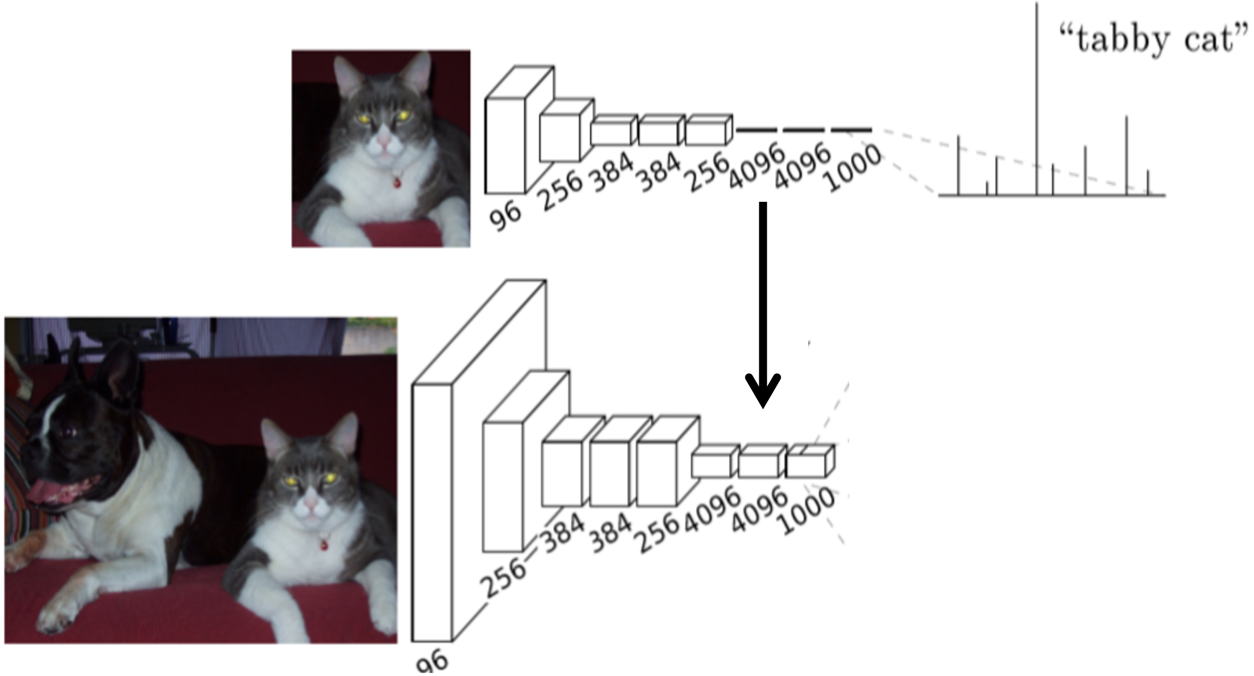


Fully Convolutional Networks



Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on CVPP. 2015: 3431-3440

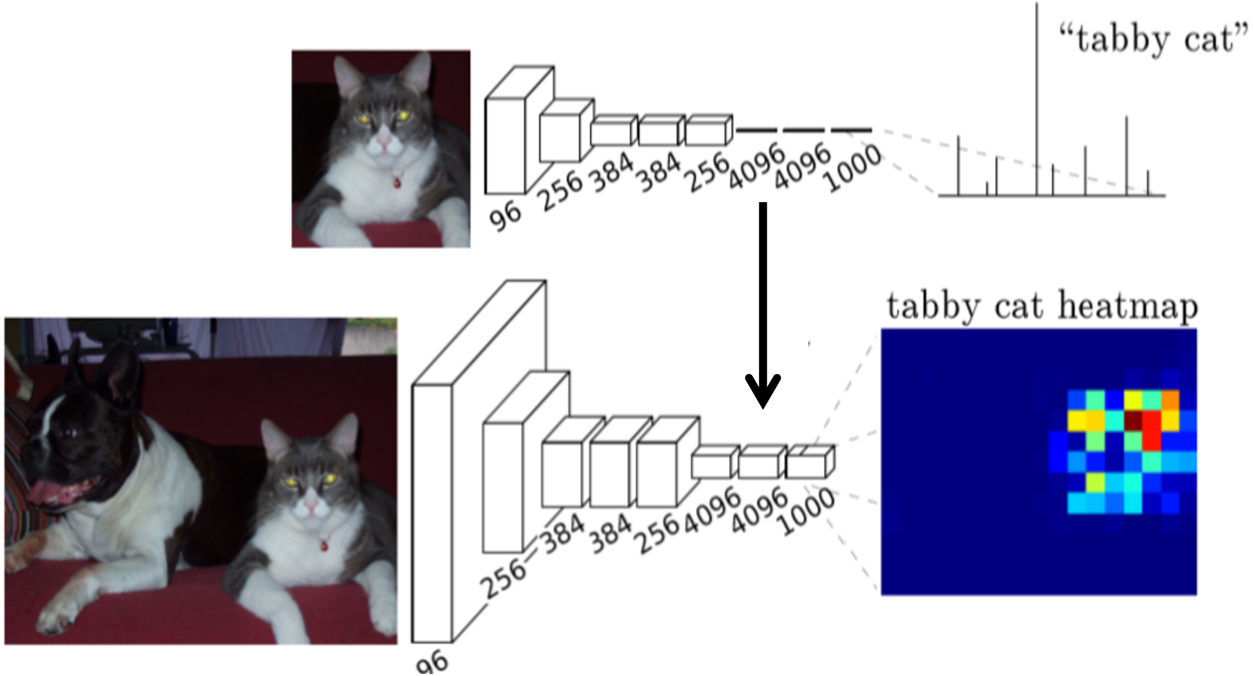
Fully Convolutional Networks



UNIVERSITY OF CAMBRIDGE

Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on CVPP. 2015: 3431-3440

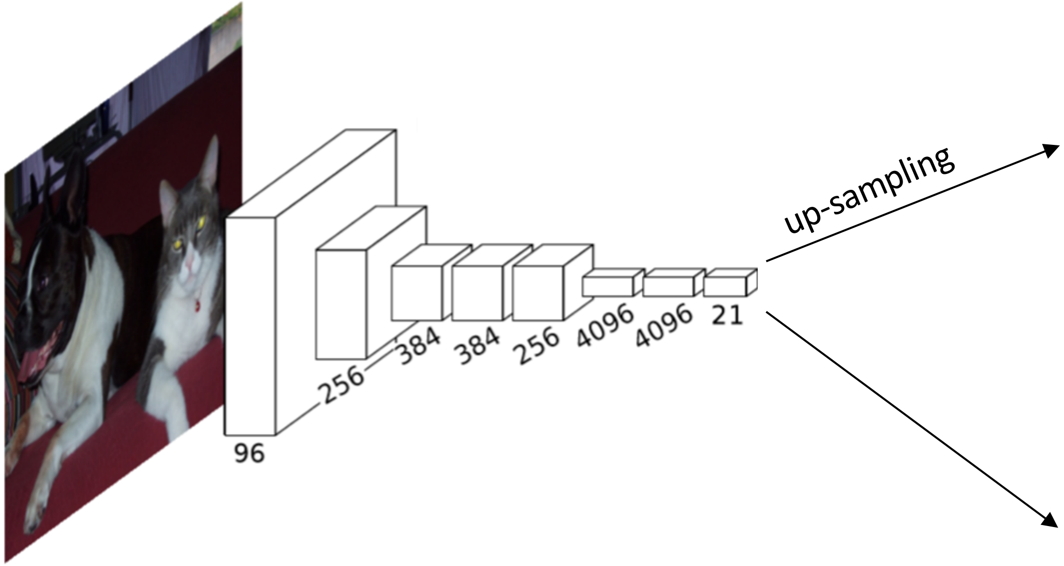
Fully Convolutional Networks



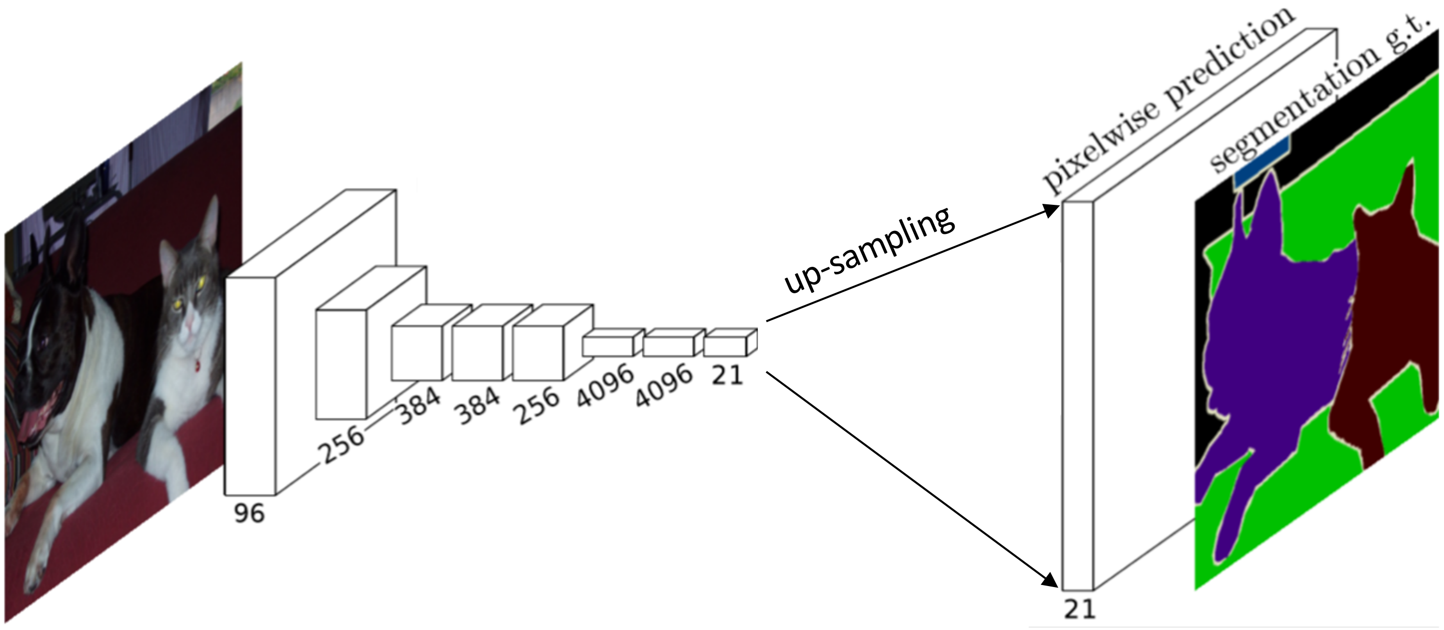
Coarse feature map

Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on CVPP. 2015: 3431-3440

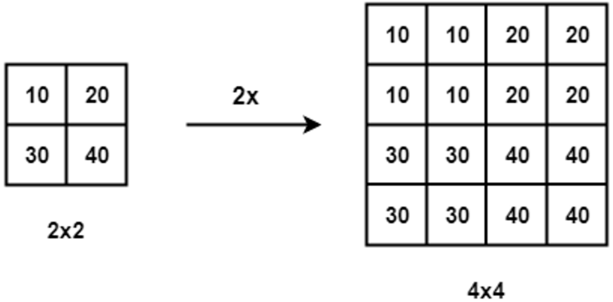
Fully Convolutional Networks



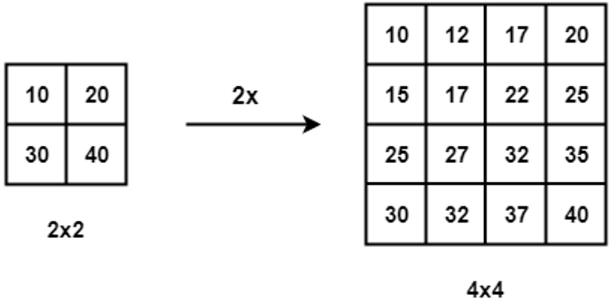
Fully Convolutional Networks



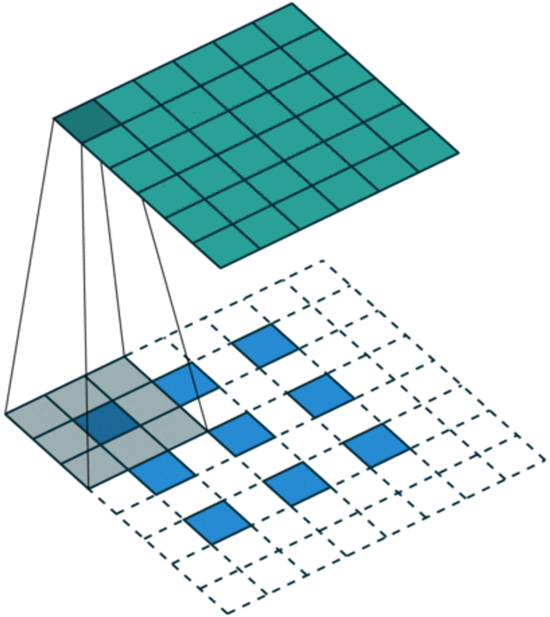
Upsampling



Nearest neighbor interpolation

