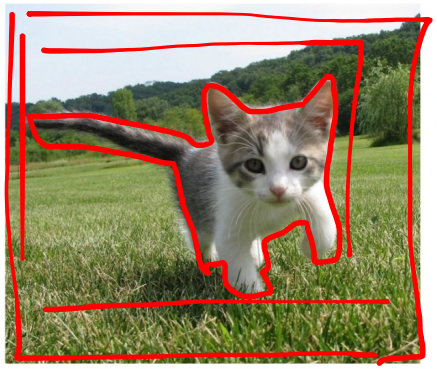
Computer Vision:
Image segmentation

Siheng Chen 陈思衡

Common computer vision tasks

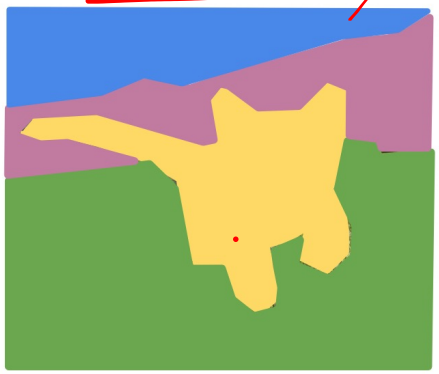
Classification



CAT

No spatial extent

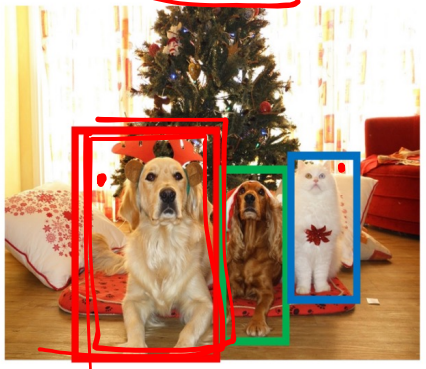
Semantic Segmentation



GRASS, CAT, TREE, SKY

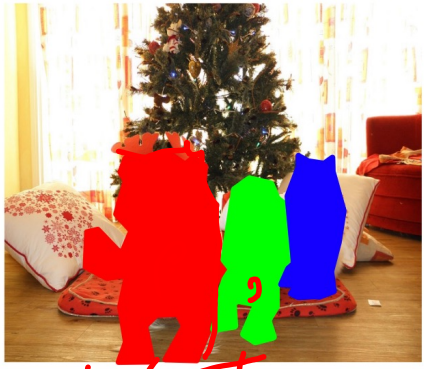
No objects, just pixels

Object Detection



DOG, DOG, CAT

Instance Segmentation



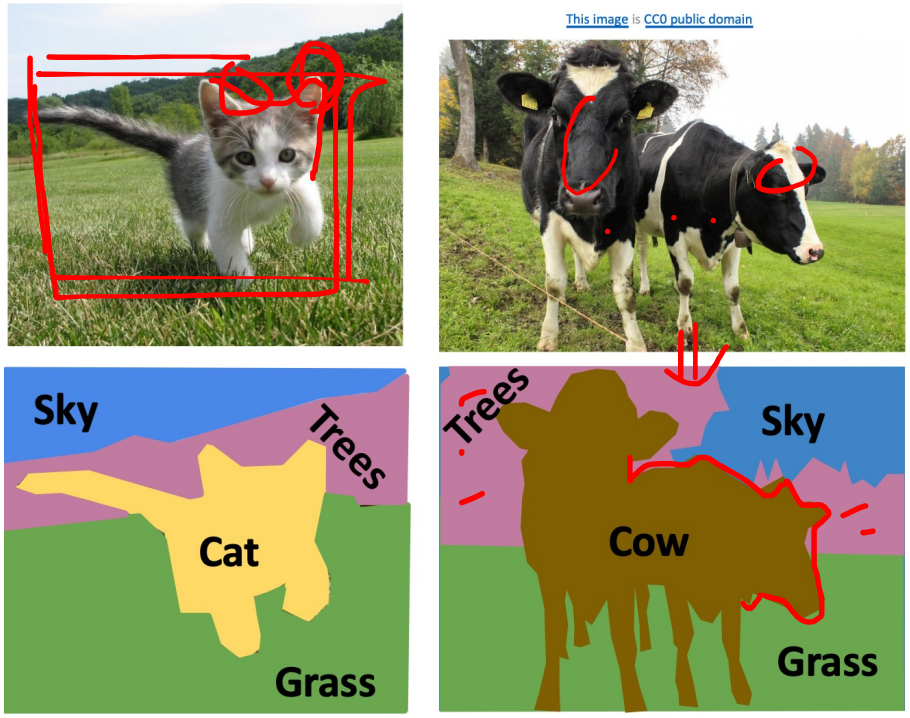
DOG, DOG, CAT

Multiple Objects

Semantic segmentation

Image segmentation is the process of partitioning a digital image into multiple image segments, also known as image regions or image objects (sets of pixels)

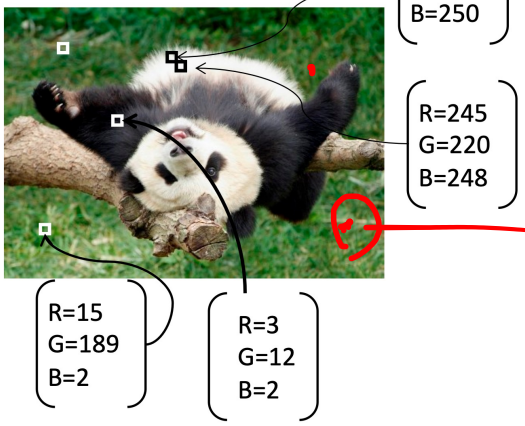
- Label each pixel in the image with a category label
- Semantic: Don't differentiate instances, only care about pixels



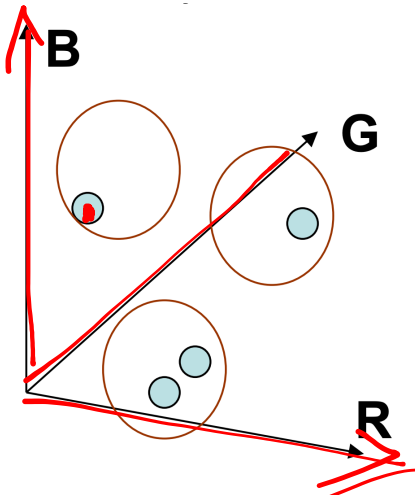
Semantic segmentation

- Segmentation as clustering
- Segmentation as energy minimization
- Pixel-wise end2end learning

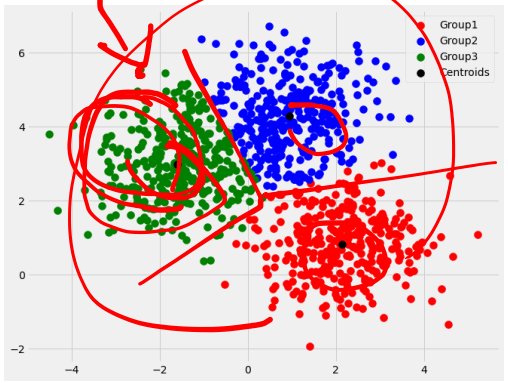
Segmentation as clustering



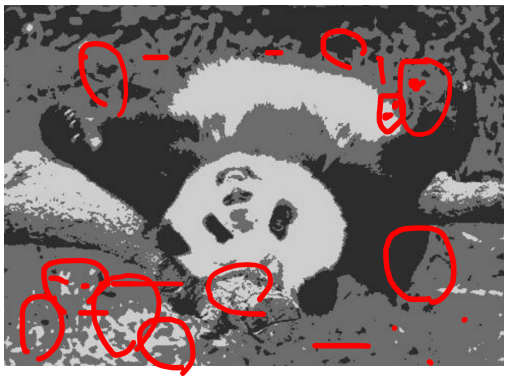
feature representation



clustering



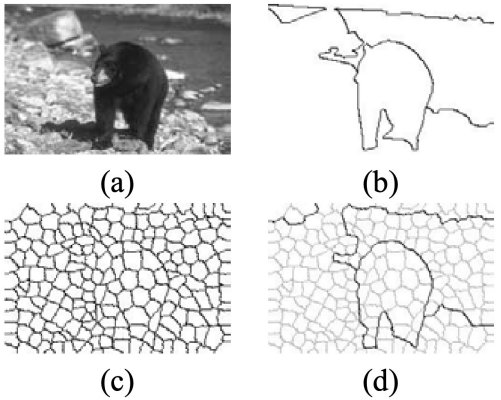
segmentation result



Segmentation as clustering

Superpixel

In this section we present a preprocessing stage to group pixels into "superpixels". The motivations of this preliminary grouping are: (1) pixels are not natural entities; they are merely a consequence of the discrete representation of images; and (2) the number of pixels is high even at moderate resolutions; this makes optimization on the level of pixels intractable. We would like to work with "superpixels' which are local, coherent, and which preserve most of the structure necessary for segmentation at the scale of interest.

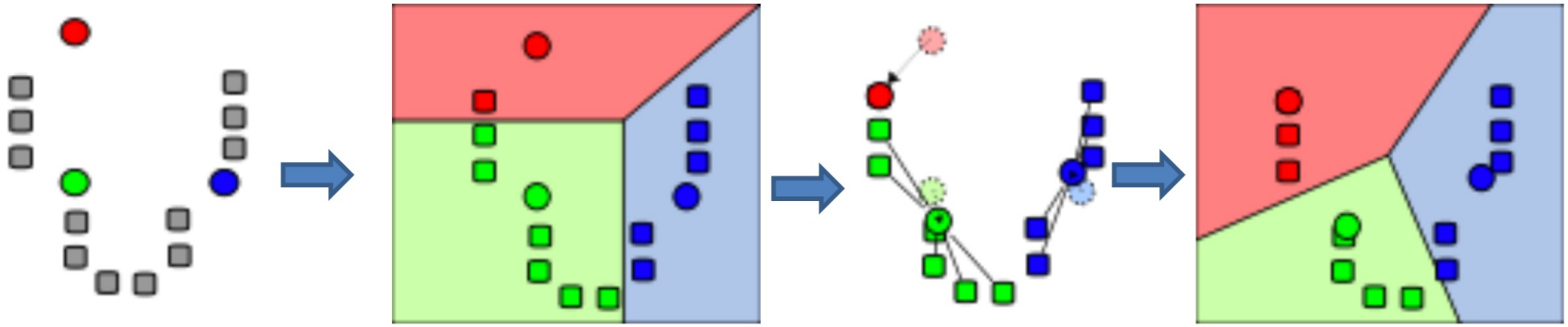


Xiaofeng Ren and Jitendra Malik, Learning a Classification Model for Segmentation, ICCV 2003
Zhengqin Li, Jiansheng Chen, Superpixel Segmentation using Linear Spectral Clustering, CVPR 2015

Segmentation as clustering

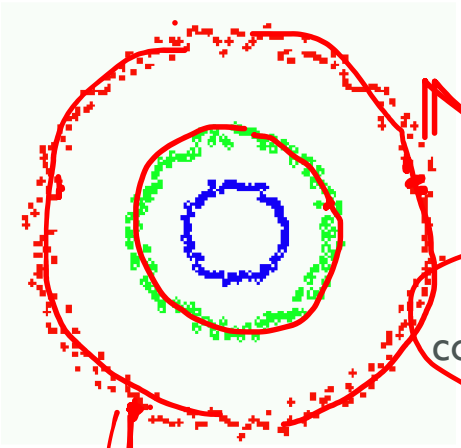
K-means clustering

$$\mathbf{c}^*, \delta^* = \arg\min_{\mathbf{c},\delta} \sum_{x_i} \sum_{c_j} \delta_{i,j}(x_i - c_j)^2 \quad \delta \in \{0,1\}^{N \times K}, \delta 1 = 1$$

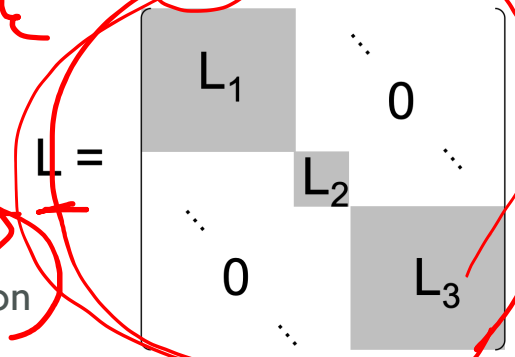


Segmentation as clustering

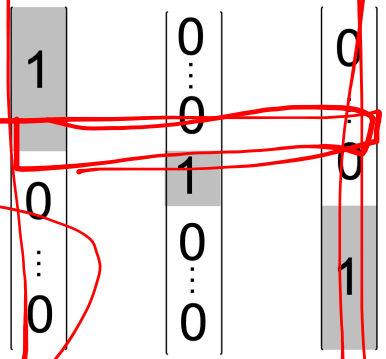
Spectral clustering

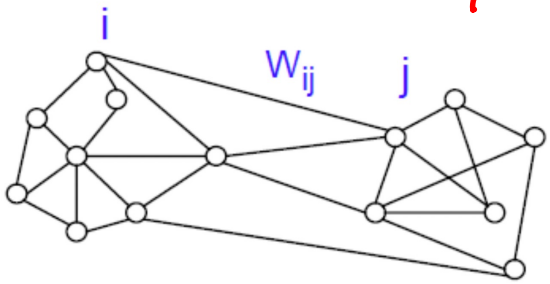


Graph construction



Eigen-decomposition





$G = \{V,E\}$

minimize $\text{Ncut}(A, B) = \frac{\mathbf{f}^T\mathbf{L}\mathbf{f}}{\mathbf{f}^T\mathbf{D}\mathbf{f}}$

subject to $\mathbf{f}^T\mathbf{D}1 = 0$

$$\mathbf{L}\mathbf{f} = \lambda\mathbf{D}\mathbf{f}$$

Semantic segmentation

- Segmentation as clustering
- **Segmentation as energy minimization**
- Pixel-wise end2end learning

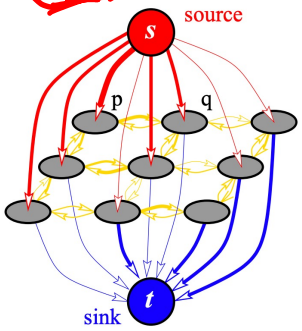
Segmentation as energy minimization

Min-cut/Max-flow algorithms on graphs

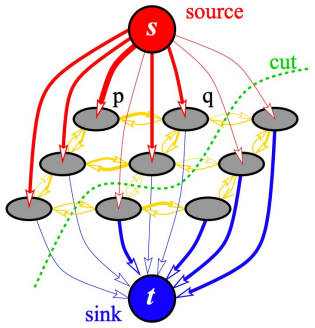
$$E(L) = \sum_{p \in \mathcal{P}} D_p(L_p) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(L_p, L_q)$$

