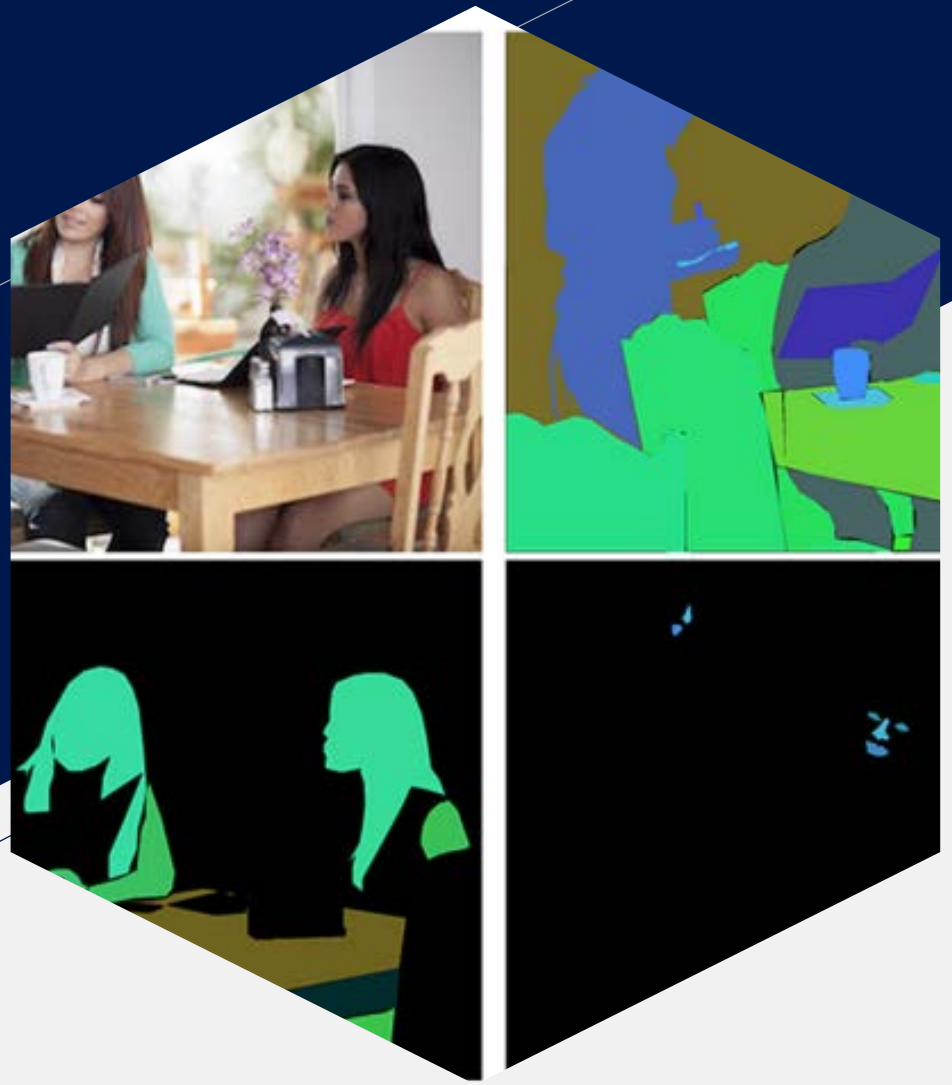


Semantic Segmentation

U-Net, PSPNet (Case Studies)

1. U-Net

Architecture, Objective Function, etc.

1.1 U-Net

Definition

- Extract intrinsic features of the scene, such as its overall shapes, colors and edges, and contextual information about objects in the scene
- This is a low-frequency scene reconstruction and high-level scene features filter

Encoder-Decoder

- Encoder
  - The input image I is transformed
  - Tensor of low spatial resolution
  - Large number of channels
  - Representing abstract information of the scene
- Decoder:
  - Increase the spatial resolution
  - Reduces the channel-depth
  - Recovering the overall structure, colors, etc.

1.2 U-Net

Autoencoder consists of a contracting path (left side) and an expansive path (right side)

Contracting Path:

- Repeated application of two 3x3 convolutions (unpadded) each followed by a ReLU
- 2x2 max pooling operation with stride 2 (downsampling)
- Double the number of feature channels

Expansive Path:

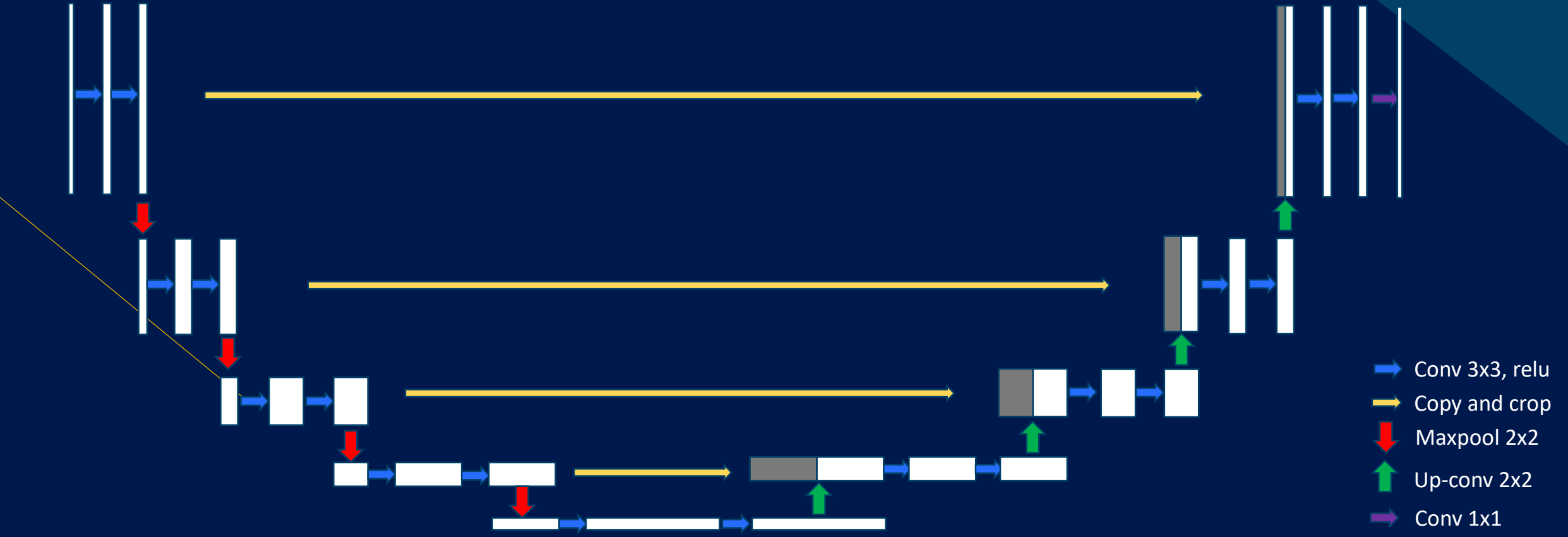
- 2x2 upconvolution to halve the number of feature channels (upsampling)
- Concatenation with the correspondingly cropped feature map from the contracting path
- Two 3x3 convolutions, each followed by a ReLU

Final Layer:

- 1x1 convolution to map each final layer component feature vector to the desired number of classes

1.3 U-Net

Autoencoder consists of a contracting path (left side) and an expansive path (right side)



1.4 U-Net

Objective Function

- Combination of two terms
  - Pixel-wise softmax
  - Cross entropy loss
- SGD optimizer
- 10 hours on a Nvidia Titan GPU (6 GB)

$$p_k(x) = \frac{e^{a_k(x)}}{\sum_{k'=1}^{K} e^{a_{k'}(x)}}$$

- *k: feature channel position*
- *x: pixel position*
- *K: number of classes*

$$E = \sum_{x \epsilon \Omega} w(x)\, log\big(p_l(x)\big)$$

- $l : \Omega \ \epsilon \ \{1, \dots, K\}$
- $w : \Omega \, \epsilon \, R$
- *w: A weight map to give some pixels more importance*

1.5 U-Net

Results

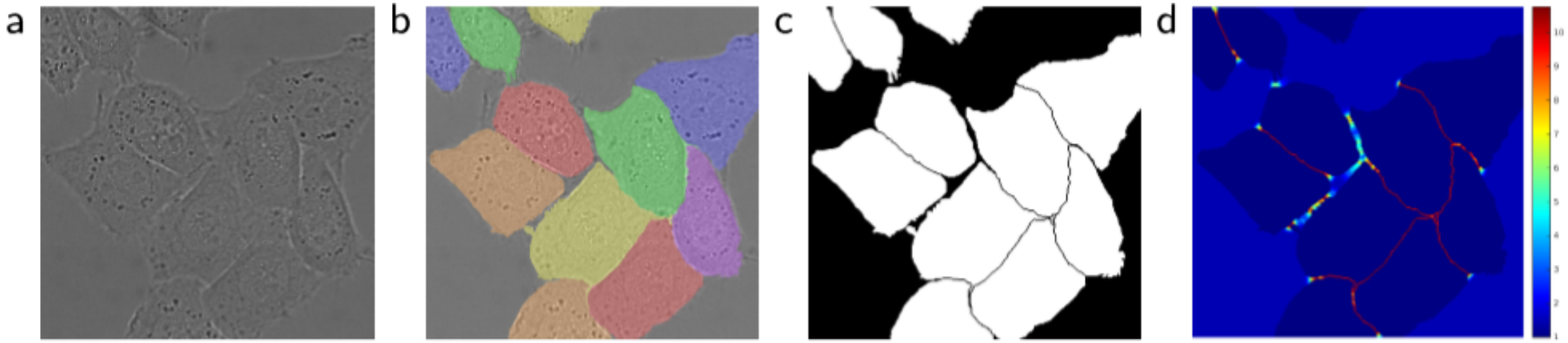


Fig. 3. HeLa cells on glass recorded with DIC (differential interference contrast) microscopy. (**a**) raw image. (**b**) overlay with ground truth segmentation. Different colors indicate different instances of the HeLa cells. (**c**) generated segmentation mask (white: foreground, black: background). (**d**) map with a pixel-wise loss weight to force the network to learn the border pixels.

