# Semantic segmentation

Instructor: Seunghoon Hong
School of Computing, KAIST

# Announcement

- Assignment 1 due is **midnight September 23**

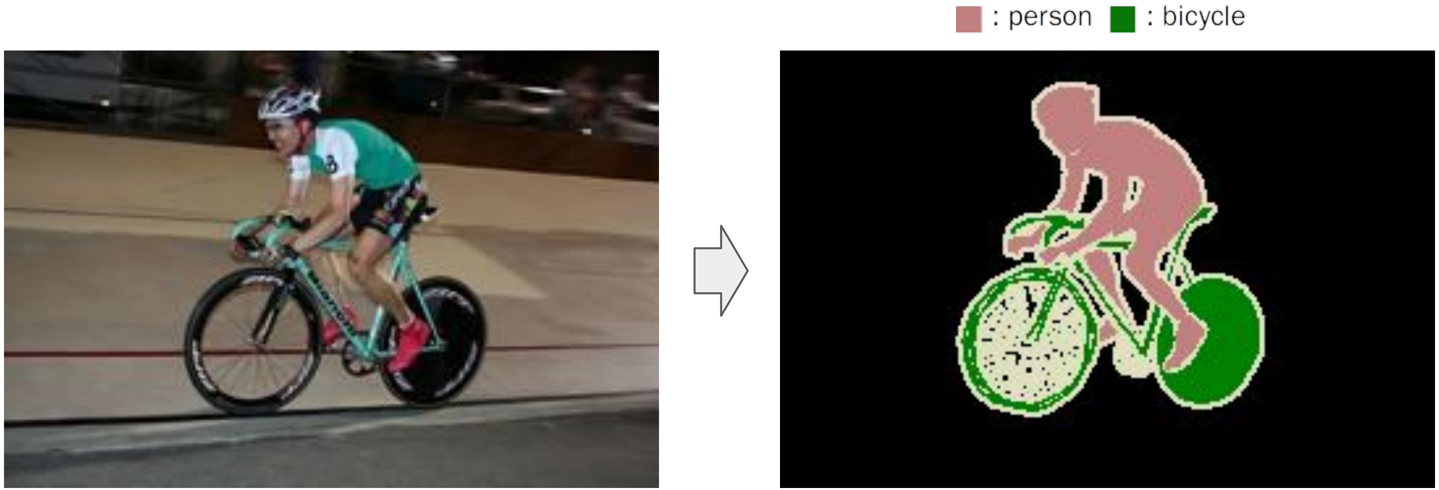
# Recap: Image Classification

- Recognition of visual concepts on an image

Is there a bicycle? **Yes**
Is there a person? **Yes**
Is there a car? **No**

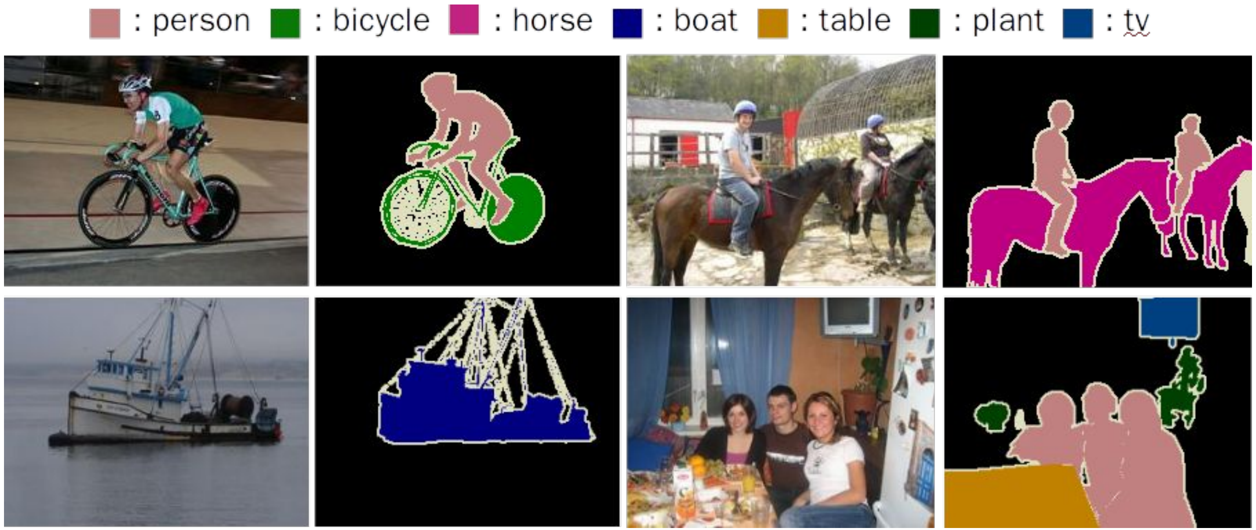
# Semantic segmentation

- Recognition of visual concepts on an image
- Recognition ***and pixel-level localization*** of visual concepts on an image



# Semantic segmentation

- Training data
  - Each image in training set is associated with pixel-level class labels
  - How can we learn to generate per-pixel class label given these training data?



# Problems

- Hand-designed representation

Remember this guy?

- Large search space for labeling
  - N: number of pixels
  - C: number of classes
  - Total $C^N$ possible labeling