Individual label-preferences of pixels based on observed intensities and pre-specified likelihood function

encourage spatial coherence by penalizing between neighboring pixels



(a) A graph $\mathcal{G}$



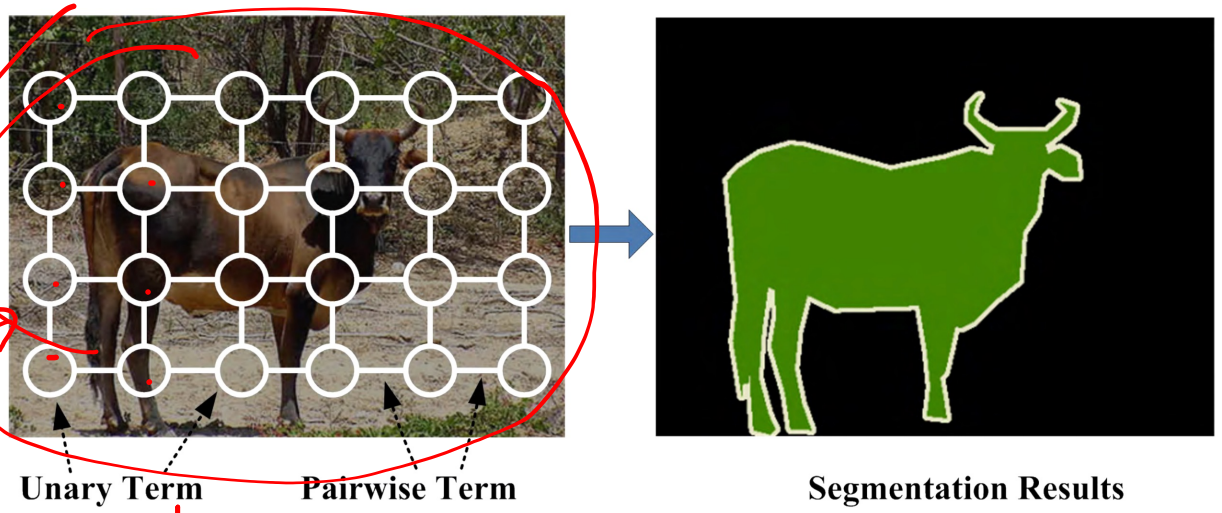
(b) A cut on $\mathcal{G}$

Segmentation as energy minimization

Conditional random field (CRF)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$



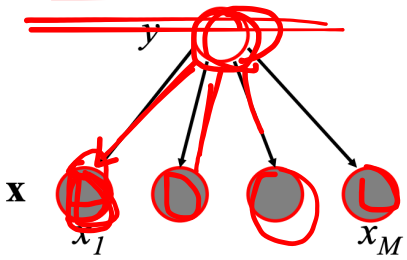
Segmentation as energy minimization

Conditional random field (CRF)

GENERATIVE

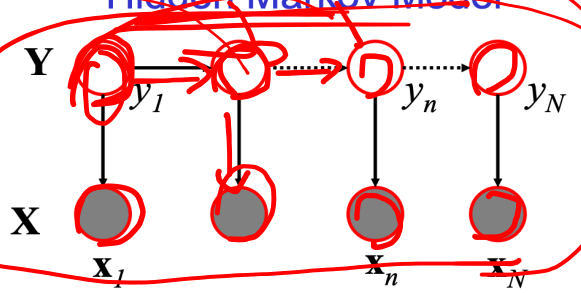
DISCRIMINATIVE

Naïve Bayes Classifier



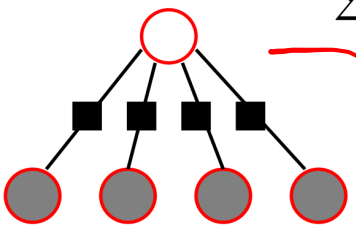
$$p(y,\mathbf{x}) = p(y)\prod_{m=1}^{M} p(x_m \mid y)$$

Hidden Markov Model



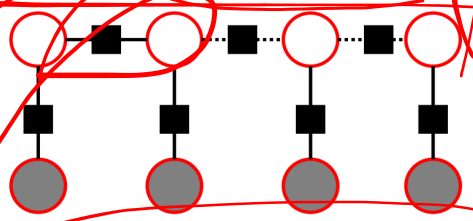
$$p(\mathbf{Y},\mathbf{X}) = \prod_{n=1}^{N} p(y_n \mid y_{n-1})\, p(\mathbf{x}_n \mid y_n)$$

$$p(y \mid \mathbf{x}) = \frac{\exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y,\mathbf{x})\right\}}{\sum_{y'} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y',\mathbf{x})\right\}}$$



Logistic Regression

$$p(\mathbf{Y} \mid \mathbf{X}) = \frac{\exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n)\right\}}{\sum_{y'} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y_n', y_{n-1}', \mathbf{x}_n)\right\}}$$

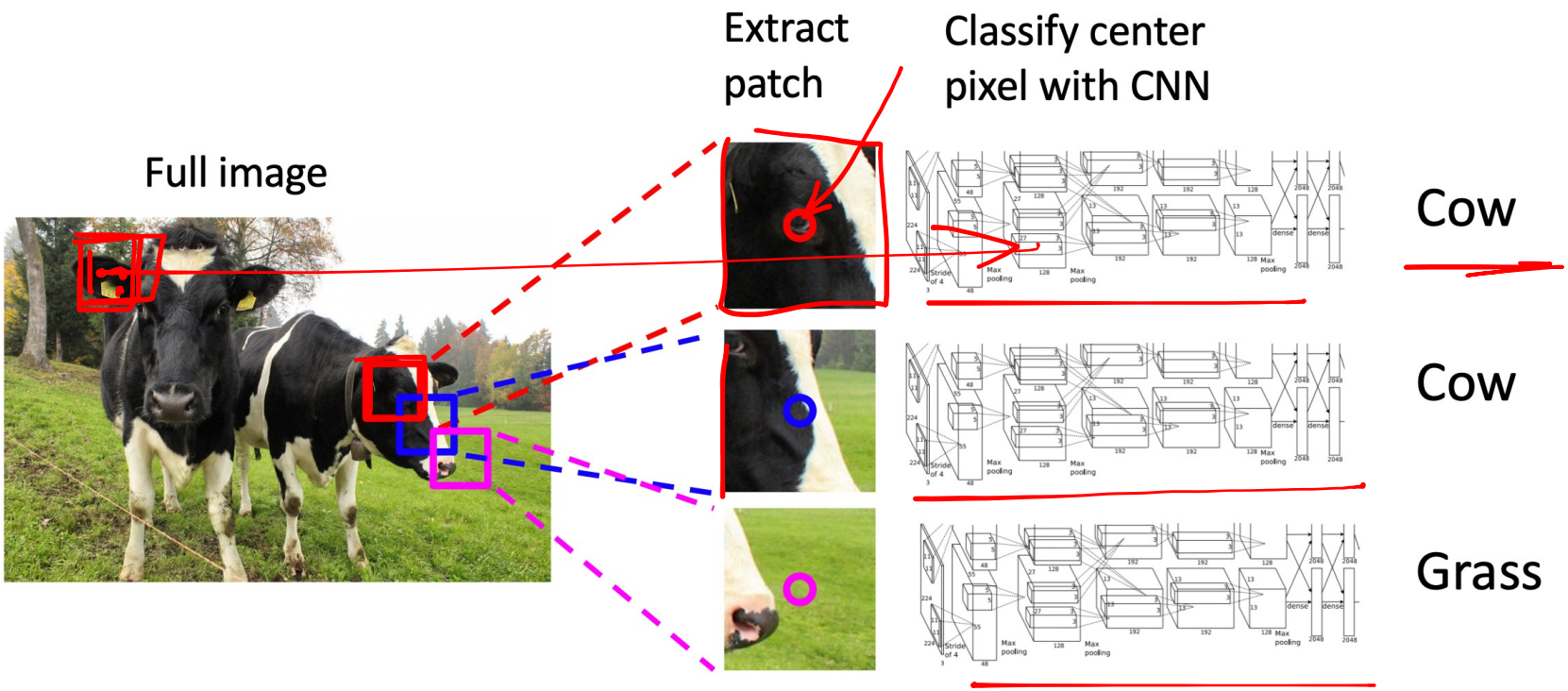


Conditional Random Field

Semantic segmentation

- Segmentation as clustering
- Graph-based segmentation
- Segmentation as energy minimization
- **Pixel-wise end2end learning**
  - Fully convolutional network
  - SegNet
  - DeepLab
  - DeepLabv3+
  - SegFormer

Semantic segmentation



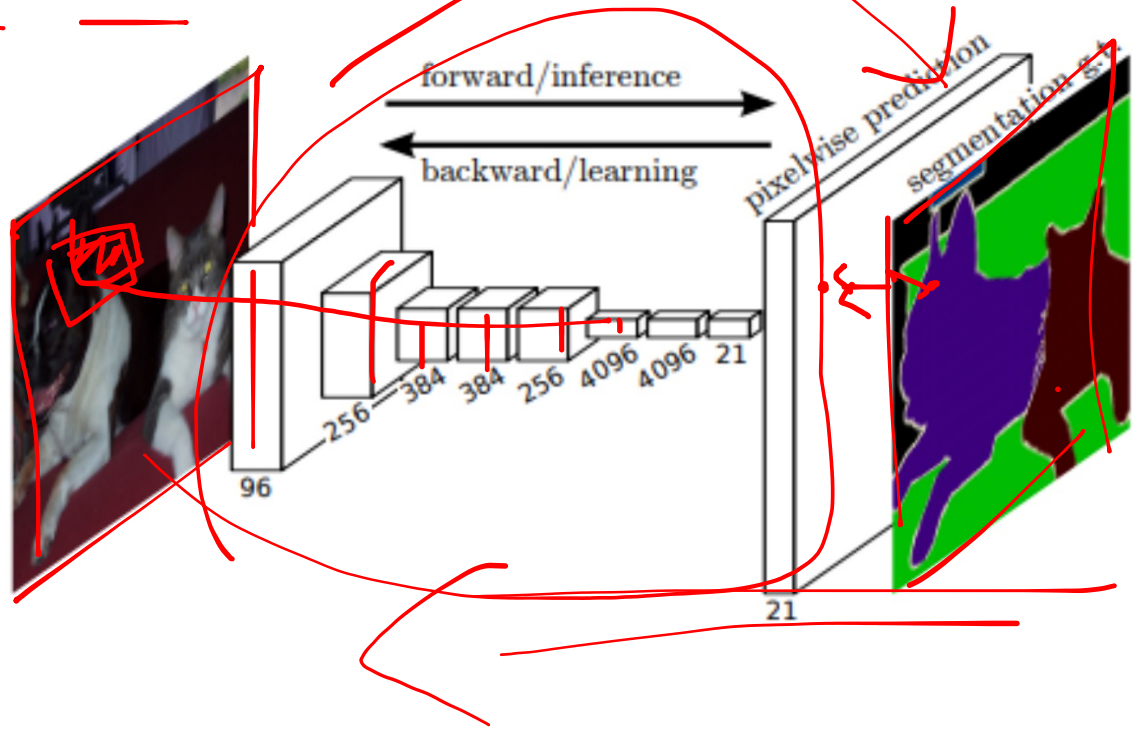
Problem:

Very inefficient! Not reusing shared features between overlapping patches

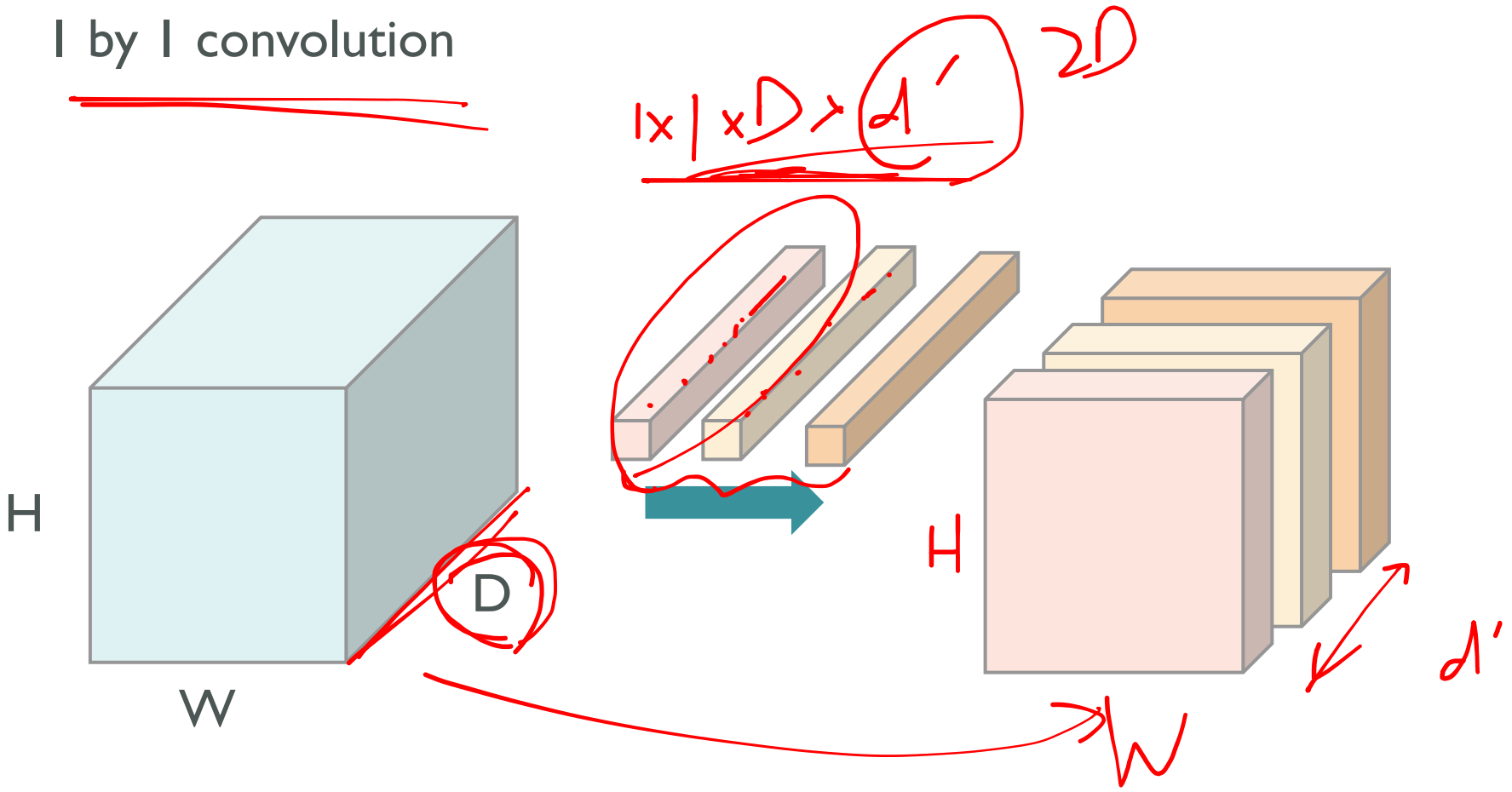
Fully convolutional network

Key ideas

- All convolutional layers.
- Convert the last layer output to the original resolution.
- Softmax-cross entropy loss between pixel-wise predictions and ground truth
- Backprop and SGD