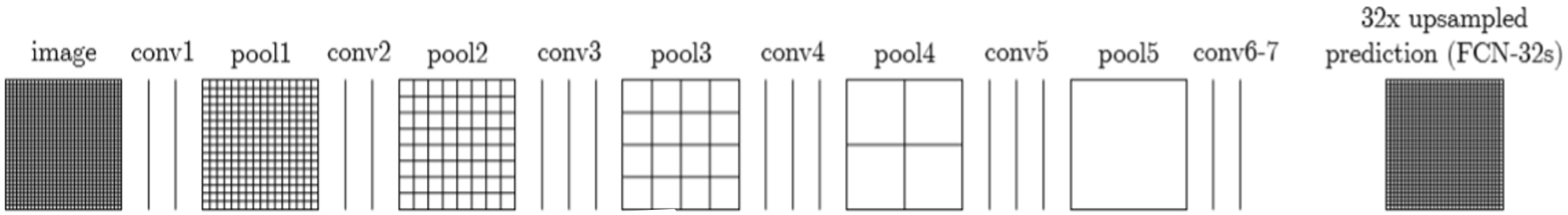


Bilinear interpolation



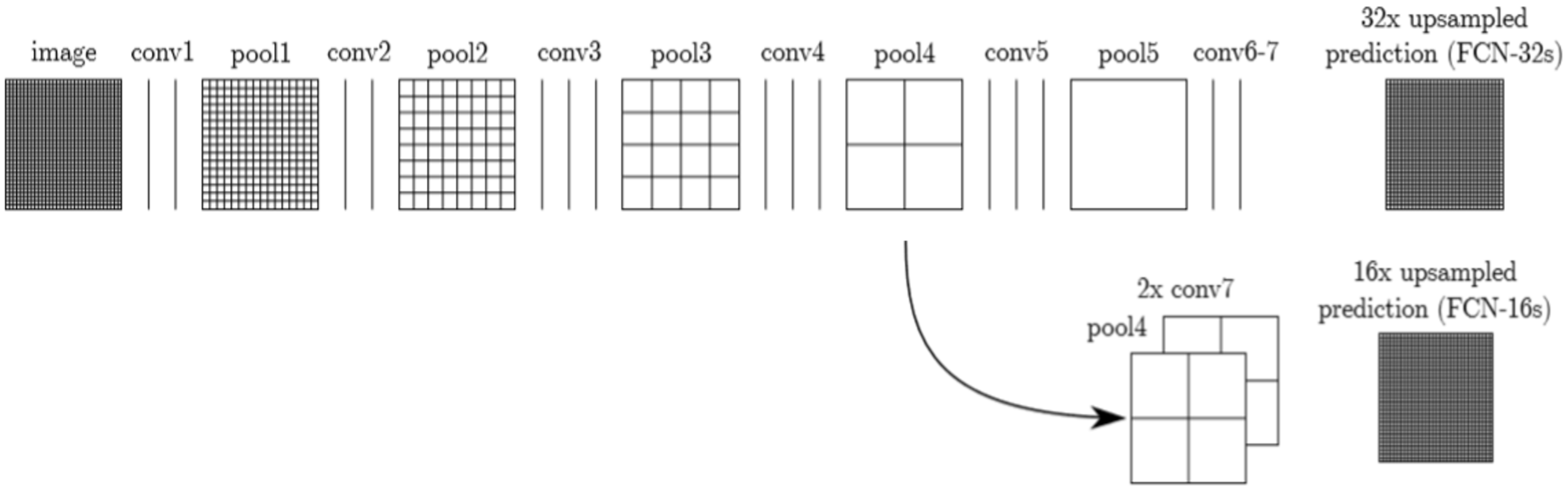
Deconvolution

Fully Convolutional Networks



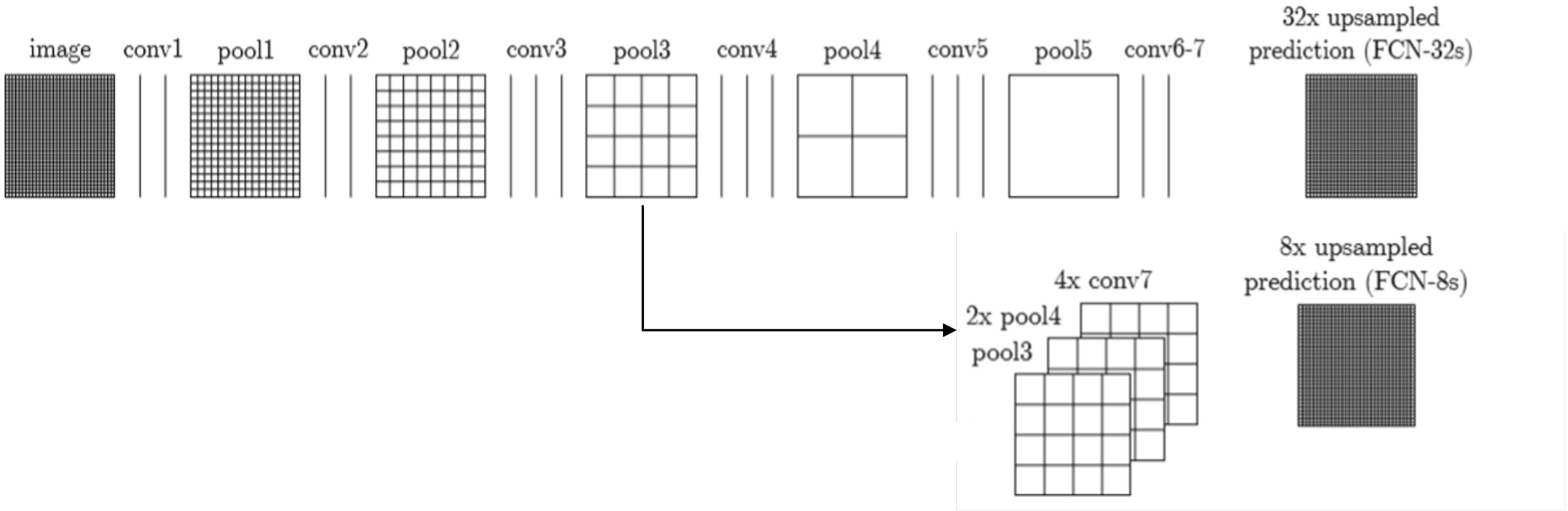
Upsampling first version with fix the parameters

Fully Convolutional Networks



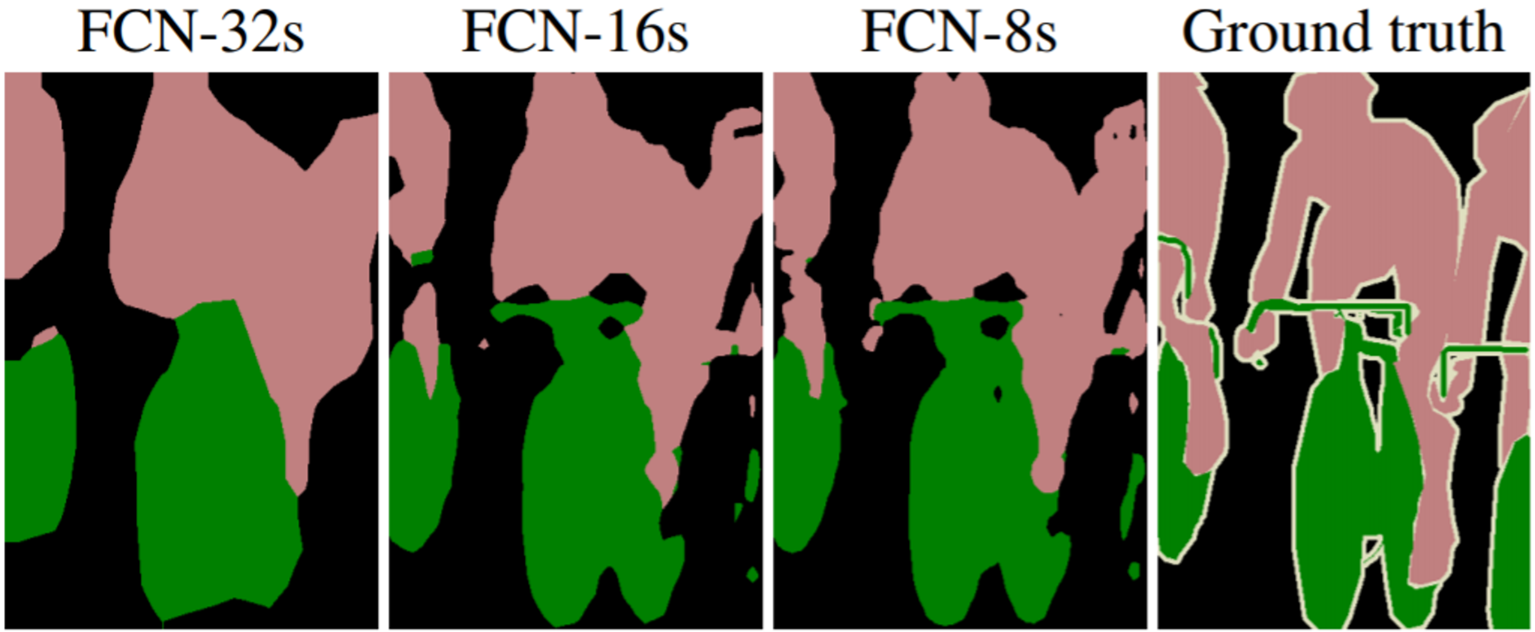
Upsampling second version with learnable parameters

Fully Convolutional Networks



Upsampling second version with learnable parameters

Results



Loss Function

- Pixel cross entropy

$$\sum_{i}^{W*H} \sum_{c=1}^{C} y_{c,i} * \log(y_{c,i}^{'})$$

Loss Function

- Pixel cross entropy

spatial size

the number of the classes

Prediction value

$$\sum_{i}^{W*H} \sum_{c=1}^{C} y_{c,i} * \log(y'_{c,i})$$

Target

index of the spatial location