1.6 U-Net

Results

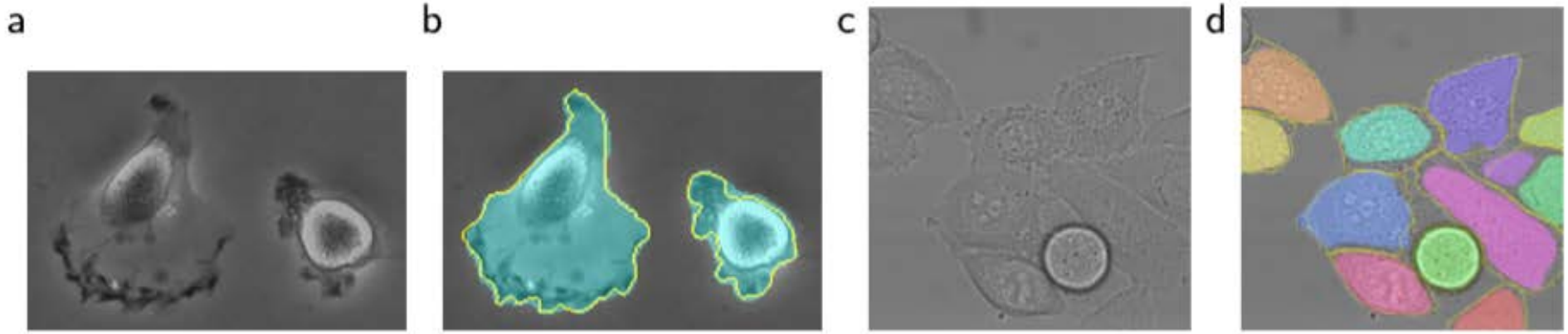


Fig. 4. Result on the ISBI cell tracking challenge. (**a**) part of an input image of the "PhC-U373" data set. (**b**) Segmentation result (cyan mask) with manual ground truth (yellow border) (**c**) input image of the "DIC-HeLa" data set. (**d**) Segmentation result (random colored masks) with manual ground truth (yellow border).

2. PSPNet

ResNet, Dilation, Supervision,
Pyramid Pooling, etc.

2.1.1 PSPNet

Definition

- It predicts label, location and the shape for each element
- Combination of CNN, Global context pooling

ADE20K Dataset Complex Scenes

Input Image Ground Truth

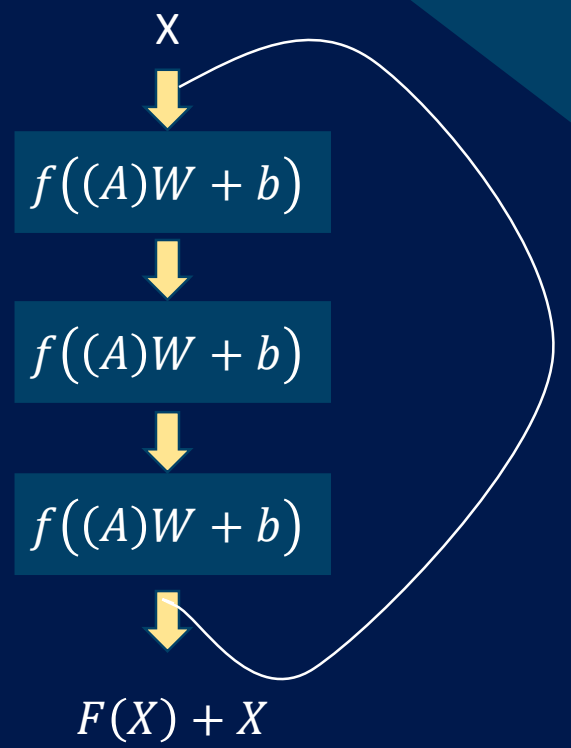
2.2.1 PSPNet

ResNet

Architecture

- Deeper is not better!
- It is easy to learn identity $f_{(x)} = x$
- We attach identity network to the exiting deep network and answer won't change
- $H(X) = F(X) + X$ is the desired mapping, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers and learn residuals:

$$H(X) - X$$



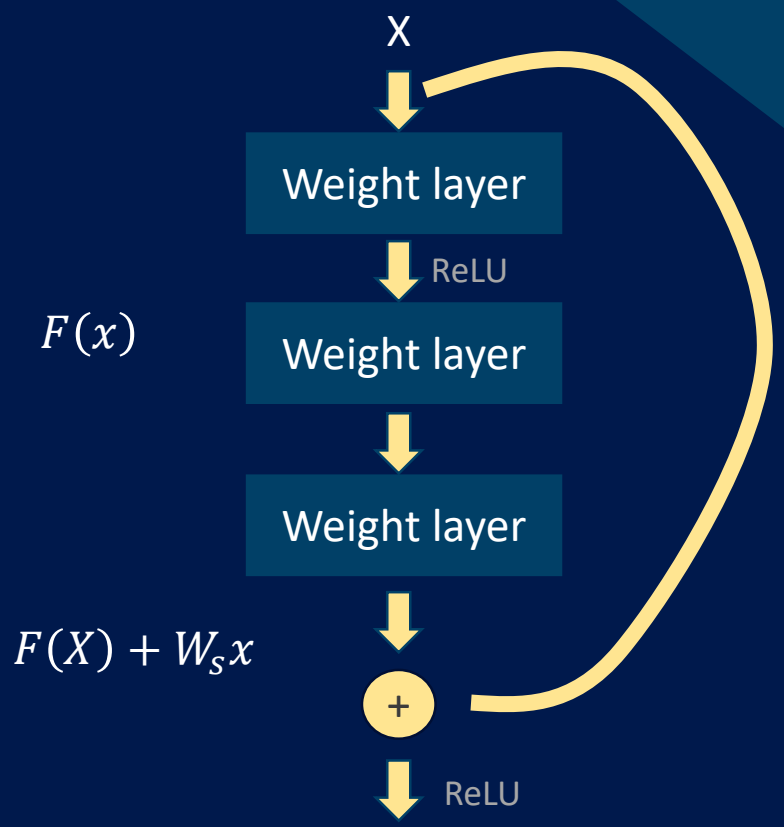
2.2.2 PSPNet

ResNet

	Plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	25.03

Architecture

- General Residual mapping
- Linear projection W_s by the shortcut connections to match the dimensions or zero padding $y = F(x, \{W_i\}) + W_s x$
- The element-wise addition is performed on two feature maps, channel by channel



2.3.1 PSPNet

Dilated ResNet

Definition

- CNNs for image classification progressively reduce resolution (7×7 is typical)
- Preserving the spatial resolution while increasing the receptive field to help to the contribution of small and thin objects
- [F. Yu et al. 2017]

- Generalized discrete convolution operator:

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)\, k(t)$$

Model	Top-1	Top-5
DRN-42	50.7	46.8
ResNet-101	54.6	51.9

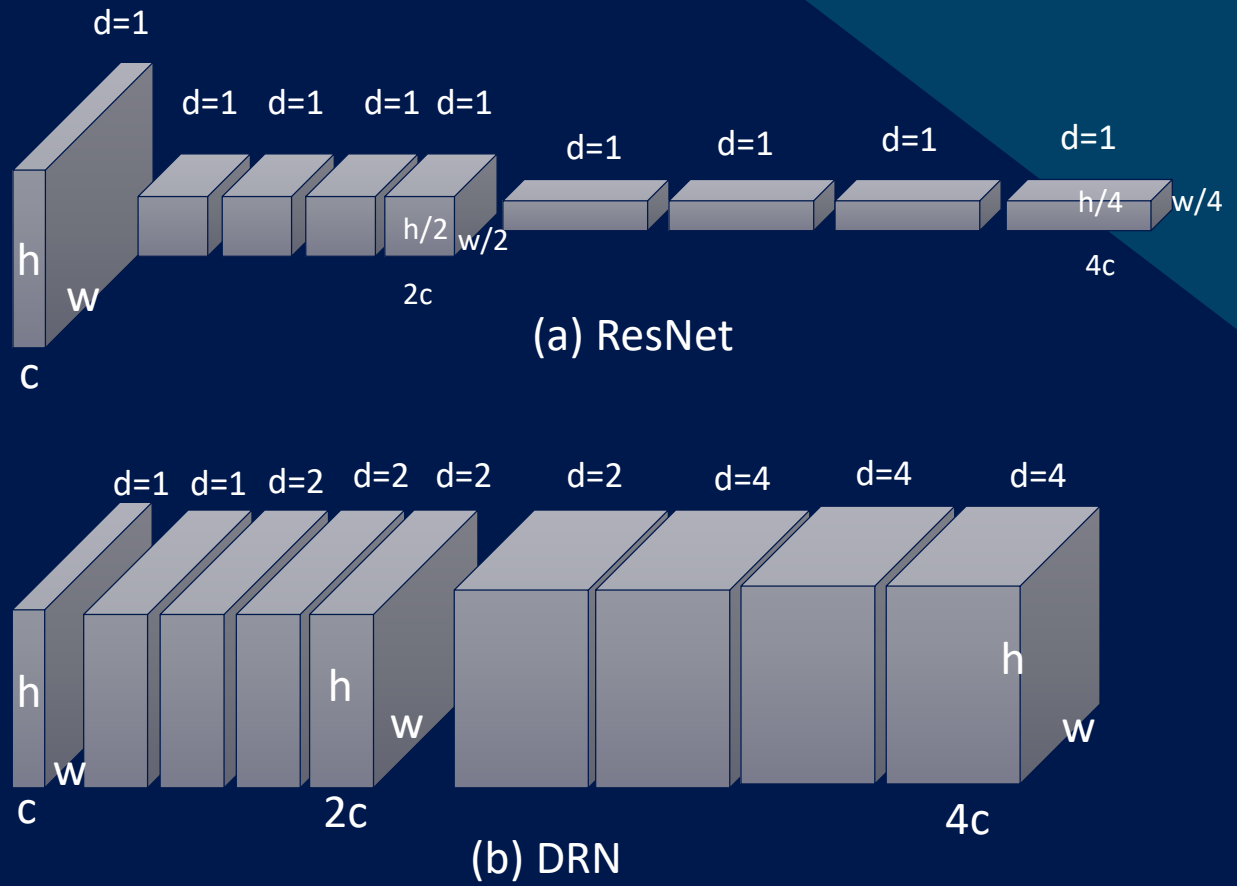
Weakly-supervised object localization error rates on the ImageNet validation set