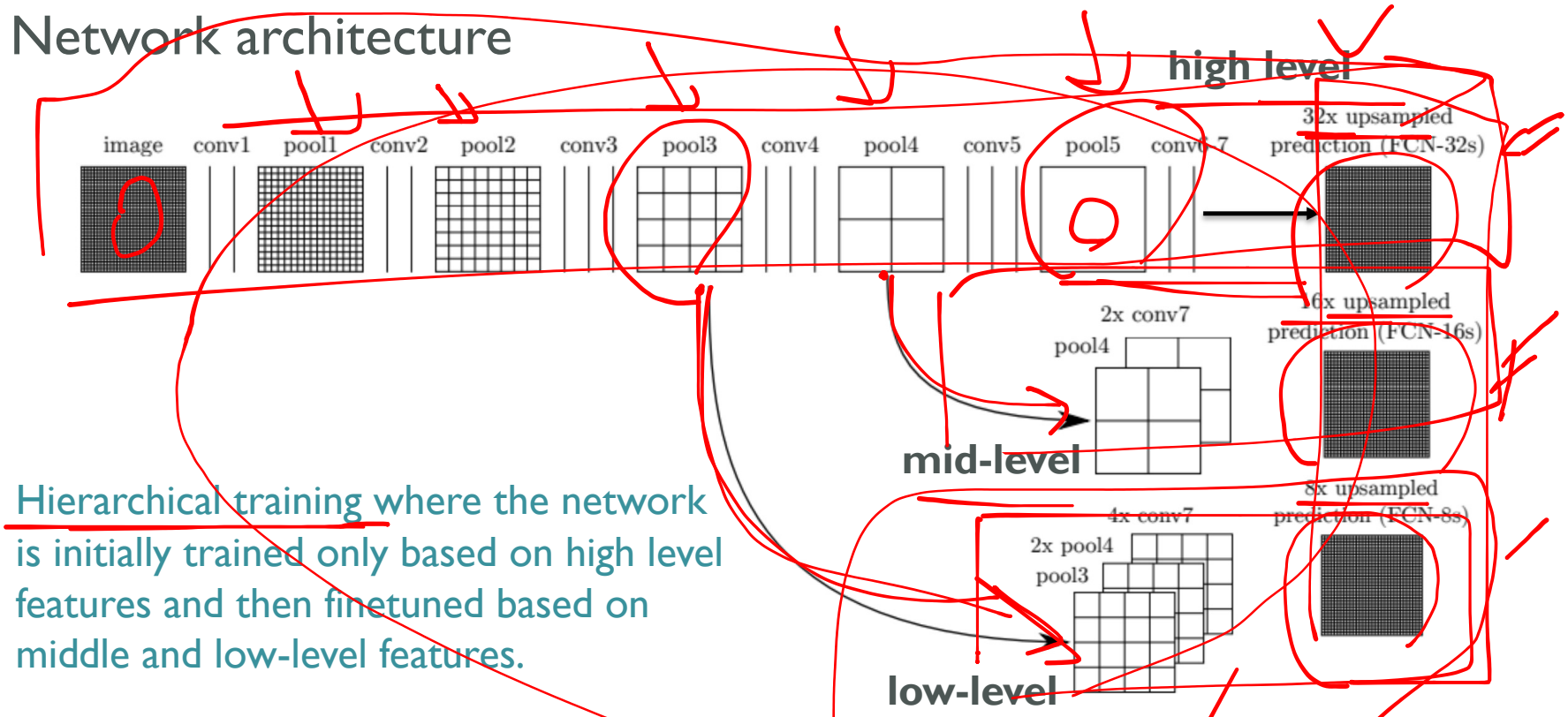
Fully convolutional network



Long, Shelhamer, Darrell - Fully Convolutional Networks for Semantic Segmentation, CVPR 2015, PAMI 2016

Fully convolutional network

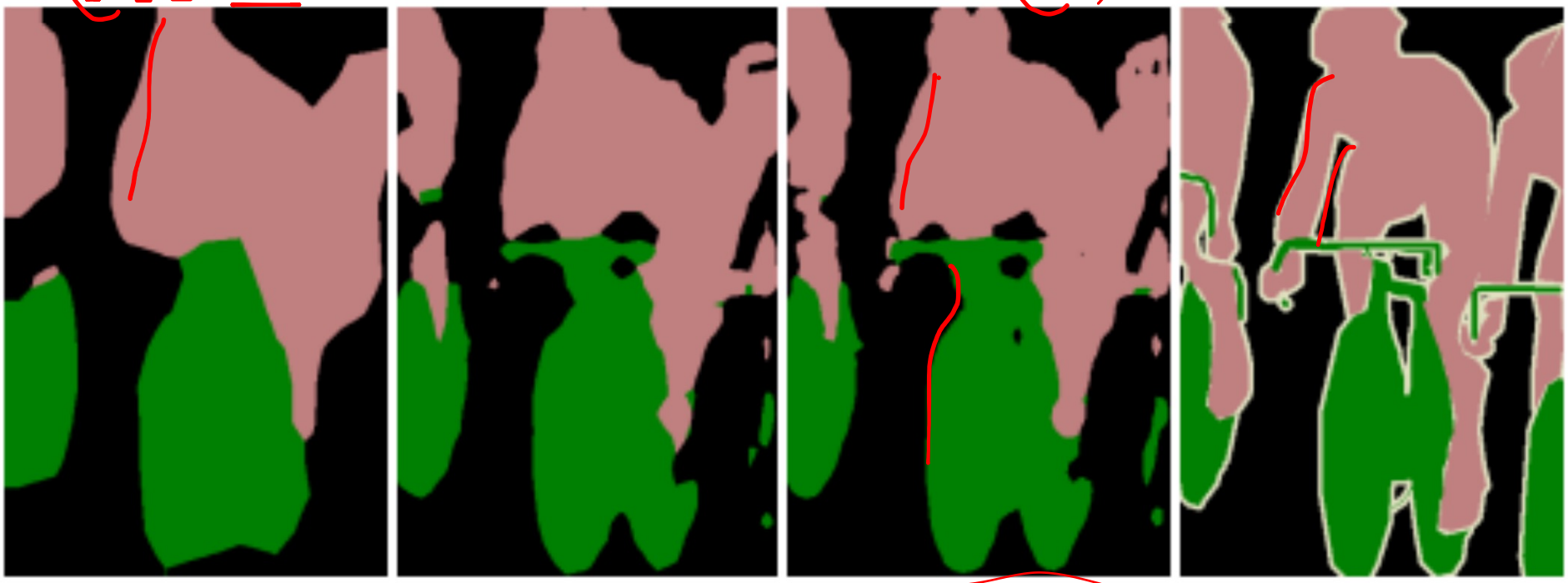
Network architecture



Hierarchical training where the network is initially trained only based on high level features and then finetuned based on middle and low-level features.

This is important because it allows the network to also learn the mid and low-level details of the image, in addition to high level ones.

Fully convolutional network

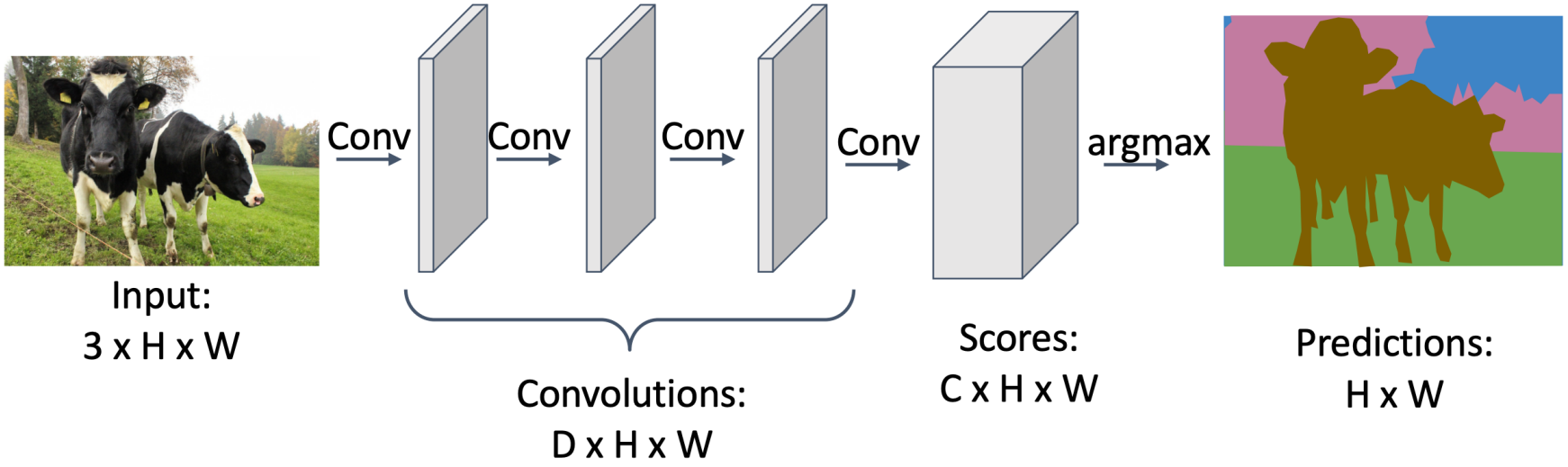


Good **Better** **Even better**

Long, Shelhamer, Darrell - Fully Convolutional Networks for Semantic Segmentation, CVPR 2015, PAMI 2016

Fully convolutional network

Design a network as a bunch of convolutional layers to make predictions for pixels all at once!



Loss function: Per-Pixel cross-entropy

Problem #1: Effective receptive field size is linear in number of conv layers: With L 3x3 conv layers, receptive field is 1+2L

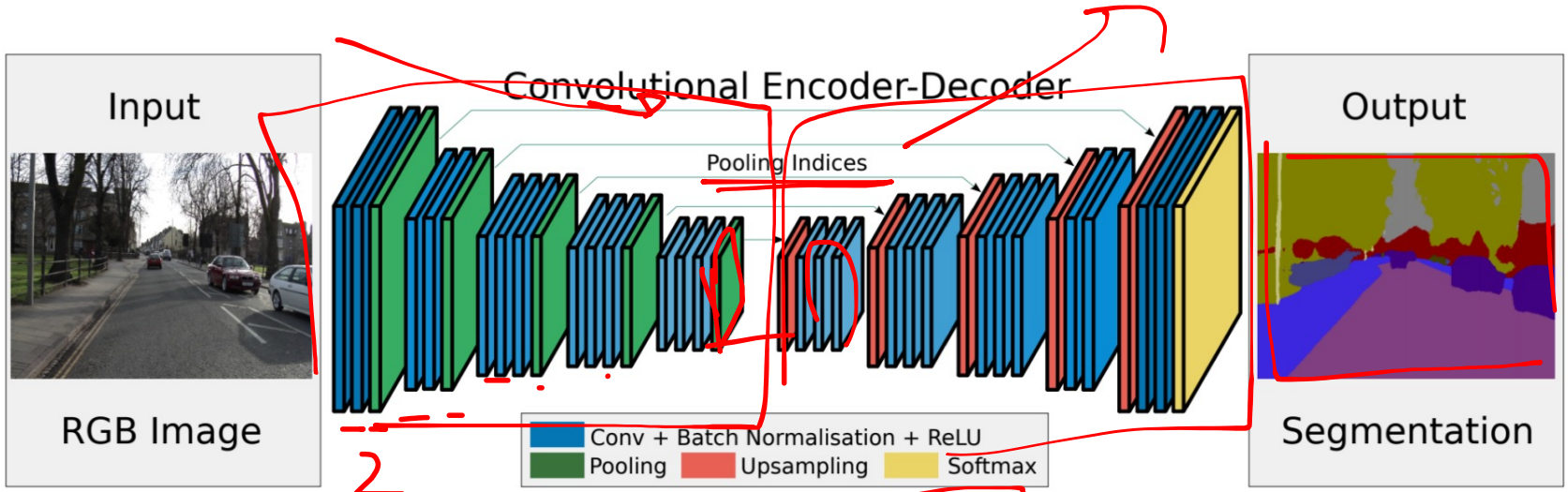
Problem #2: Convolution on high res images is expensive!

Long et al, "Fully convolutional networks for semantic segmentation", CVPR 2015

SegNet

Design network as a bunch of convolutional layers, with pooling (downsampling) and upsampling inside the network!

- **Encoder:** Normal convolutional filters + pooling
- **Decoder:** Upsampling + convolutional filters
- **Softmax layer:** The output of the soft-max classifier is a K channel image of probabilities where K is the number of classes

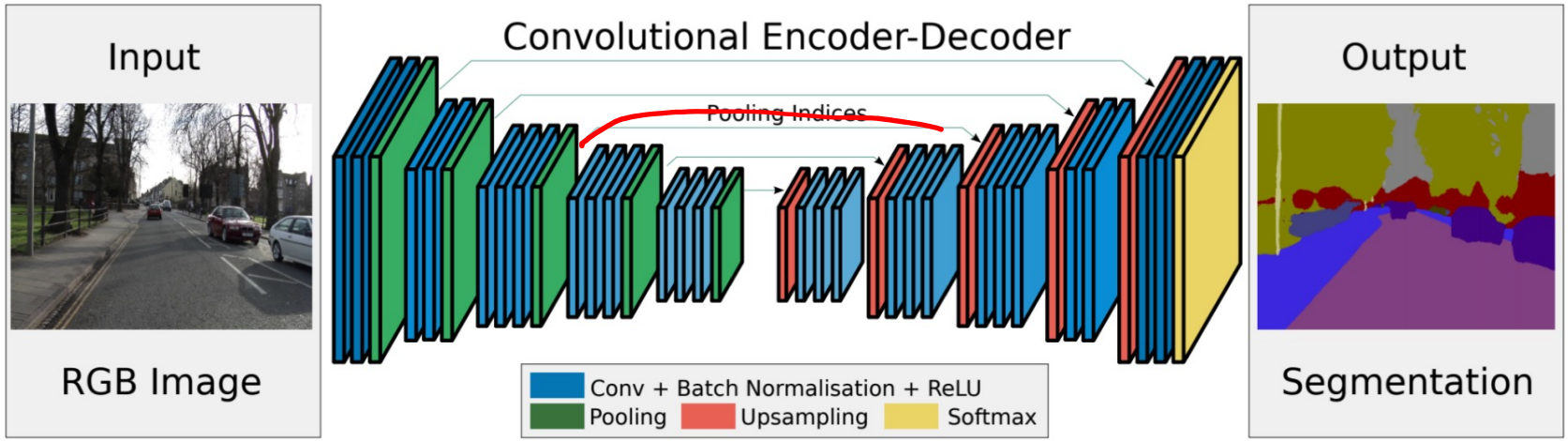


The convolutional filters in the decoder are learned using backprop and their goal is to refine the upsampling

SegNet

Design network as a bunch of convolutional layers, with pooling (downsampling) and upsampling inside the network!