index of the class

Loss Function

$$\sum_{i}^{W*H} \sum_{c=1}^{C} y_{c,i} * \log(y'_{c,i})$$

Logits on position i: 0.1 0.1 0.1 0.02 0.03 0.5 0.15
Target on position i : 0 0 0 0 0 1 0

$$\ell_i = 1 * \log(0.5)$$

Loss Function

- Other loss functions:
  weighted cross entropy
  - Add different weights for different classes
  - Widely-used in long-tail class distributions

$$\sum_{i}^{W*H}\sum_{c=1}^{C} w_c * y_{c,i} * \log(y^{'}_{c,i})$$

Loss Function

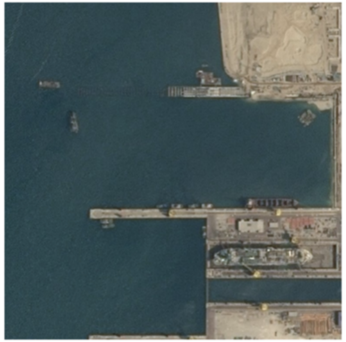
- Other loss functions:
  Dice coefficient
  - Focus more on small regions
  - Widely-used in medical image processing
  - Dice loss for the class c:

$$1 - 2 * \frac{\sum_{i}^{W*H} y_{c,i} * y_{c,i}^{'}}{\sum_{i}^{W*H} (y_{c,i})^2 + \sum_{i}^{W*H} (y_{c,i}^{'})^2}$$