2.3.1 PSPNet

Dilated ResNet

Architecture

- Remove striding in the first layer of each group
  - Receptive field of first layers won't change
  - Subsequent layers receptive field reduced by factor of 2
- Replace convolution in subsequent layers by 2-dilated convolutions
- It generates some artifacts which is removed by removing some of residual skip-connections and adding convolution layers at the end



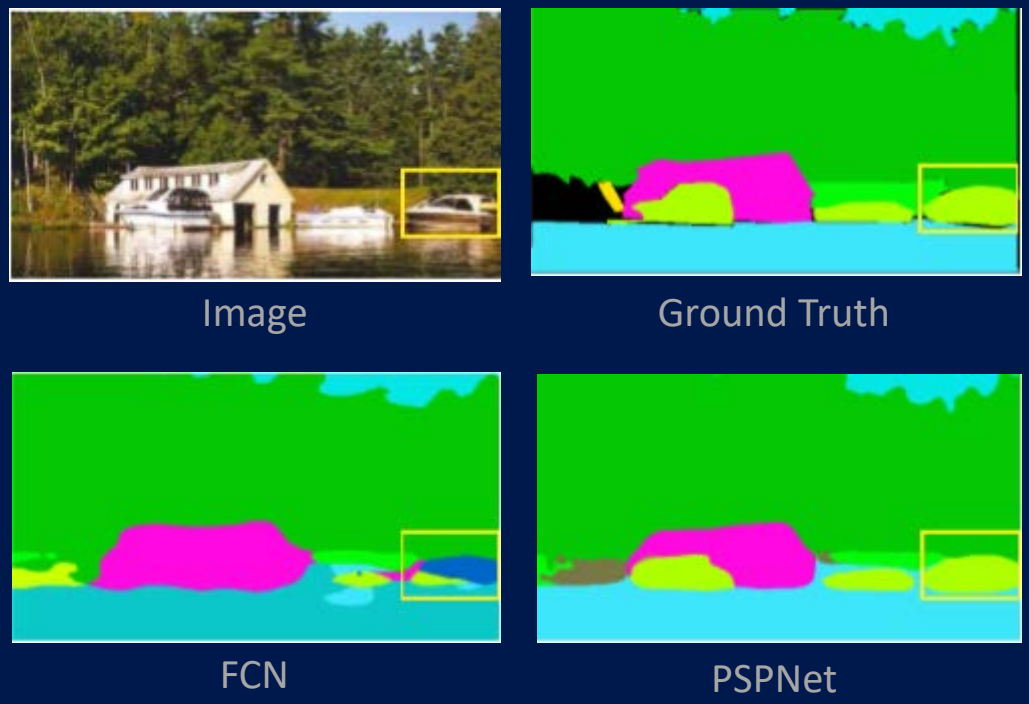
2.4.1 PSPNet

Pyramid Pooling

Definition

- CNNs face challenges considering diverse scenes and unrestricted vocabulary
  - Mismatched relations
  - Confusion categories
  - Inconspicuous classes
- Exploit the capability of global context information by different-region-based context aggregation
- [H. Zhao, et al. 2017]