- **Encoder:** Normal convolutional filters + pooling
- **Decoder:** Upsampling + convolutional filters
- **Softmax layer:** The output of the soft-max classifier is a K channel image of probabilities where K is the number of classes



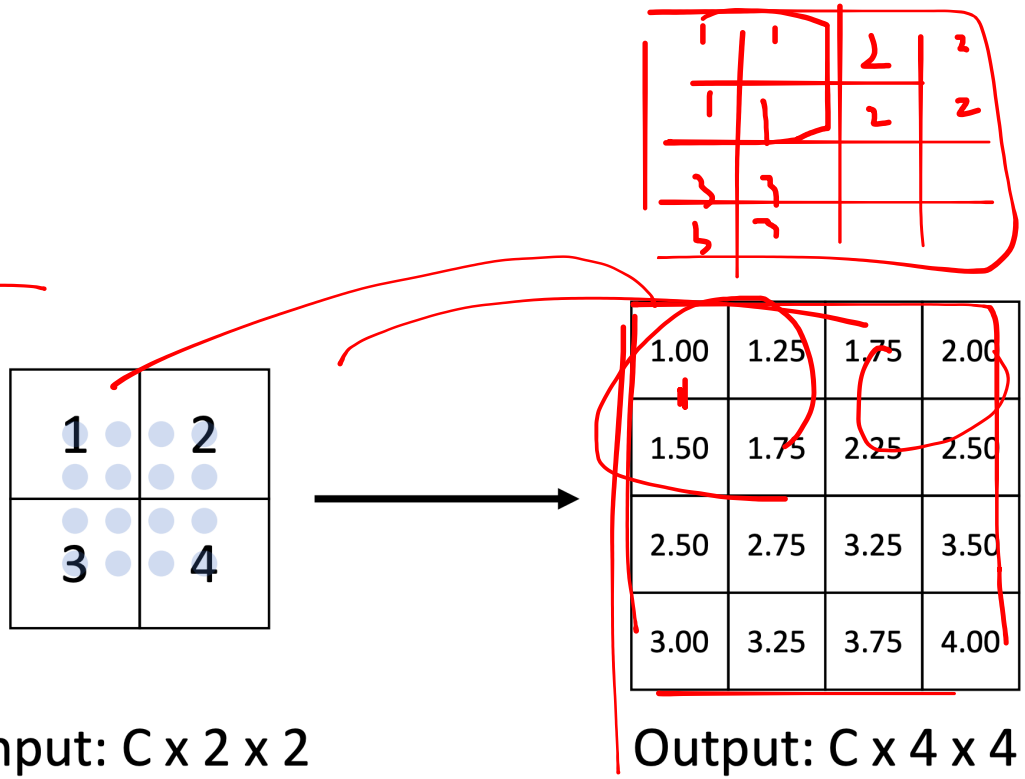
The convolutional filters in the decoder are learned using backprop and their goal is to refine the upsampling

Badrinarayanan et al. „SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation“. TPAMI 2016

SegNet

Upsampling

- Interpolation



Use two closest neighbors in x and y to construct linear approximations

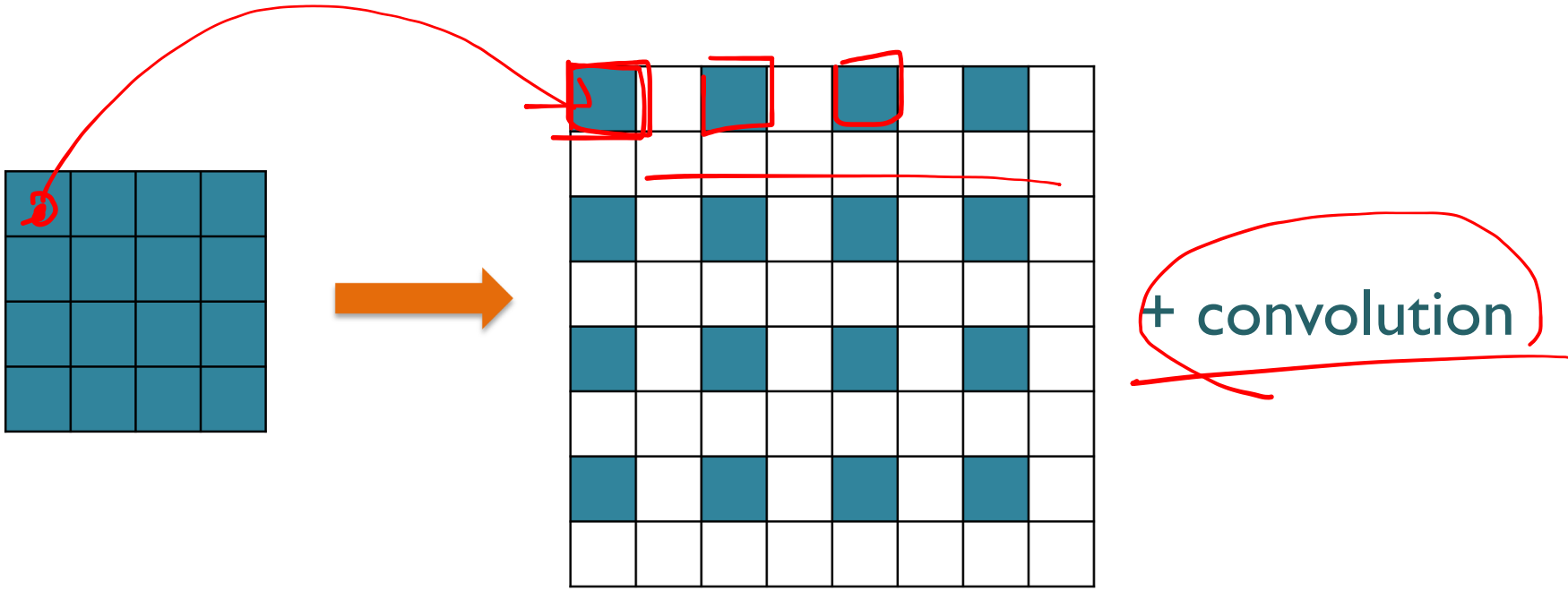
$$f_{x,y} = \sum_{i,j} f_{i,j} \max(0, 1 - |x - i|) \max(0, 1 - |y - j|) \quad \begin{array}{l} i \in \{\lfloor x \rfloor - 1, \ldots, \lceil x \rceil + 1\} \\ j \in \{\lfloor y \rfloor - 1, \ldots, \lceil y \rceil + 1\} \end{array}$$

Badrinarayanan et al. „SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation“. TPAMI 2016

SegNet

Upsampling

- Interpolation
- Fixed unpooling

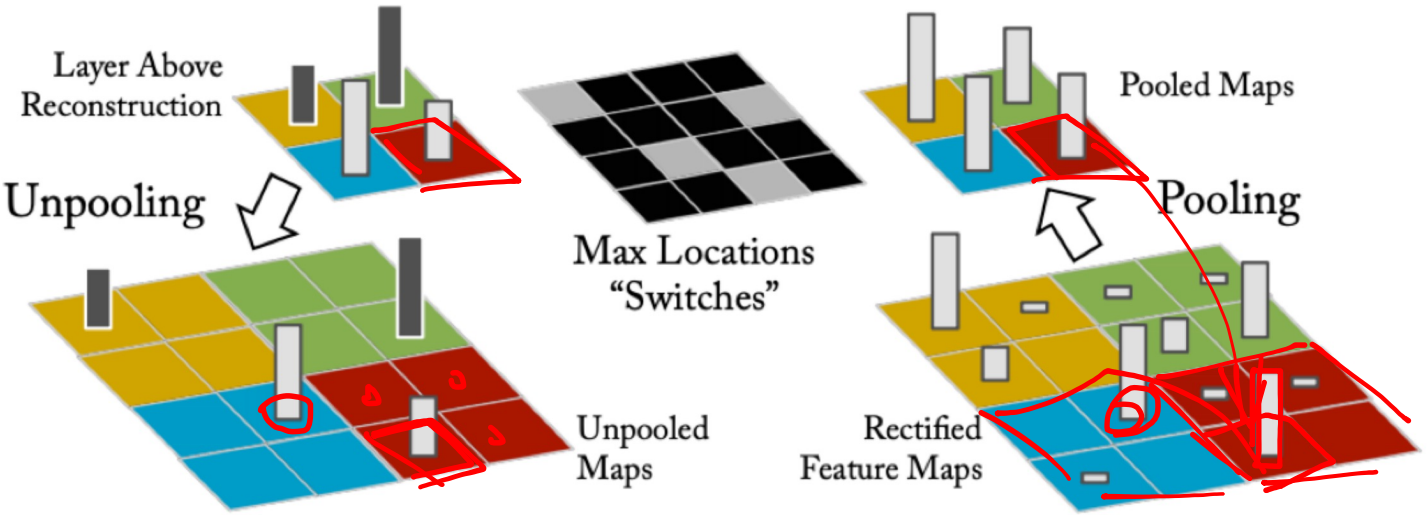


Badrinarayanan et al. „SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation“. TPAMI 2016

SegNet

Upsampling

- Interpolation
- Fixed unpooling
- Unpooling in DeconvNet

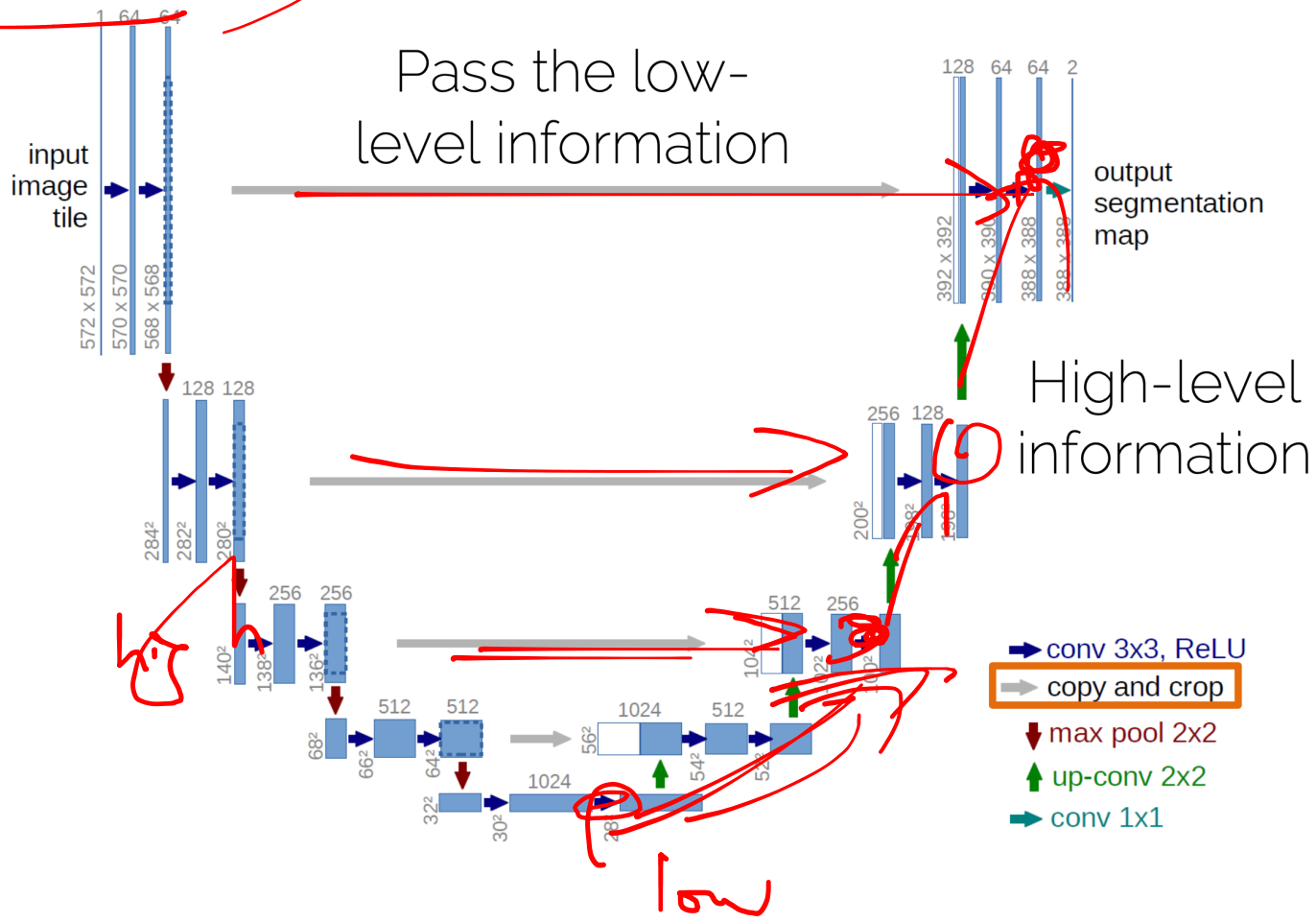


Keep the locations where the max came from

Badrinarayanan et al. „SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". TPAMI 2016