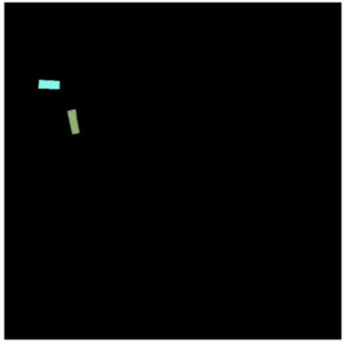
Repeat for all classes and average the score

Evaluation Metric

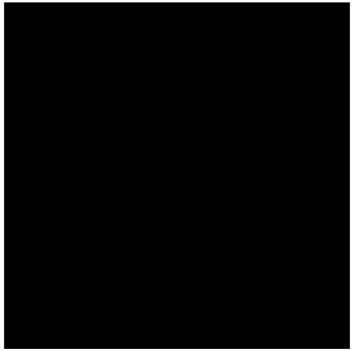
Pixel accuracy: the percent of pixels in your image that are classified correctly



Input



Ground truth



Prediction

Evaluation Metric

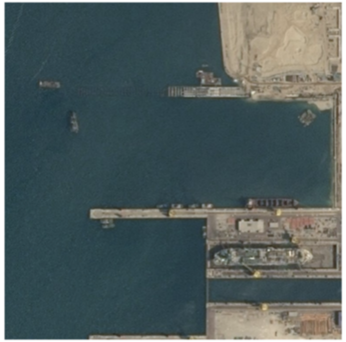
IoU: the area of overlap between the predicted segmentation and the ground truth divided by the area of the union between the predicted segmentation and the ground truth

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

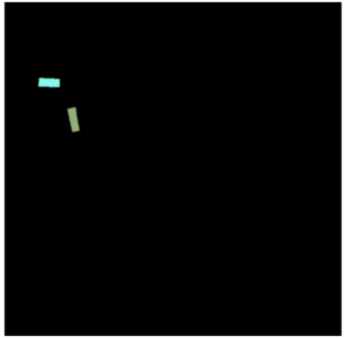
Evaluation Metric

mIoU: the area of overlap between the predicted segmentation and the ground truth divided by the area of the union between the predicted segmentation and the ground truth

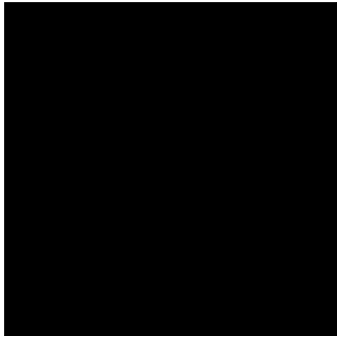
Assume we calculate a two classes mIoU:
(0 / 5 + 95 / 100) / 2
= 47.5



Input



Ground truth



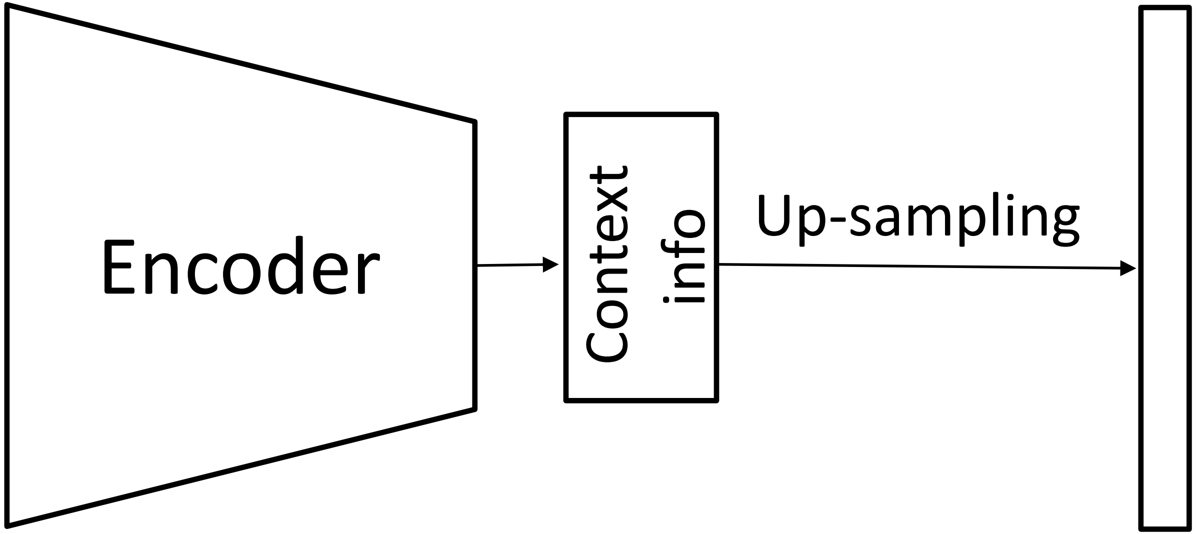
Prediction

UNIVERSITY OF CAMBRIDGE

General Problem

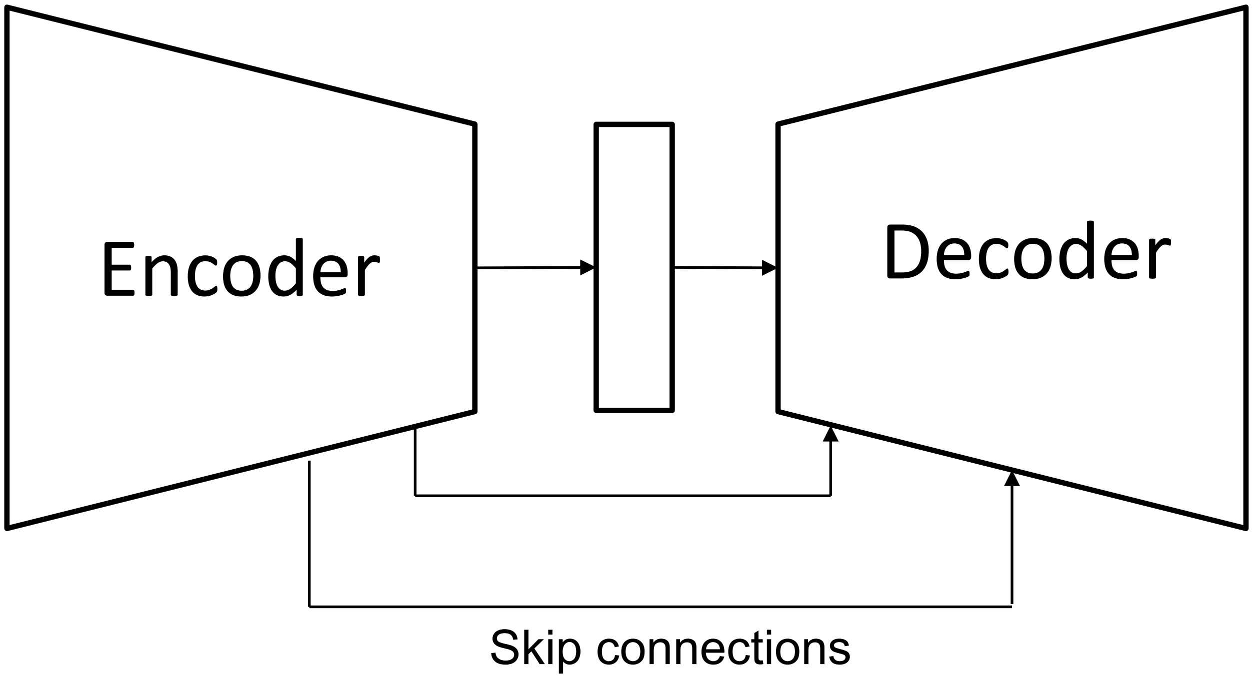
- Classification: global information
  - We need to get rich context information for semantic classes
- Localization: local information
  - We need to predict fine-grained results for each pixel.