Example: search space on small image

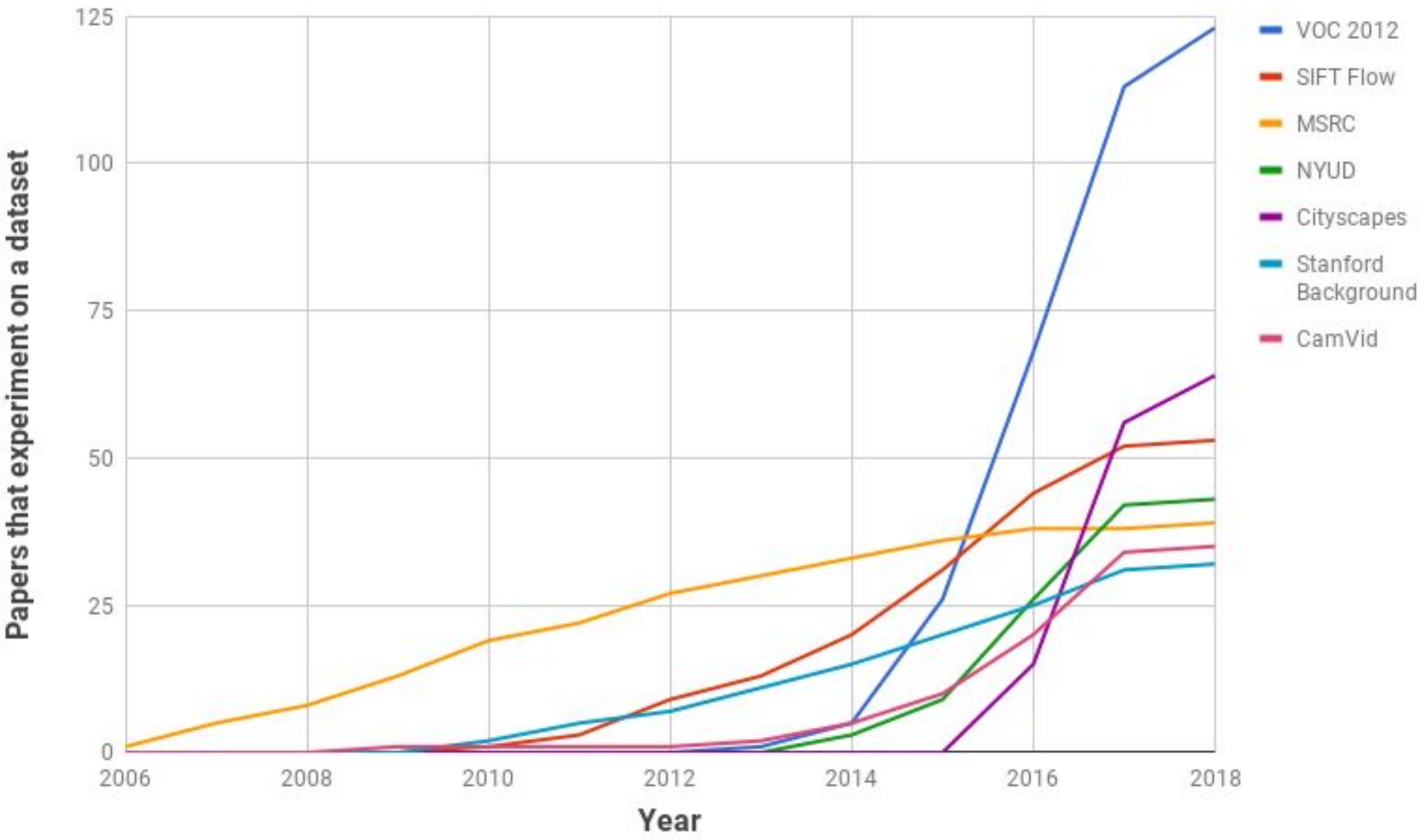
$N = 32 \times 32 = 1024$

$C = 20$

Size of possible label combinations:

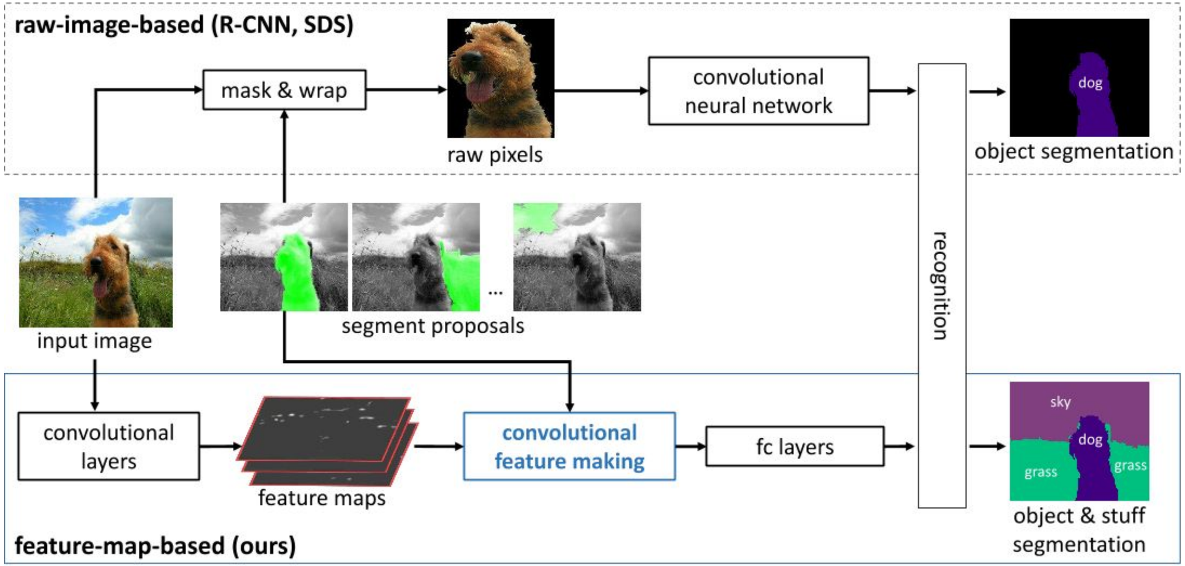
$$20^{1024}$$

# Accumulated dataset importance



# Semantic segmentation with CNN

- Early approaches