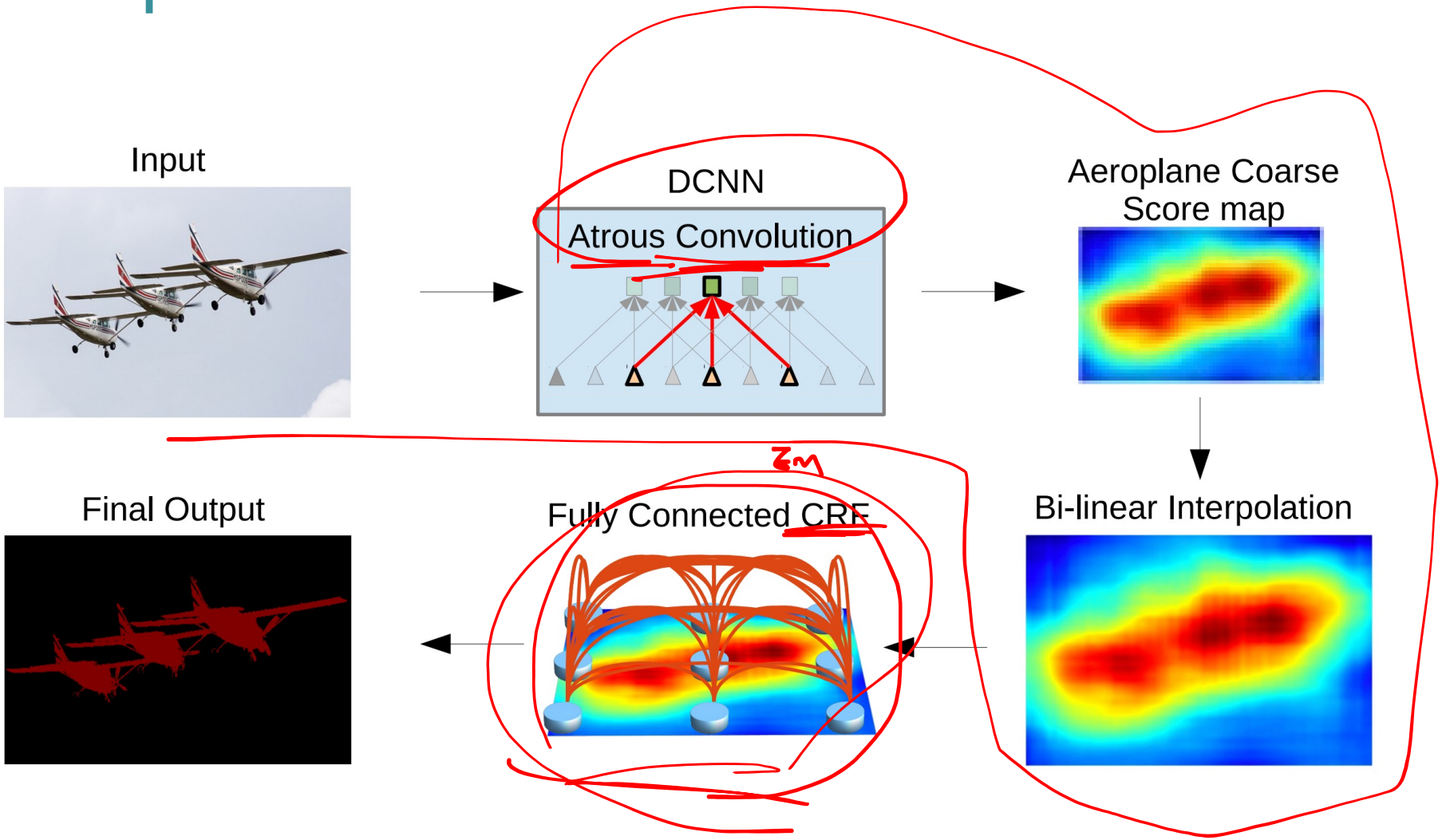
U-Net

Skip Connections



O. Ronneberger et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. MICCAI 2015

DeepLab



Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

Reduced feature resolution

Solution: Atrous convolutions

Objects exist at multiple scales

Solution: Pyramid pooling

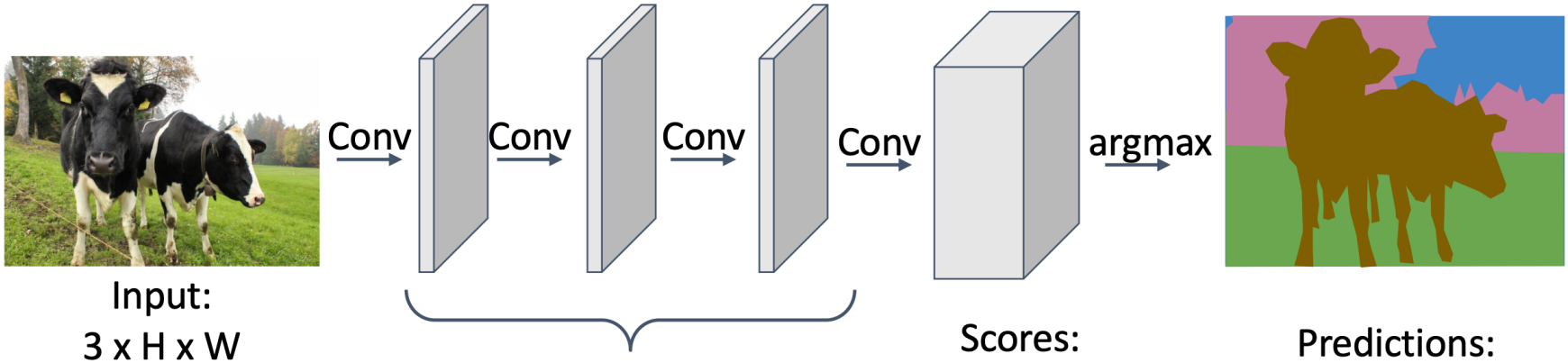
Poor localization of the edges

Solution: Refinement with Conditional Random Field (CRF)

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

Design a network as a bunch of convolutional layers to make predictions for pixels all at once!



Loss function: Per-Pixel cross-entropy

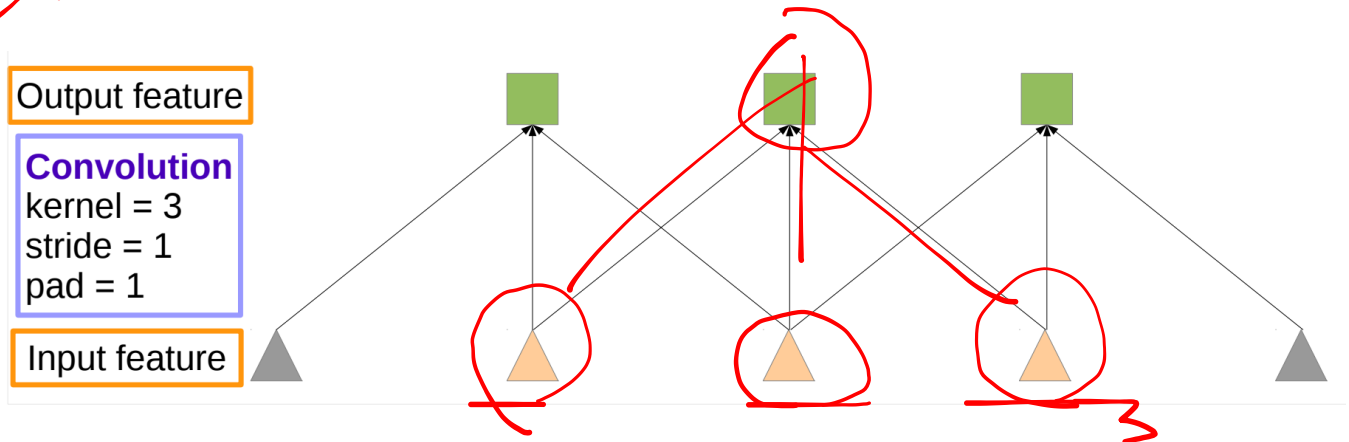
Problem #1: Effective receptive field size is linear in number of conv layers: With L 3x3 conv layers, receptive field is 1+2L

Problem #2: Convolution on high res images is expensive!

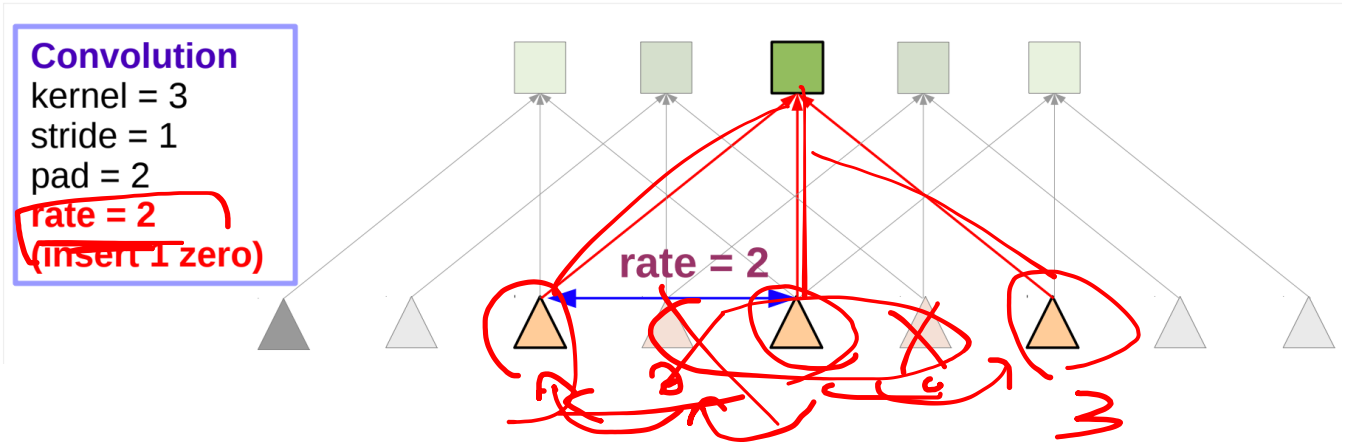
Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

Dilated (atrous) convolutions 1D

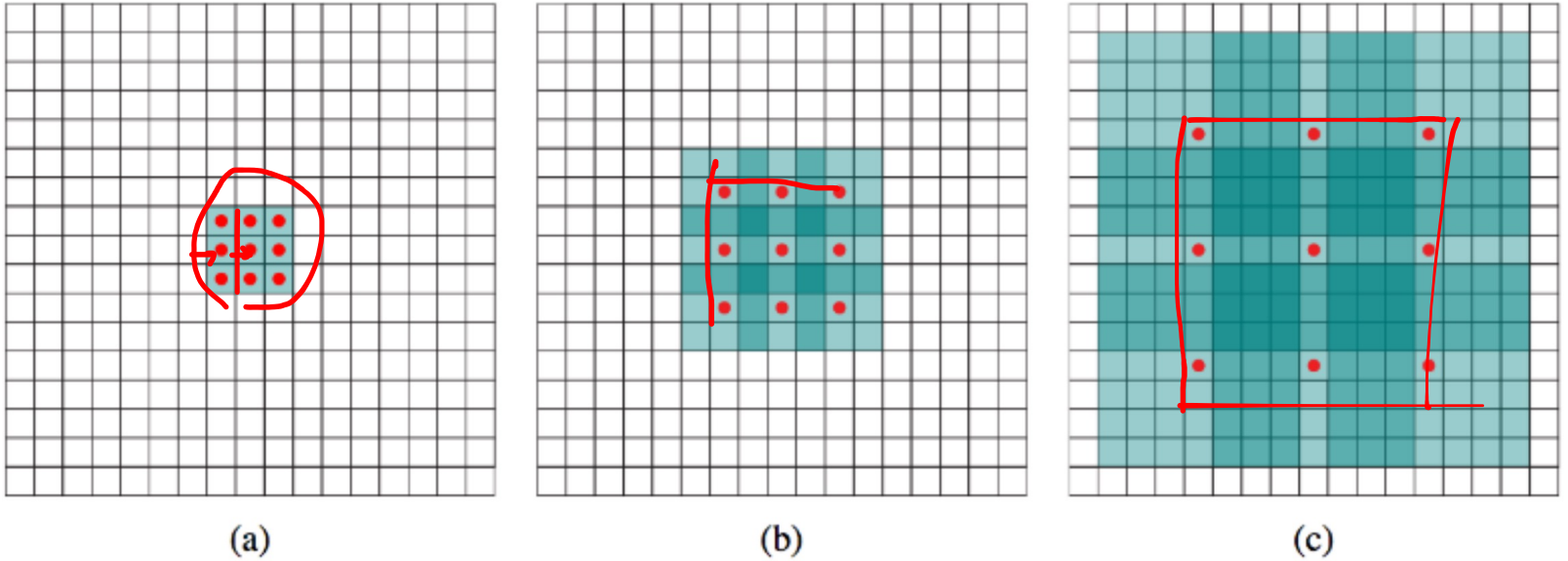


(a) Sparse feature extraction



DeepLab

Dilated (atrous) convolutions 2D



(a) the dilation parameter is 1, and each element produced by this filter has reception field of 3x3.

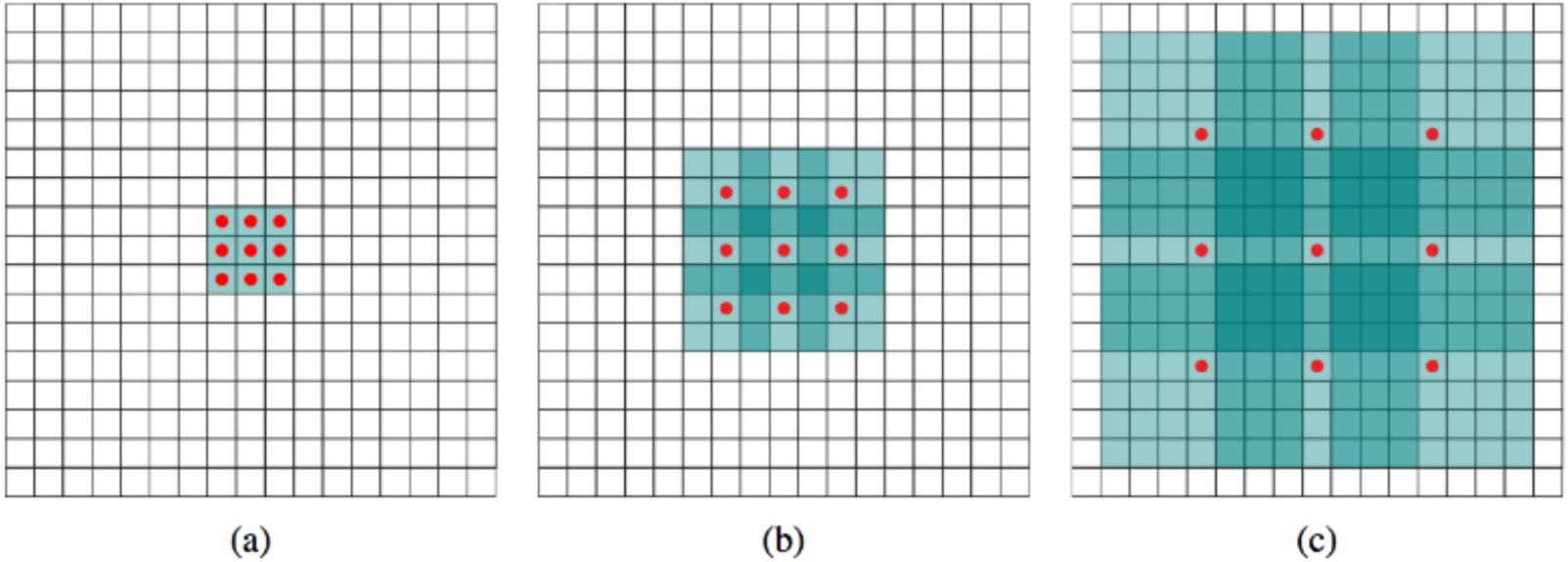
(b) the dilation parameter is 2, and each element produced by it has reception field of 7x7.