General Structure



Focus on solving the classification problem

General Structure



Fine-grained localization

Outline

Baseline

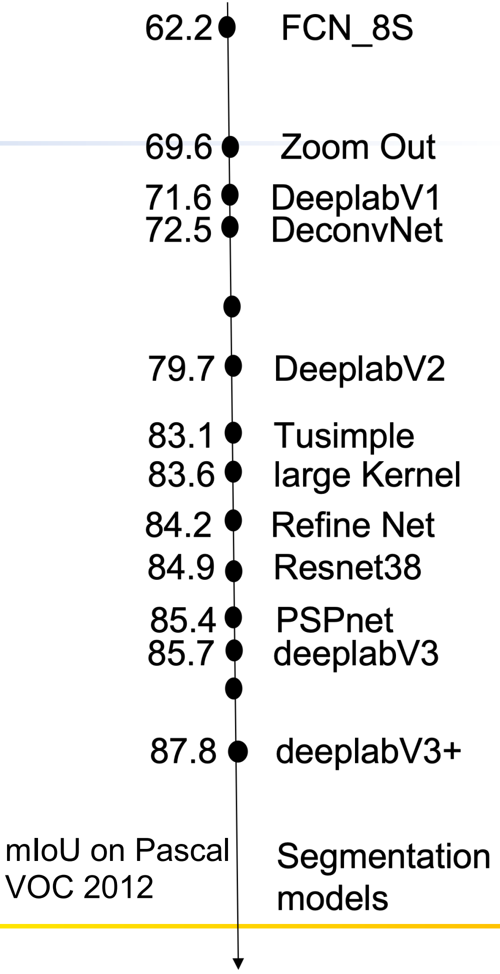
Fully convolutional network

Rich context information (Encoder)

Multi scale & Enlarge the reception field:
Deeplab, DenseNet, PSPNet

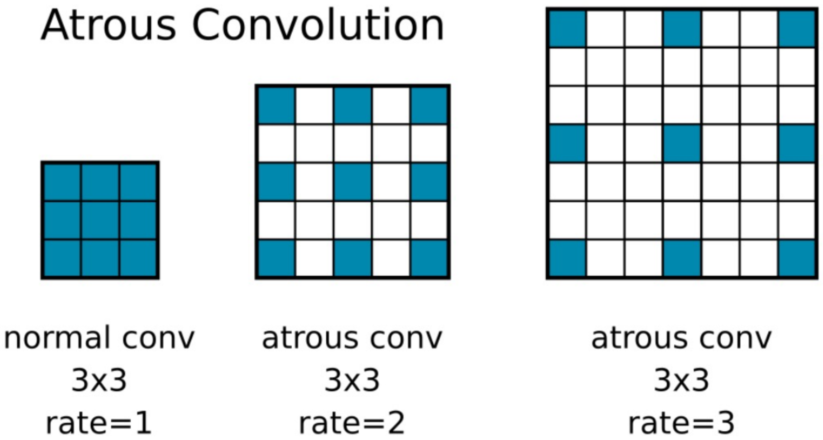
Decoder

DeconvNet, SegNet, Tusimple, DeelabV3+



Rich Context Information (Encoder)

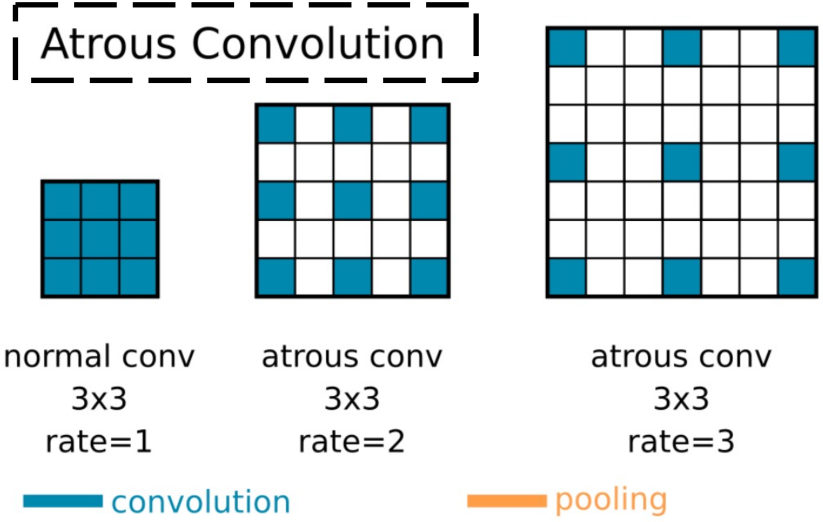
Deeplab v1-v3

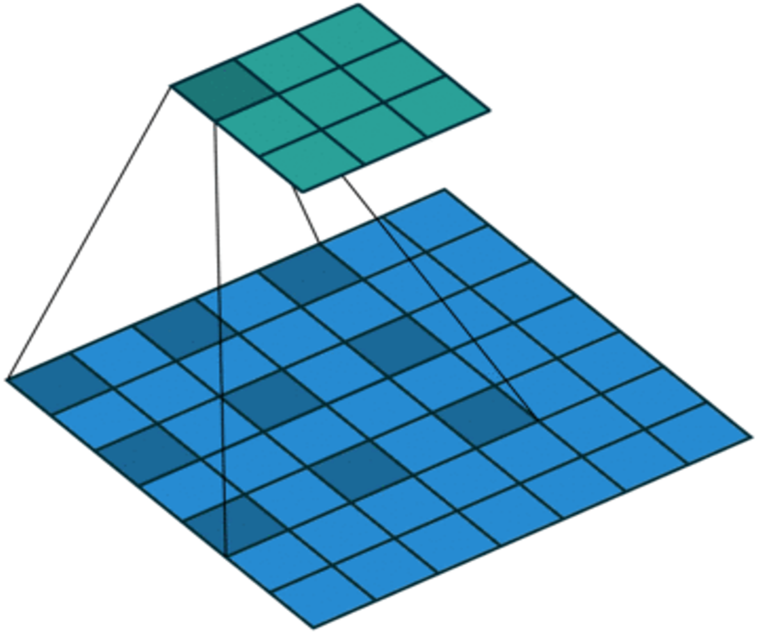


Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. ICLR. 2015

Rich Context Information (Encoder)