  - Region-based proposal + classification



# Semantic segmentation with CNN

- Early approaches
  - Region-based proposal + classification
- Limitations?

# Semantic segmentation with CNN

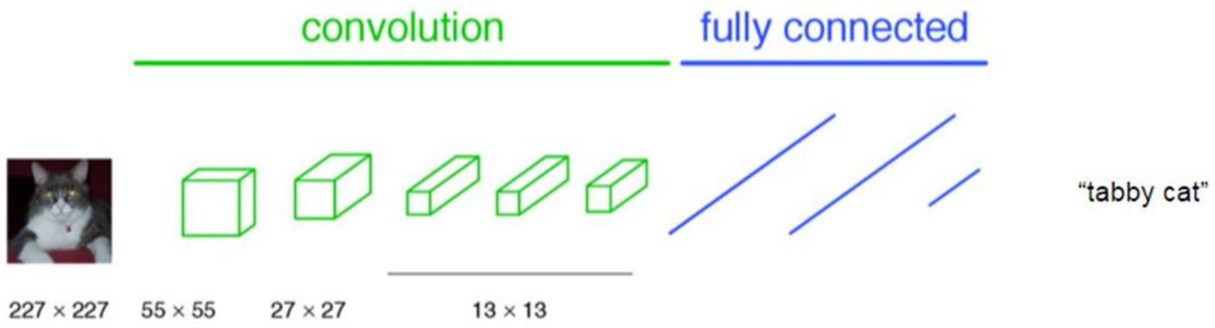
- Early approaches
  - Region-based proposal + classification