(c) the dilation parameter is 4, and each element produced by it has reception field of 15x15

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

Dilated (atrous) convolutions 2D

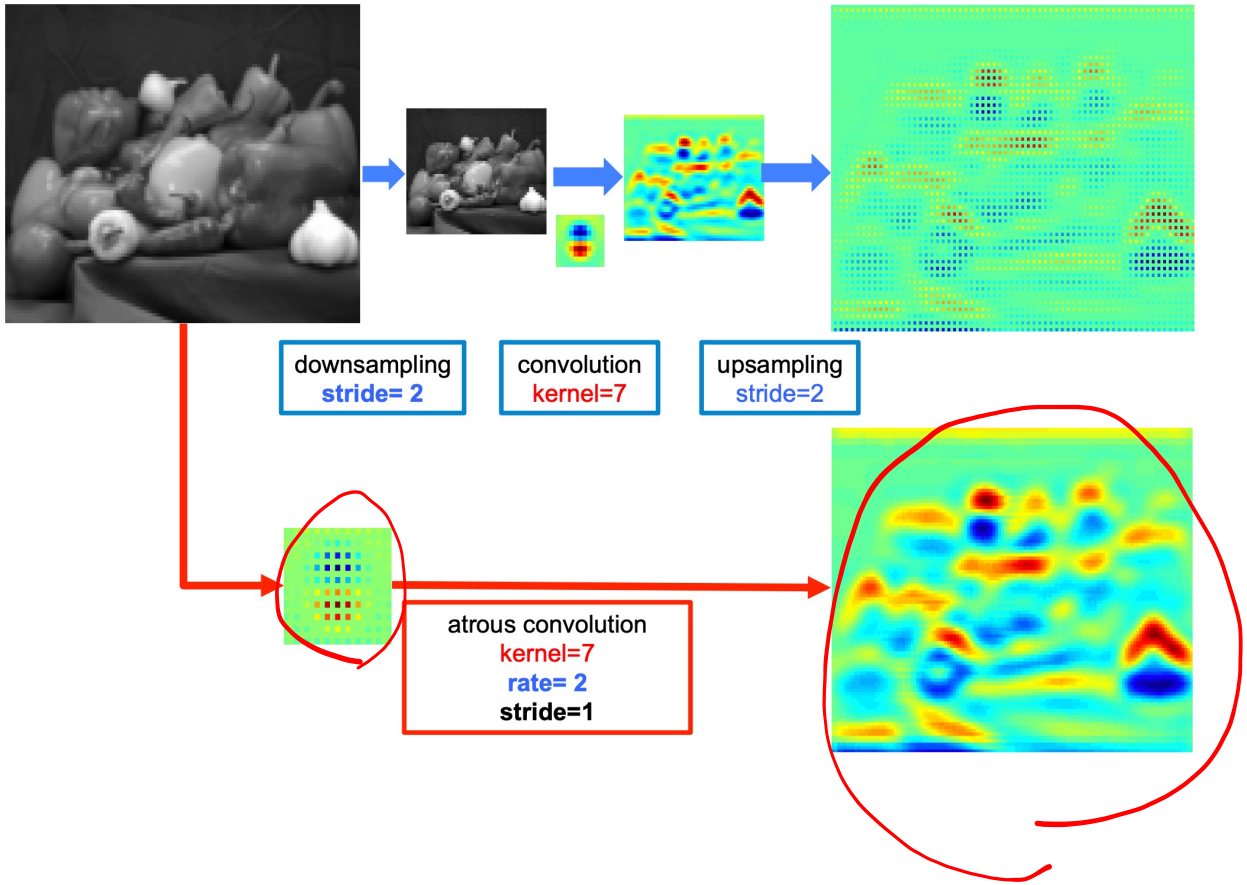


Each layer has the same number of parameters, but the receptive field grows exponentially while the number of parameters grows linearly.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

Dilated (atrous) convolutions



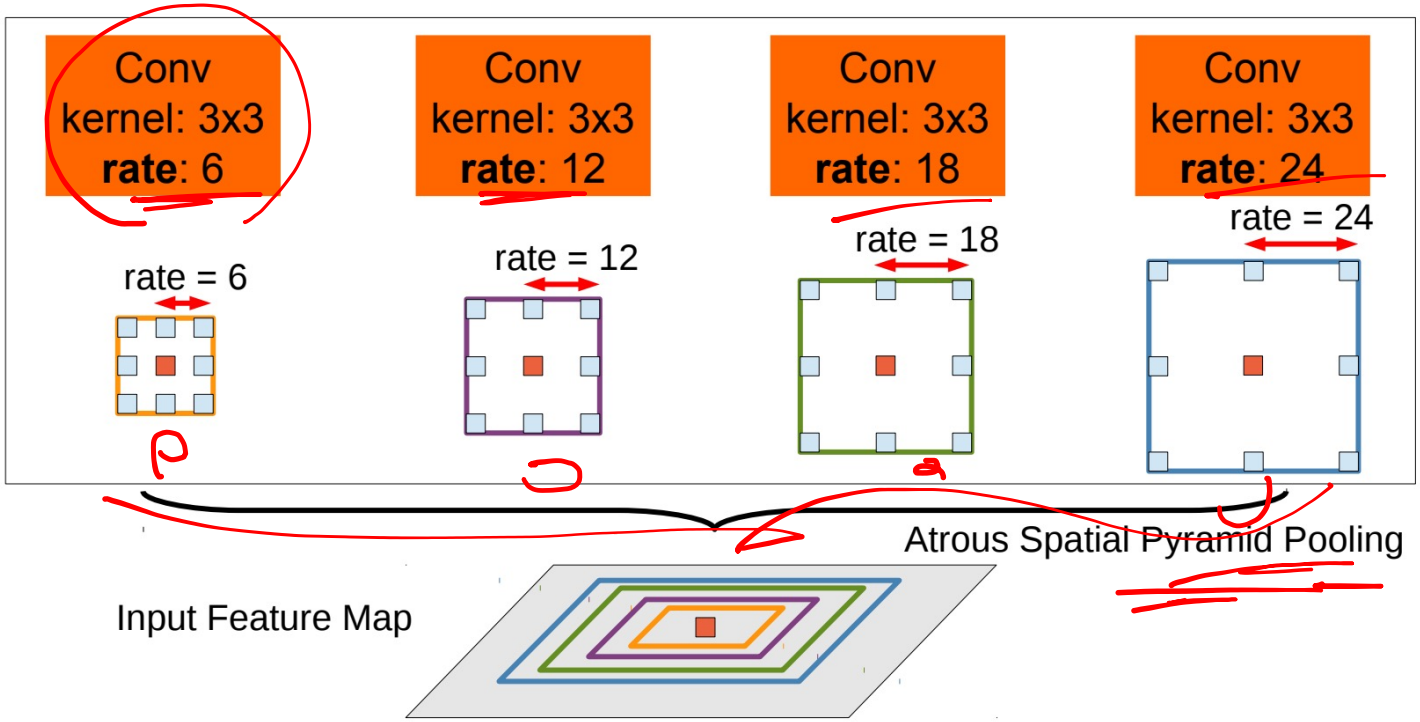
Sparse feature extraction with standard convolution on a low resolution input feature map

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

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

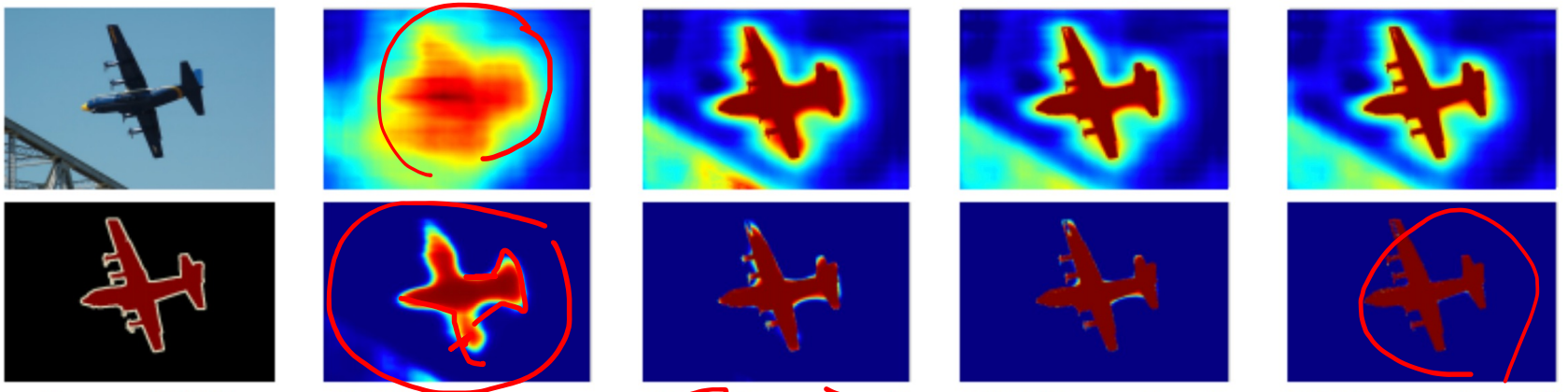
Atrous Spatial Pyramid Pooling



Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLab

Conditional Random Field (CRF)



Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. The image shows the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference.

DeepLab

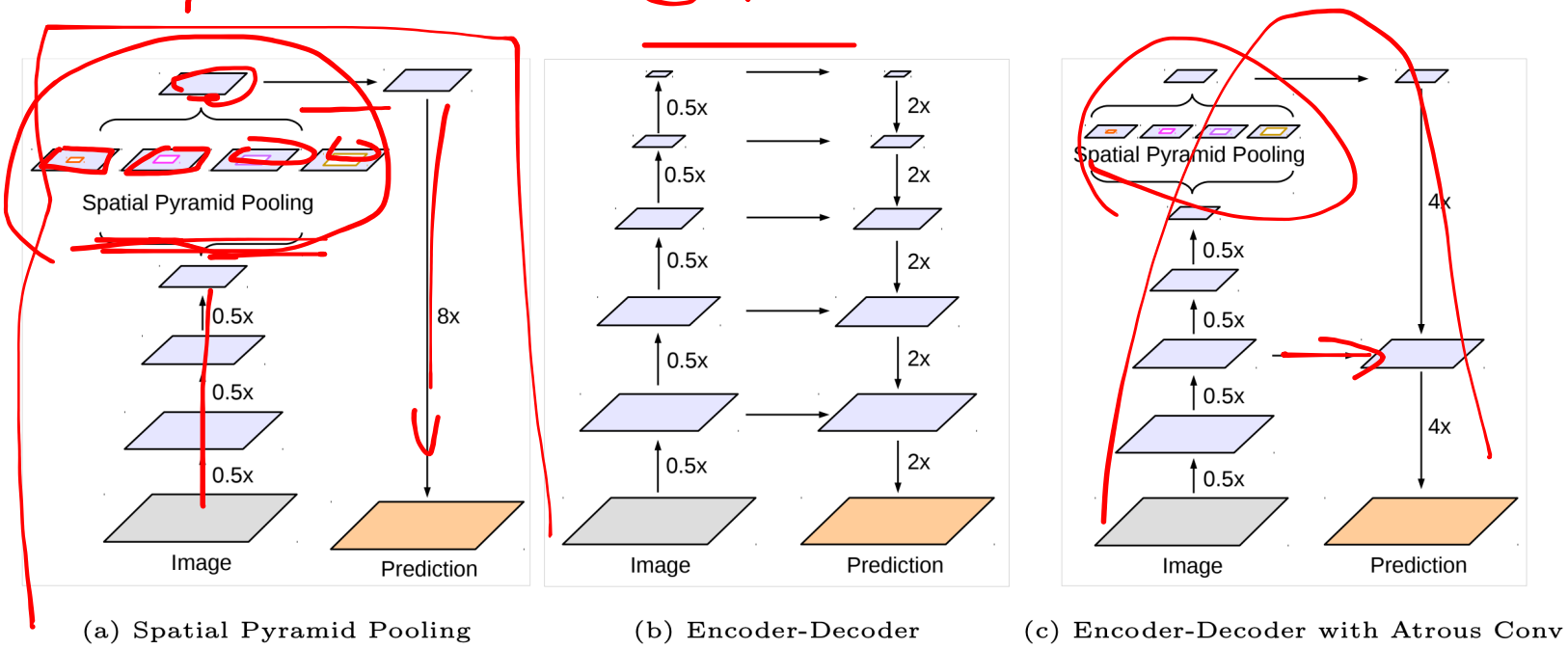
Fig. 6: PASCAL VOC 2012 *val* results. Input image and our DeepLab results before/after CRF.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, T-PAMI 2017