Deeplab v1-v3

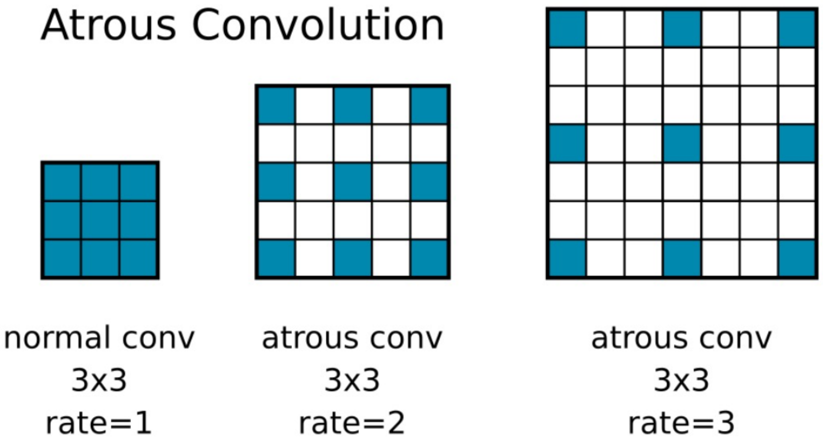


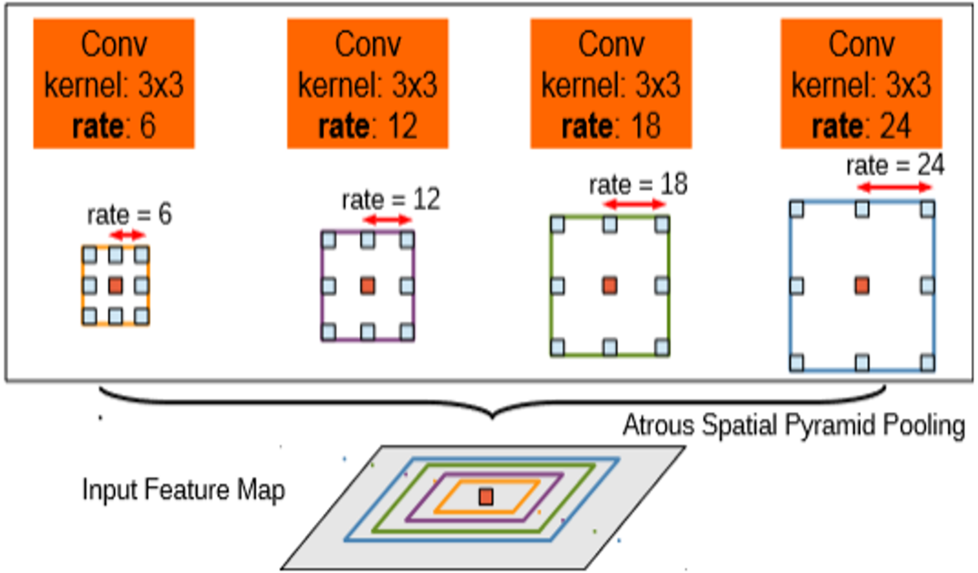


Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. ICLR. 2015

Rich Context Information (Encoder)

Deeplab v2

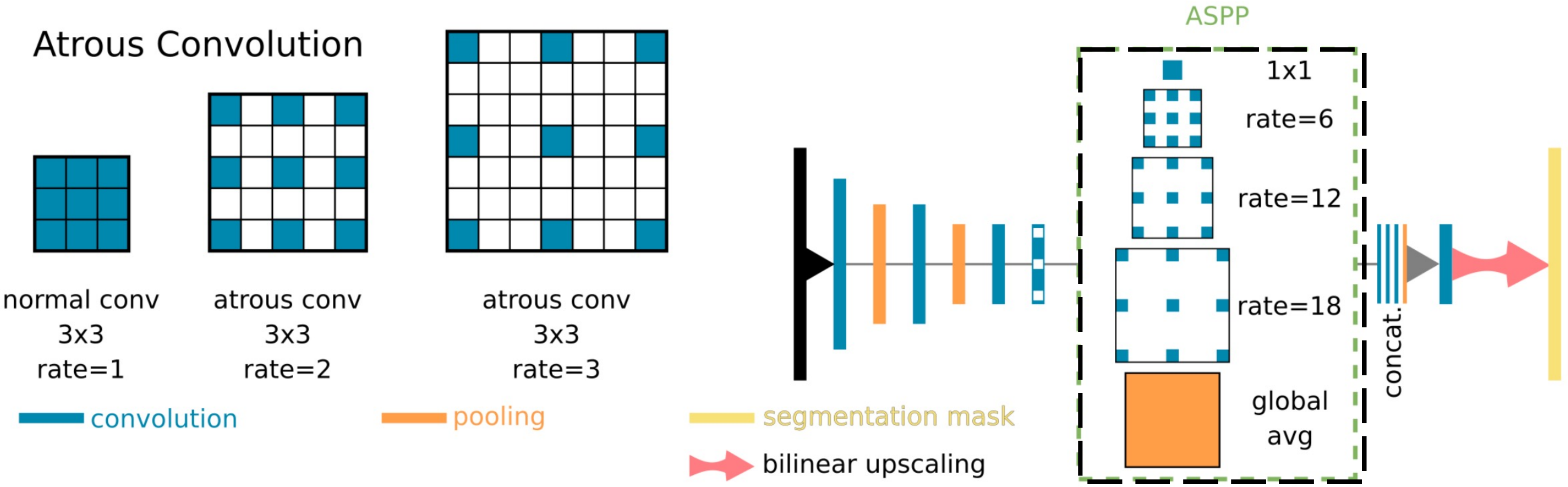




Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. ICLR. 2015

Rich Context Information (Encoder)

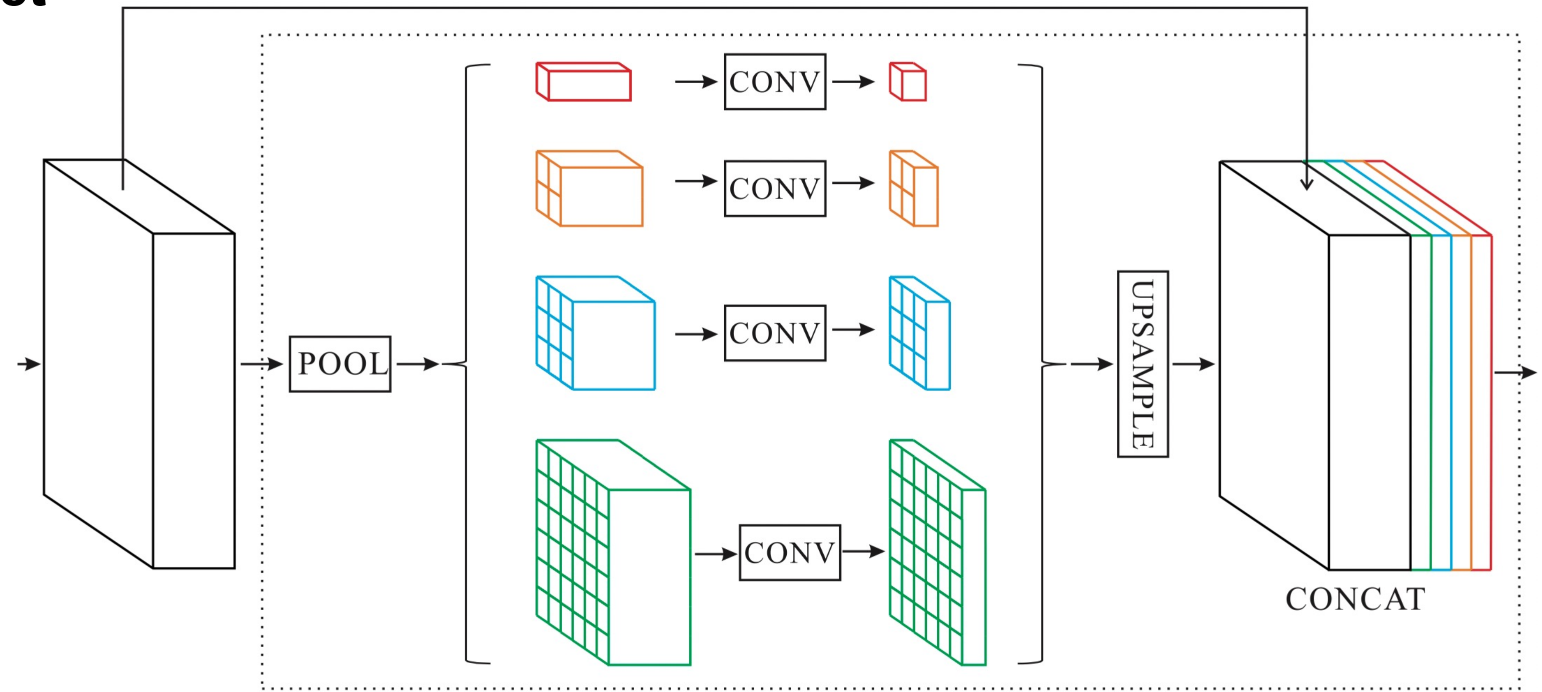
Deeplab v3



Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. ICLR. 2015

Rich Context Information (Encoder)