Scene parsing issues on ADE20K dataset



2.4.1 PSPNet

Pyramid Pooling

Scene parsing issues on ADE20K dataset

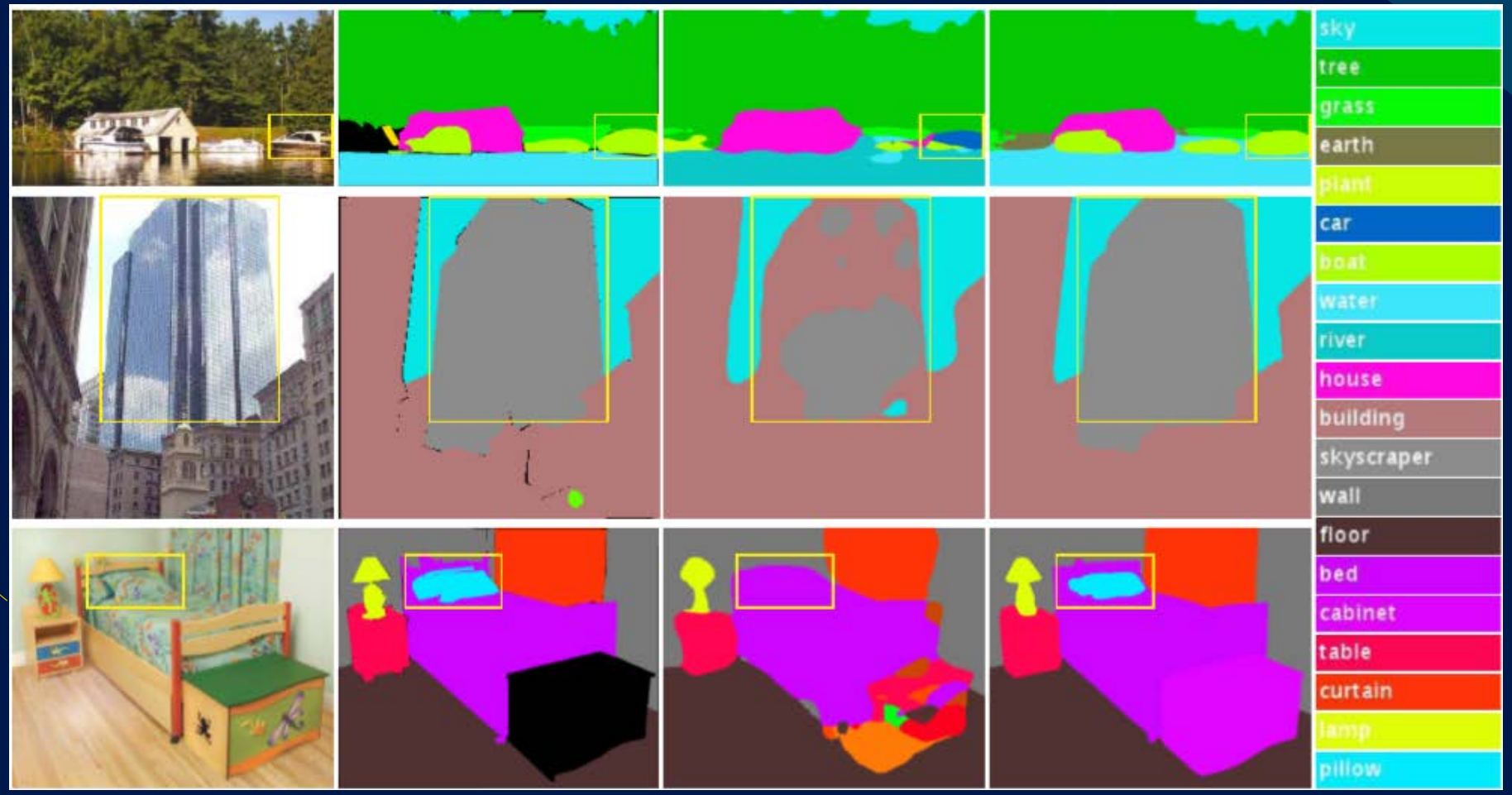


Image Ground Truth FCN PSPNet

2.4.2 PSPNet

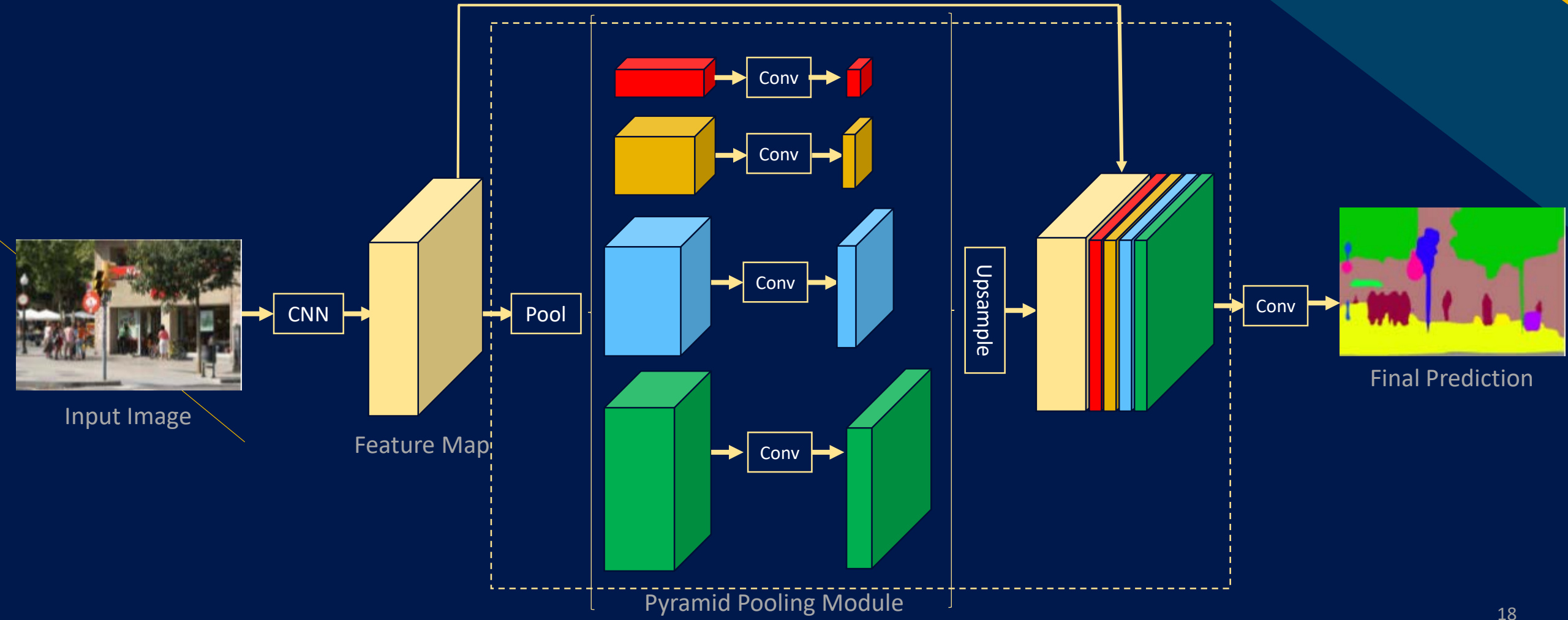
Pyramid Pooling

Architecture

- CNN to get the feature map of last convolutional layer
- Pyramid parsing module is applied to harvest different sub-region representations
- Upsampling and concatenation layers to form the final feature representation, which carries both local and global context information
- The representation is fed into a convolution layer to get the final per-pixel prediction
- The pyramid pooling module fuses features under four different pyramid scales
- It separates the feature map into different sub-regions and forms pooled representation for different locations