DeepLabv3

Multiresolution

atrous convolutions + spatial pyramid pooling + skip connection



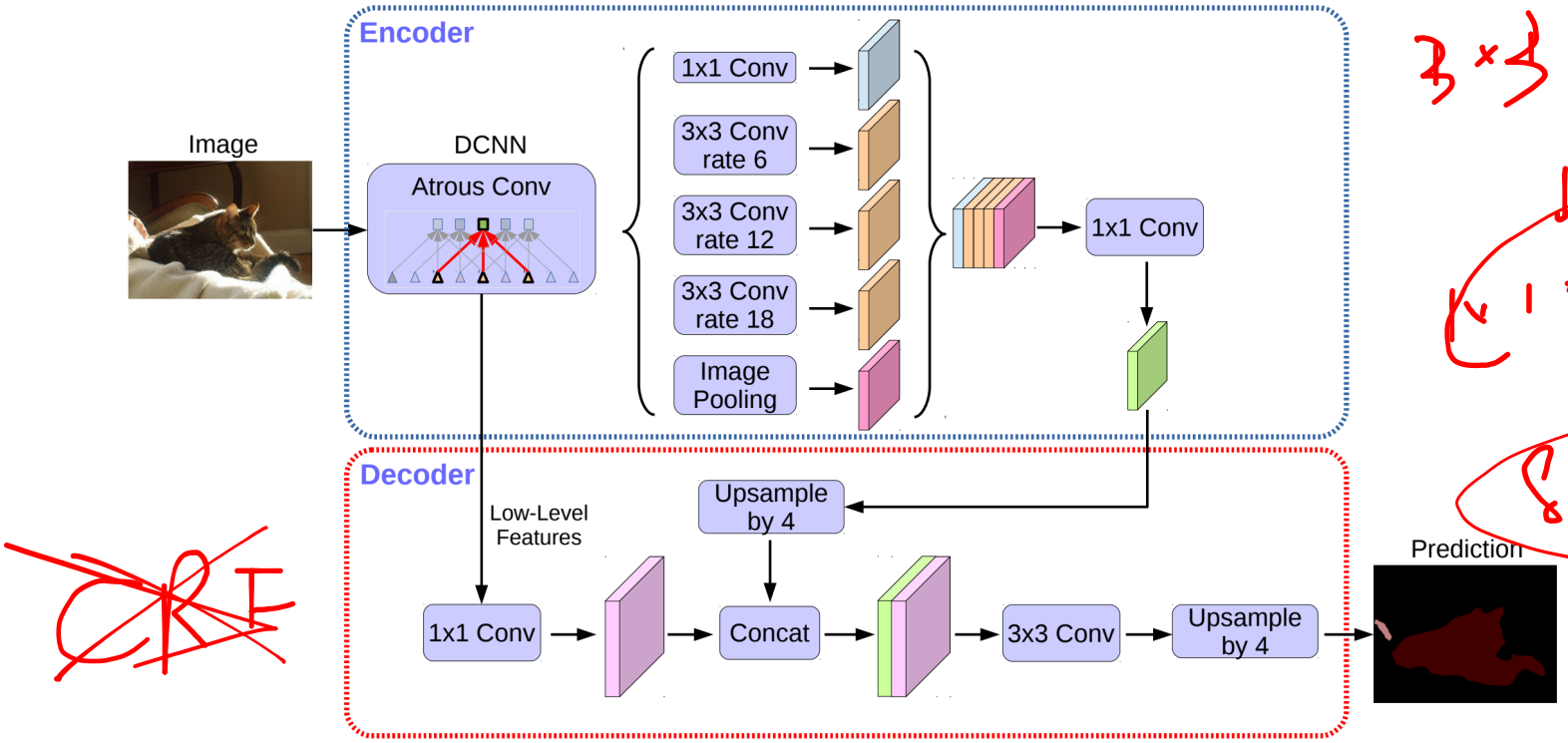
(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Combine atrous convolutions and spatial pyramid pooling with an encoder-decoder module.

DeepLabv3

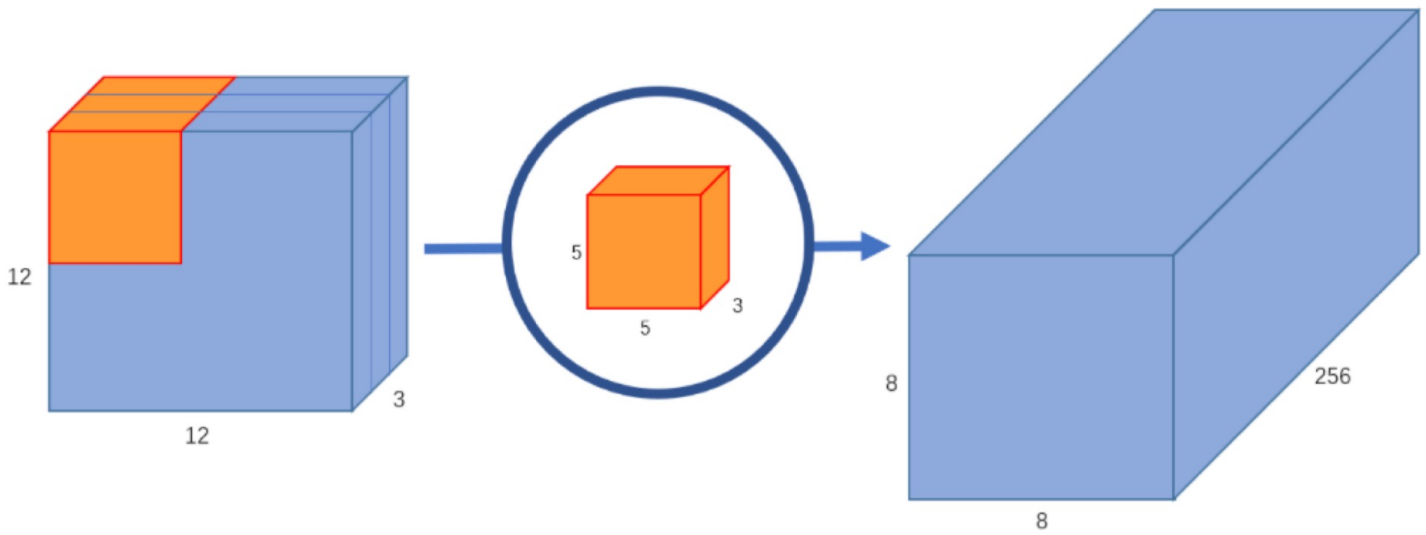
Network architecture

Encode multi-scale contextual information by applying atrous convolution at multiple scales



refine the segmentation results along object boundaries

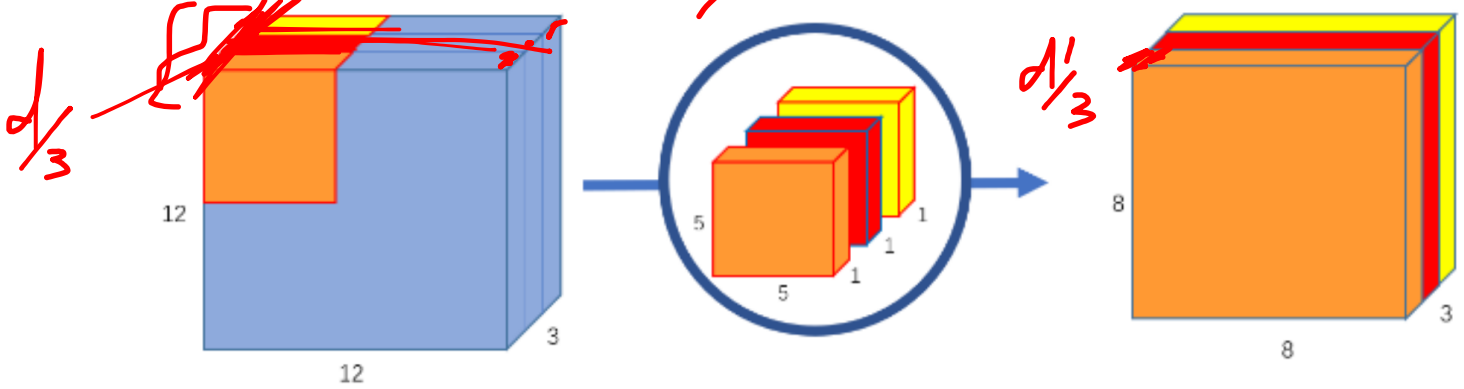
DeepLabv3



Normal convolutions act on all channels.

DeepLabv3

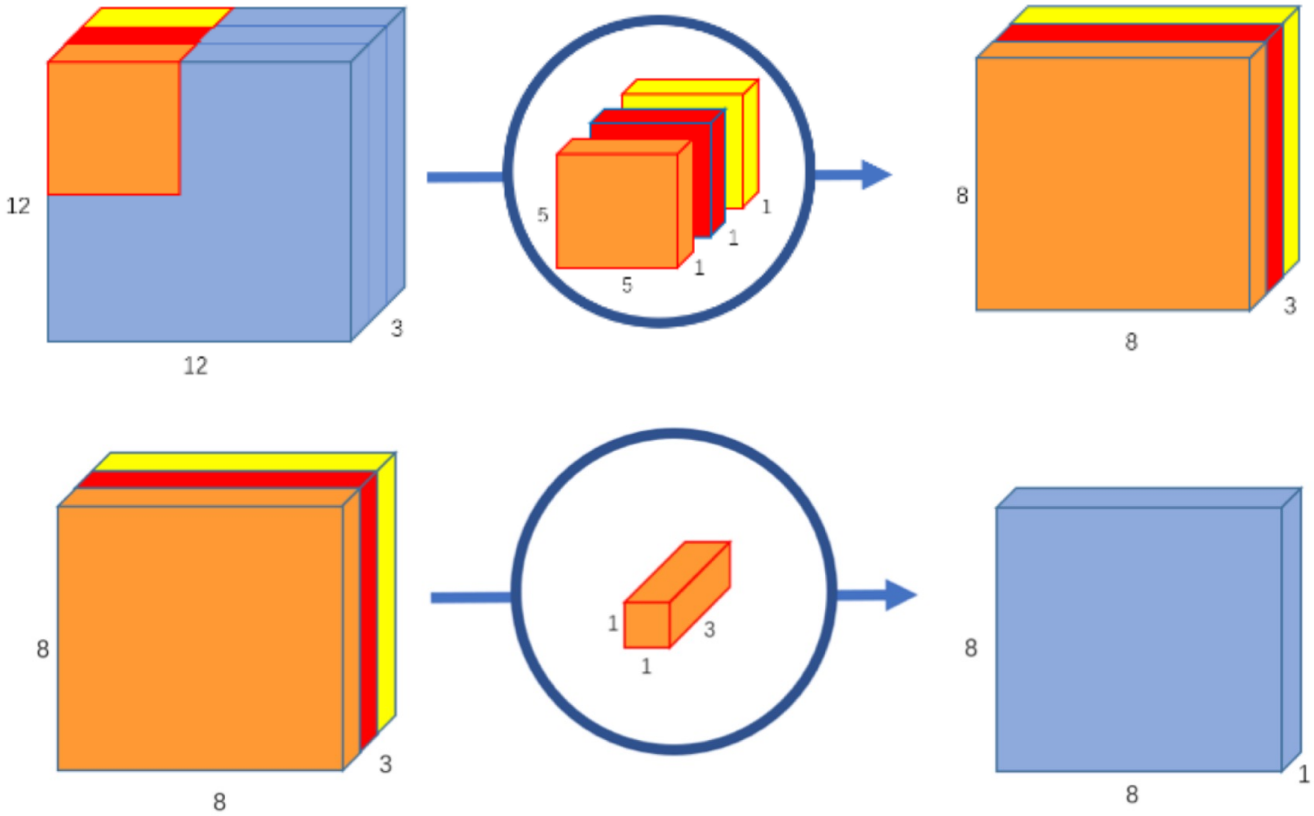
Problem: the depth size is always the same!



Filters are applied only at certain depths of the features.
Normal convolutions have groups set to 1, the convolutions used in this image have groups set to 3.

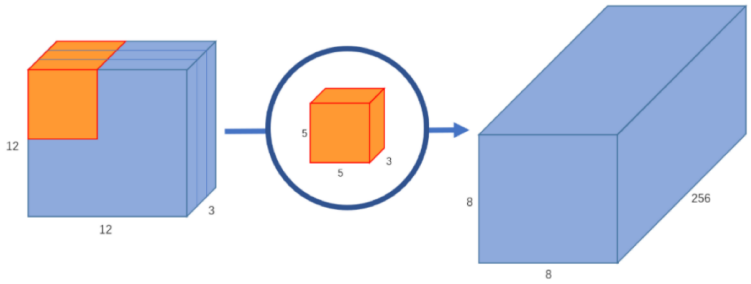
DeepLabv3

Problem: the depth size is always the same!



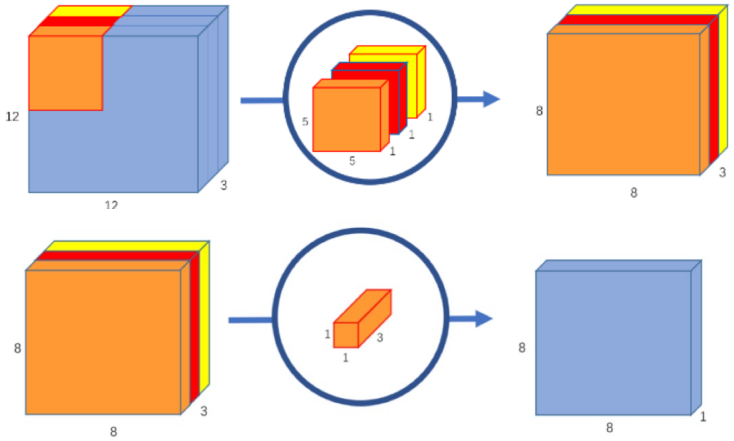
Solution: 1x1 convs!