PSPNet



Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

Rich Context Information (Encoder)

Further reading:

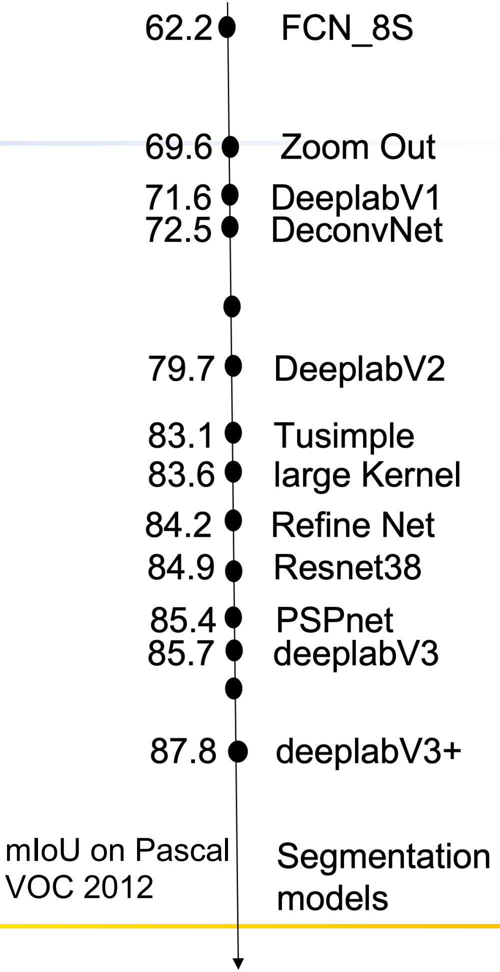
- DenseASPP for Semantic Segmentation in Street Scenes
- Context Encoding for Semantic Segmentation
- Representative Graph Neural Network
- Object-Contextual Representations for Semantic Segmentation
- Not All Pixels Are Equal: Difficulty-Aware Semantic Segmentation via Deep Layer Cascade
- Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes

Outline

Baseline
Fully convolutional network

Rich context information (Encoder)
Multi scale & Enlarge the reception field: Deeplab, DenseNet, PSPNet

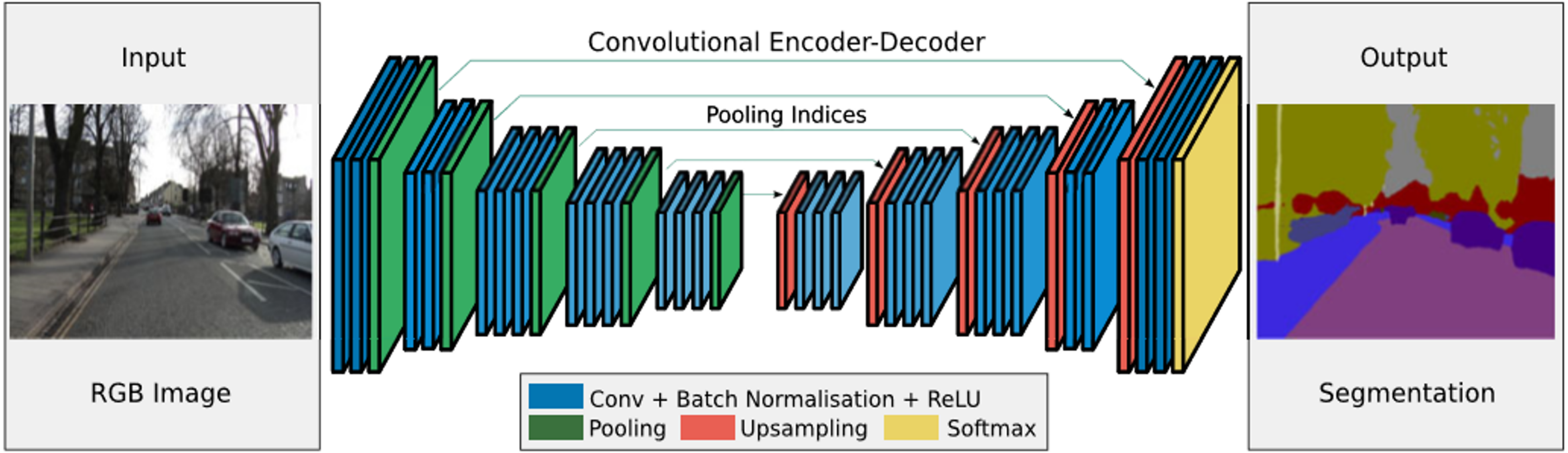
Decoder
DeconvNet, SegNet, Tusimple, DeelabV3+