2.4.3 PSPNet

Pyramid Pooling



2.5.1 PSPNet

Deep Supervision

Definition

- In order to train deeper networks, we propose to add auxiliary supervision branches after certain intermediate layers during training
- Encourage the feature maps at lower layers to be directly predictive of the final labels
- At test time, we remove supervision branches, So both models have the exact same structure at test time
- Supervision as a form of regularization

2.5.2 PSPNet

Deep Supervision

Architecture

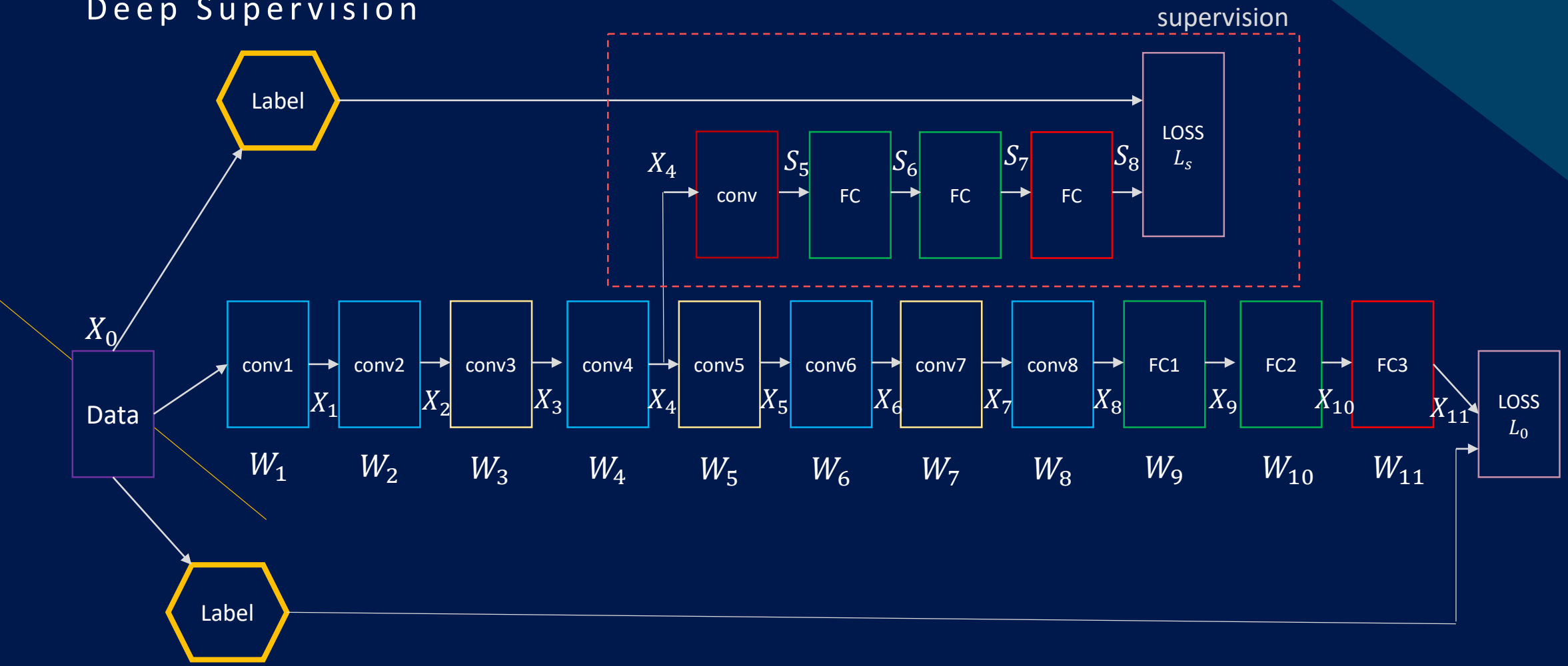
- Supervision is a small neural network composed of a convolutional layer, several fully connected layers, and a softmax classifier
- Optimize a loss function that is a sum of the overall (final-layer) loss and companion losses associated with all intermediate layers