DeepLabv3



Original convolution
256 kernels of size 5x5x3

Multiplications:
256x5x5x3 x (8x8 locations) = **1,228,800**



Depth-wise convolution
3 kernels of size 5x5x1

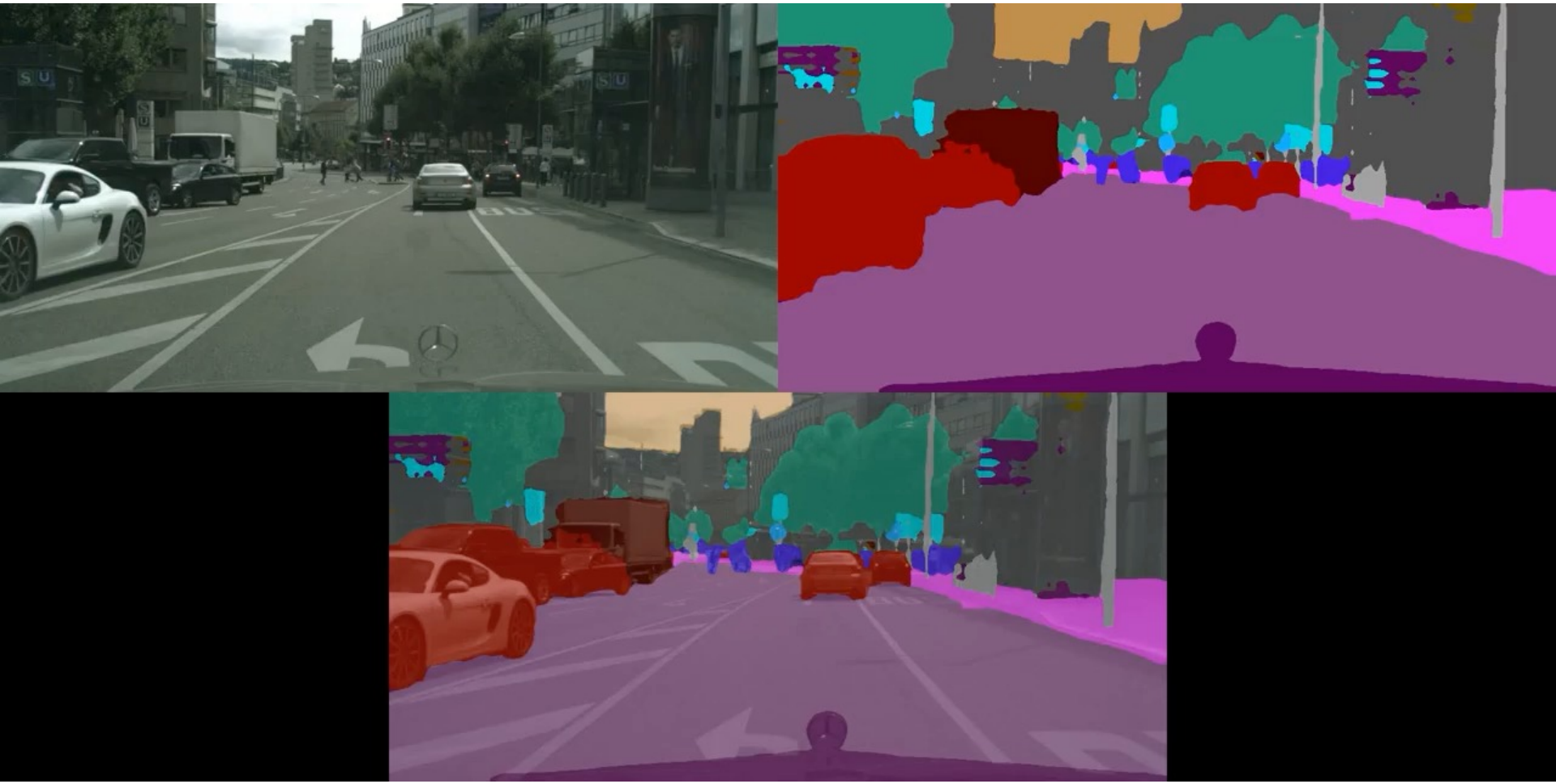
Multiplications:
5x5x3 x (8x8 locations) = 4800

1x1 convolution
256 kernels of size 1x1x3

Multiplications:
256x1x1x3x (8x8 locations) = **49,152**

Liang-Chieh Chen et al, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV 2018

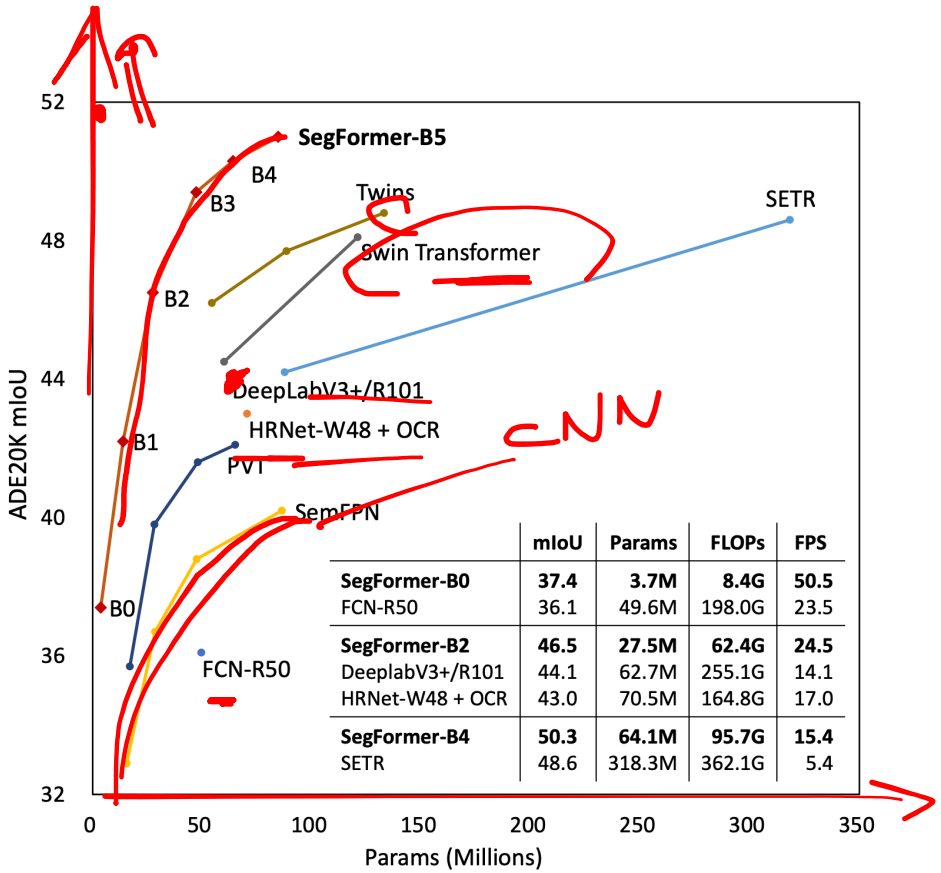
DeepLabv3



Liang-Chieh Chen et al, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV 2018

Semantic segmentation

SegFormer



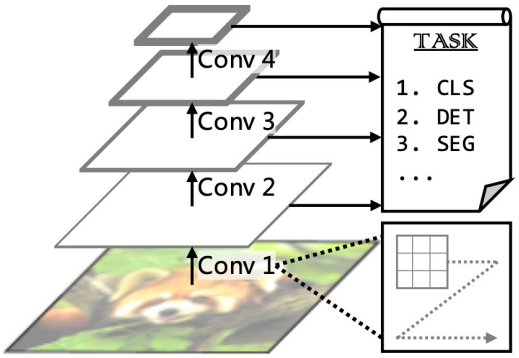
	mIoU	Params	FLOPs	FPS
SegFormer-B0	**37.4**	**3.7M**	**8.4G**	**50.5**
FCN-R50	36.1	49.6M	198.0G	23.5
SegFormer-B2	**46.5**	**27.5M**	**62.4G**	**24.5**
DeeplabV3+/R101	44.1	62.7M	255.1G	14.1
HRNet-W48 + OCR	43.0	70.5M	164.8G	17.0
SegFormer-B4	**50.3**	**64.1M**	**95.7G**	**15.4**
SETR	48.6	318.3M	362.1G	5.4

Figure 1: **Performance *vs.* model efficiency on ADE20K.** All results are reported with single model and single-scale inference. SegFormer achieves a new state-of-the-art 51.0% mIoU while being significantly more efficient than previous methods.

Enze Xie, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, 2021

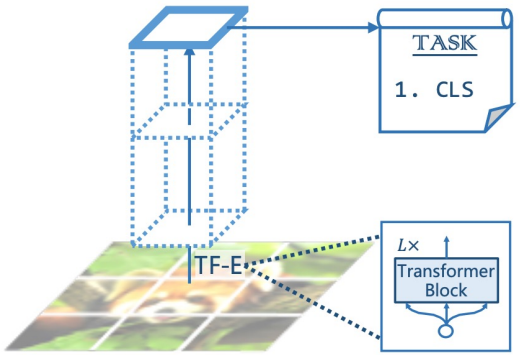
Semantic segmentation

SegFormer



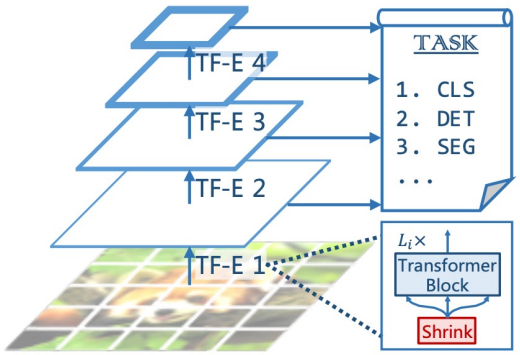
(a) CNNs: VGG [54], ResNet [22], *etc.*

Many CNN backbones use a pyramid structure for dense prediction tasks such as object detection, instance and semantic segmentation



(b) Vision Transformer [13]

a "columnar" structure specifically designed for image classification (CLS)

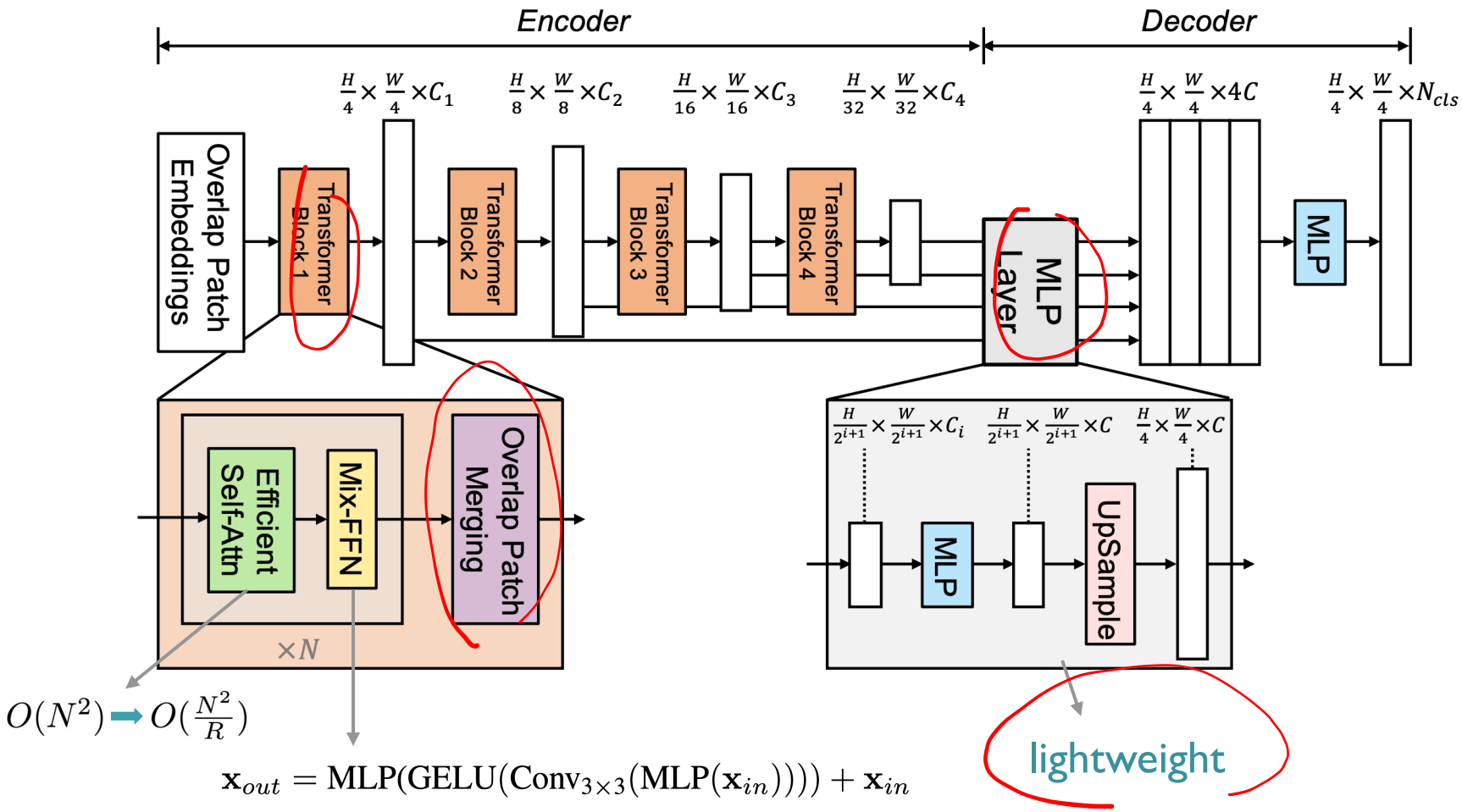


(c) Pyramid Vision Transformer (ours)

by incorporating the pyramid structure from CNNs, the Pyramid Vision Transformer (PVT) can be used as a versatile backbone for many computer vision tasks, broadening the scope and impact of ViT

Semantic segmentation

SegFormer



Enze Xie et al, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, NeurIPS 2021

Semantic segmentation

SegFormer

- For semantic segmentation, maintaining large receptive field to include context information has been a central issue
- decoder design benefits from the non-local attention in Transformers and leads to a larger receptive field without being complex

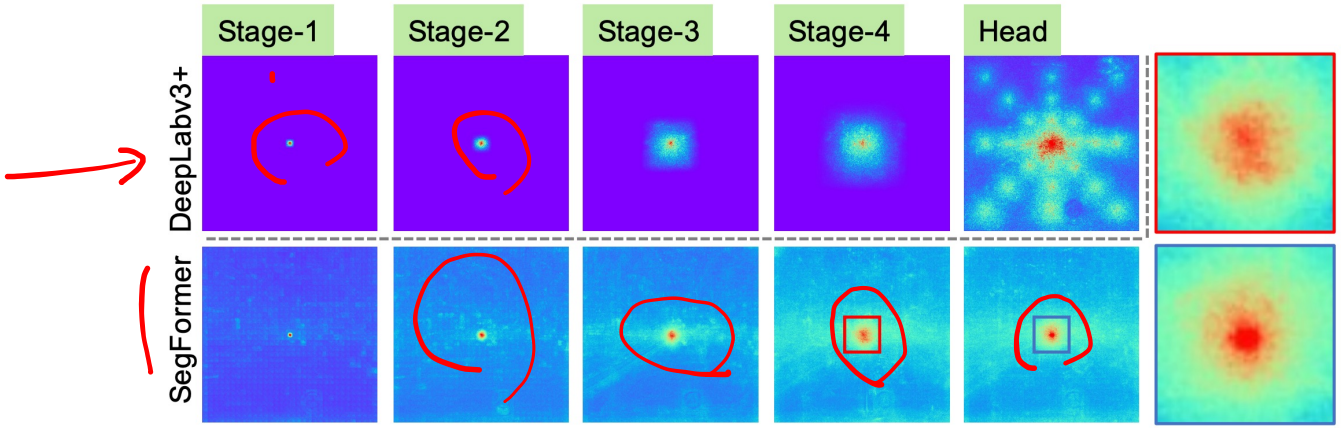


Figure 3: **Effective Receptive Field (ERF) on Cityscapes** (average over 100 images). Top row: Deeplabv3+. Bottom row: SegFormer. ERFs of the four stages and the decoder heads of both architectures are visualized. Best viewed with zoom in.

Enze Xie et al, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, NeurIPS 2021

Next lecture: Video understanding

Thank you very much!

sihengc@sjtu.edu.cn