UNIVERSITY OF CAMBRIDGE

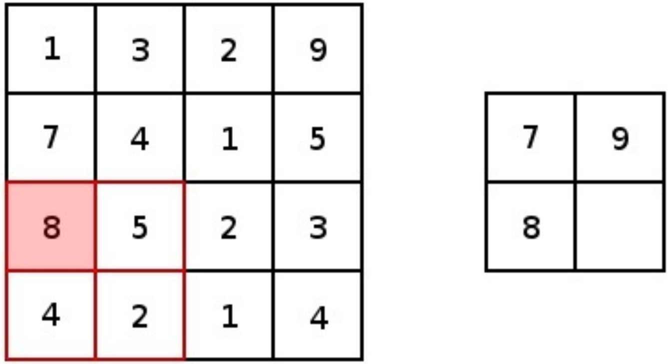
Decoder

SegNet

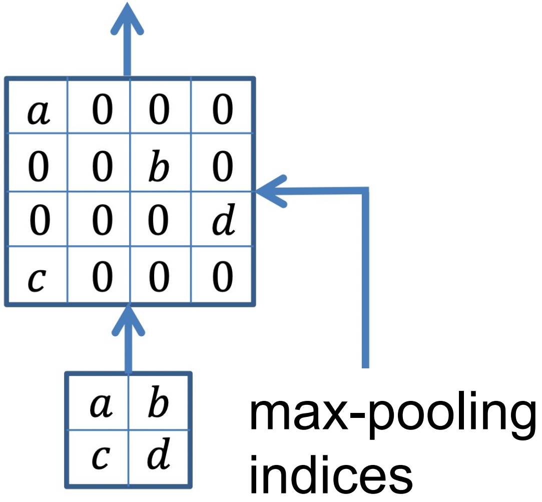


Decoder

SegNet



Max-pooling

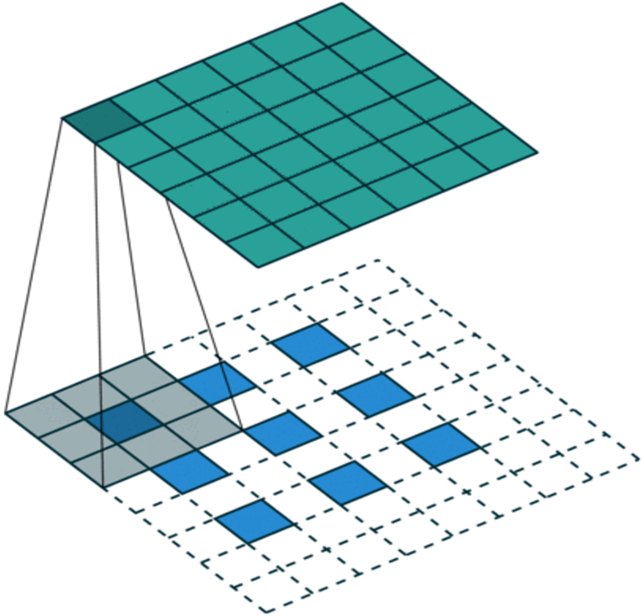


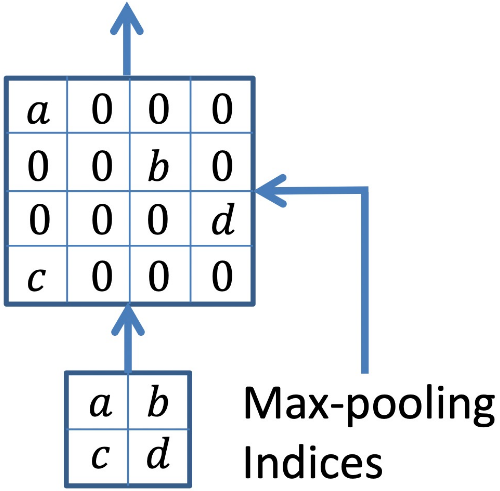
SegNet

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

Decoder

SegNet

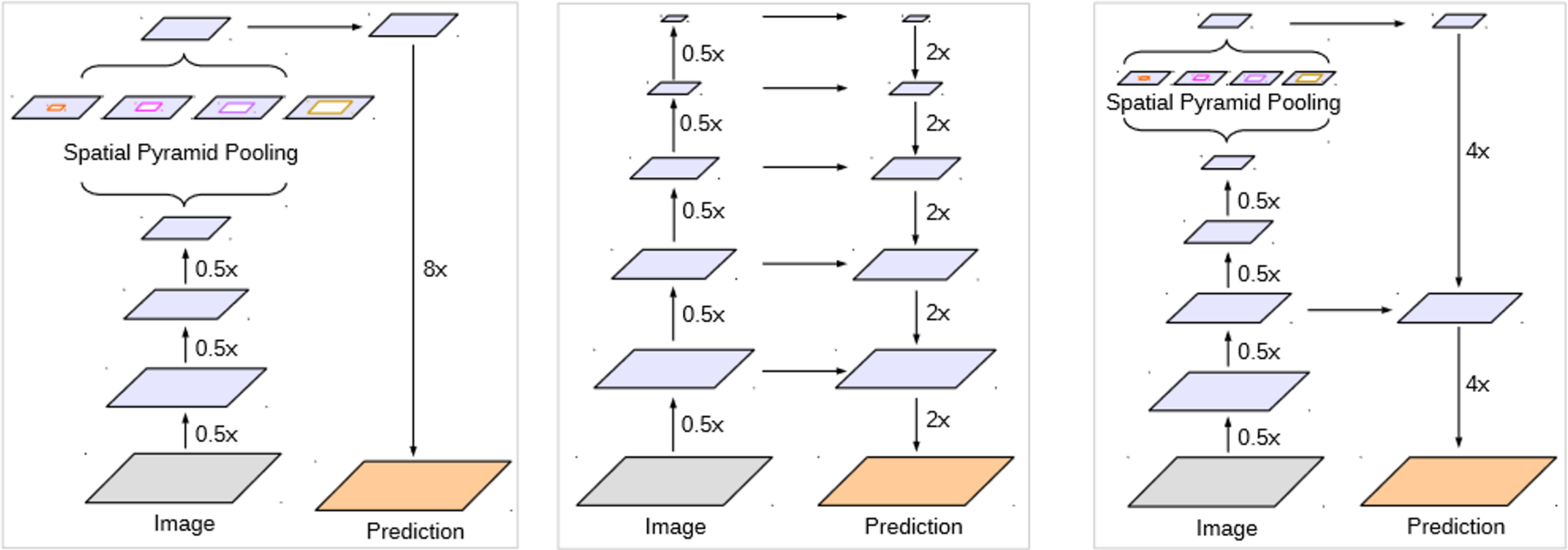




SegNet

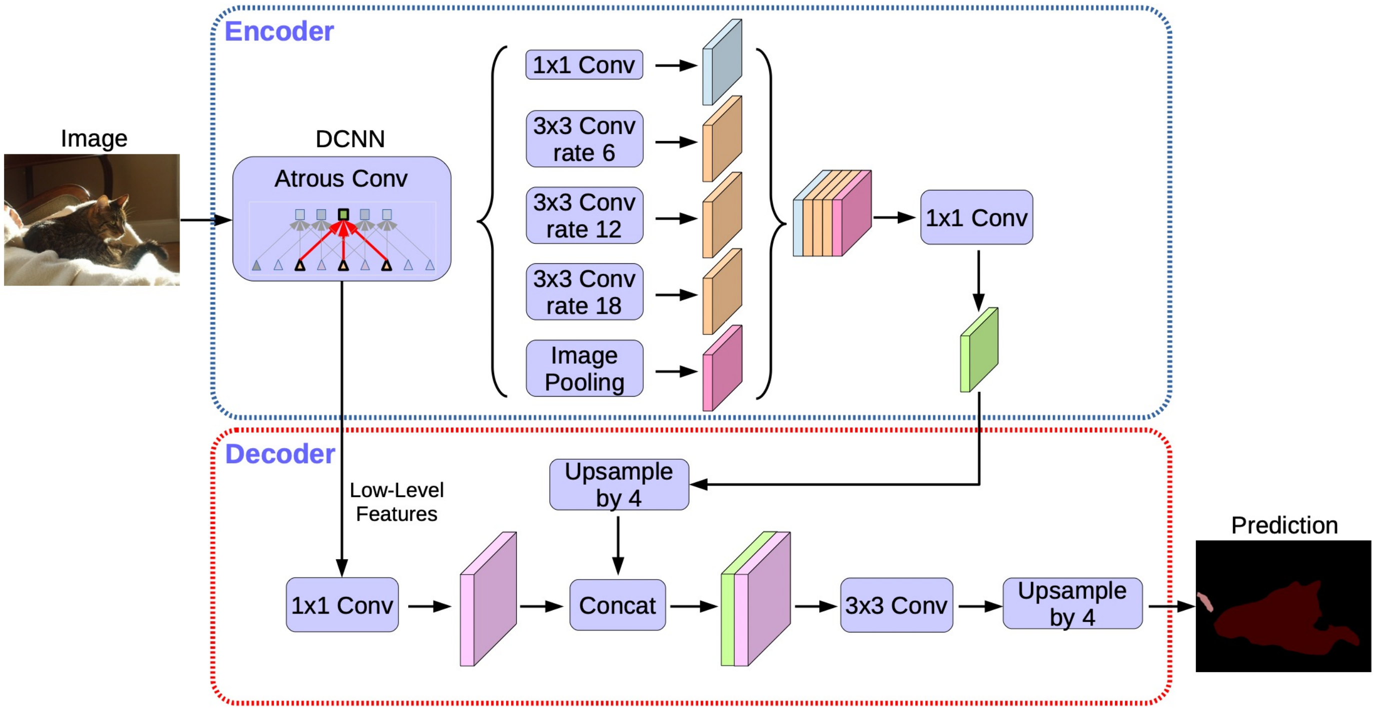
Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

Deeplab v3+



Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

Deeplab v3+



UNIVERSITY OF CAMBRIDGE

Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

Deeplab v3+

Image w/ BU w/ Decoder

UNIVERSITY OF CAMBRIDGE

Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

Conclusions

- Take away messages
  - Dense prediction problem
  - Classification: large receptive field and rich context info.
  - Segmentation: localization, fine-grained boundaries
  - Deeplabv3+ is a strong baseline.

Open Questions

- Trade-off between accuracy and efficiency
- Generalization to various classes
- Unbalanced training samples
- Semi-supervised, weak-supervised learning

References

- [1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on CVPP. 2015: 3431-3440.
- [2] Mostajabi M, Yadollahpour P, Shakhnarovich G. Feedforward semantic segmentation with zoom-out features[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3376-3385.
- [3] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.
- [4] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1520-1528.
- [5] Chen L C, Yang Y, Wang J, et al. Attention to scale: Scale-aware semantic image segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3640-3649.
- [6] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.
- [7] Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 1451-1460.
- [8] Peng C, Zhang X, Yu G, et al. Large kernel matters--improve semantic segmentation by global convolutional network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4353-4361.
- [9] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.

References

- [10] Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. Pattern Recognition, 2019, 90: 119-133.
- [11] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
- [12] \Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 7151-7160.
- [13] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017
- [14] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495..