$$L(w, w_s) = \left(\sum_{i=0} L_s^i\right) + L_0$$

- Where to add super vision:
  - We run a few (10-50) iterations of back-propagation
  - Plot the mean gradient values of intermediate layers
  - Add supervision after the layer where the mean gradient value vanishes (10^{-7})

2.5.3 PSPNet

Deep Supervision



2.6.1 PSPNet

Result

Image Ground Truth Baseline PSPNet

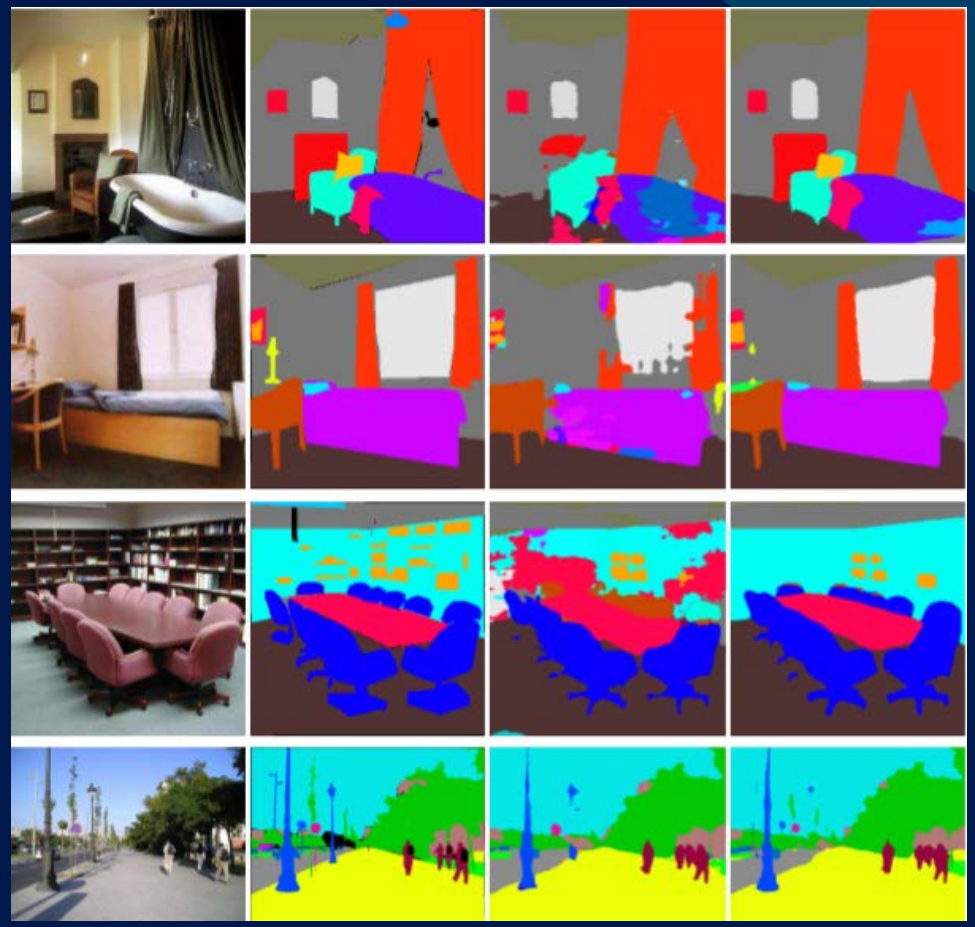


Image Ground Truth Baseline PSPNet

References

- O. Ronneberger, P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer Assisted Intervention (MICCAI) (LNCS), Vol. 9351. Springer, 234–241. (available on arXiv: 1505.04597 [cs.CV]).
- Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017b. Scene Parsing through ADE20K Dataset. In IEEE Conference on Computer Vision and Pattern Recognition.
- F. Yu, V. Koltun, T.A. Funkhouser. 2017. Dilated Residual Networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia. 2017. Pyramid Scene Parsing Network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

References

- https://github.com/hszhao/PSPNet - PSPNet implementation and trained models
- C. A. Bouman: Digital Image Processing - January 7, 2019
- http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net

Thank You

Mohammad Doosti Lakhani

+98 937 915 6599

nikan.doosti@outlook.com

nikronic.github.io

Questions