- Limitations?
  - The segmentation performance is determined by region-proposal accuracy
  - The models often employ separate classifier + feature extractor,
    which can be improved by end-to-end training

→ How can we design an ***end-to-end***, ***pixel-level prediction*** network?

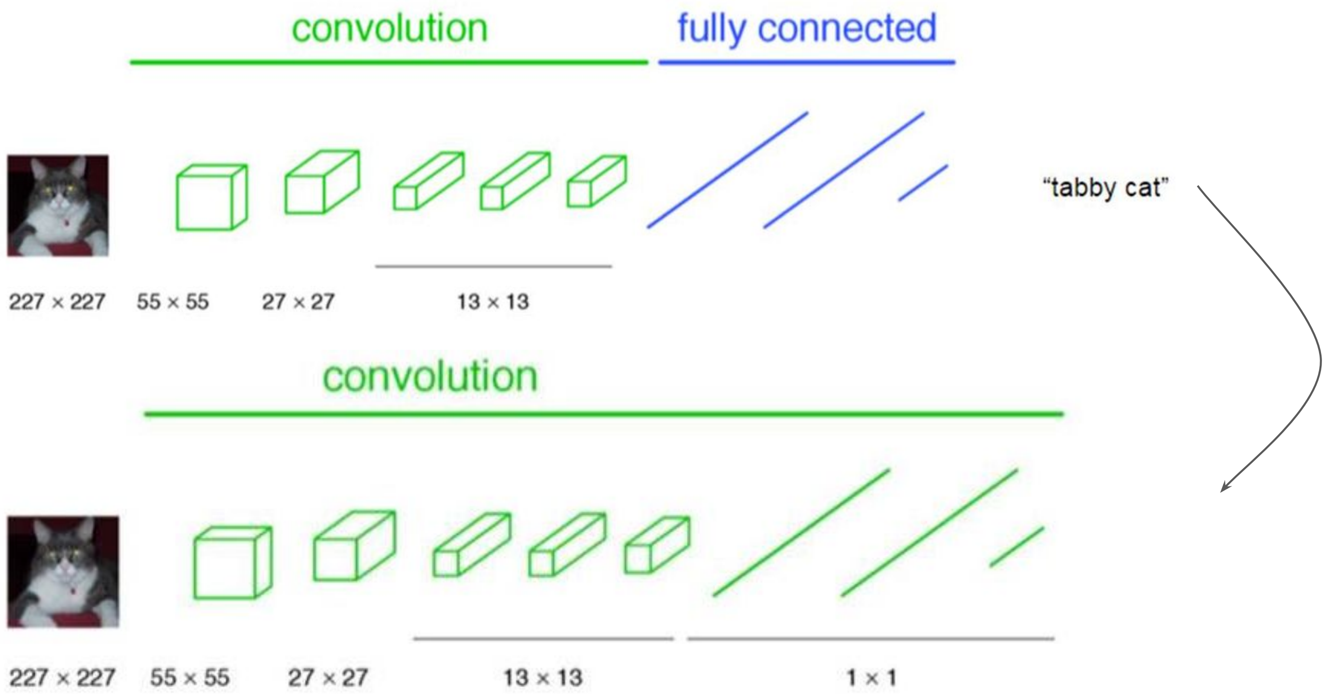
# Revisit: convnet for image classification

- Combination of convolutional + fully connected layers
  - Convolutional layers: operation is based on filtering.
    It takes an input in ***arbitrary size***,
    and the produce outputs preserving spatial information.
  - Fully-connected layers: operation is based on matrix multiplication.
    It takes a input in ***fixed size***, and produce fixed-sized output vector.

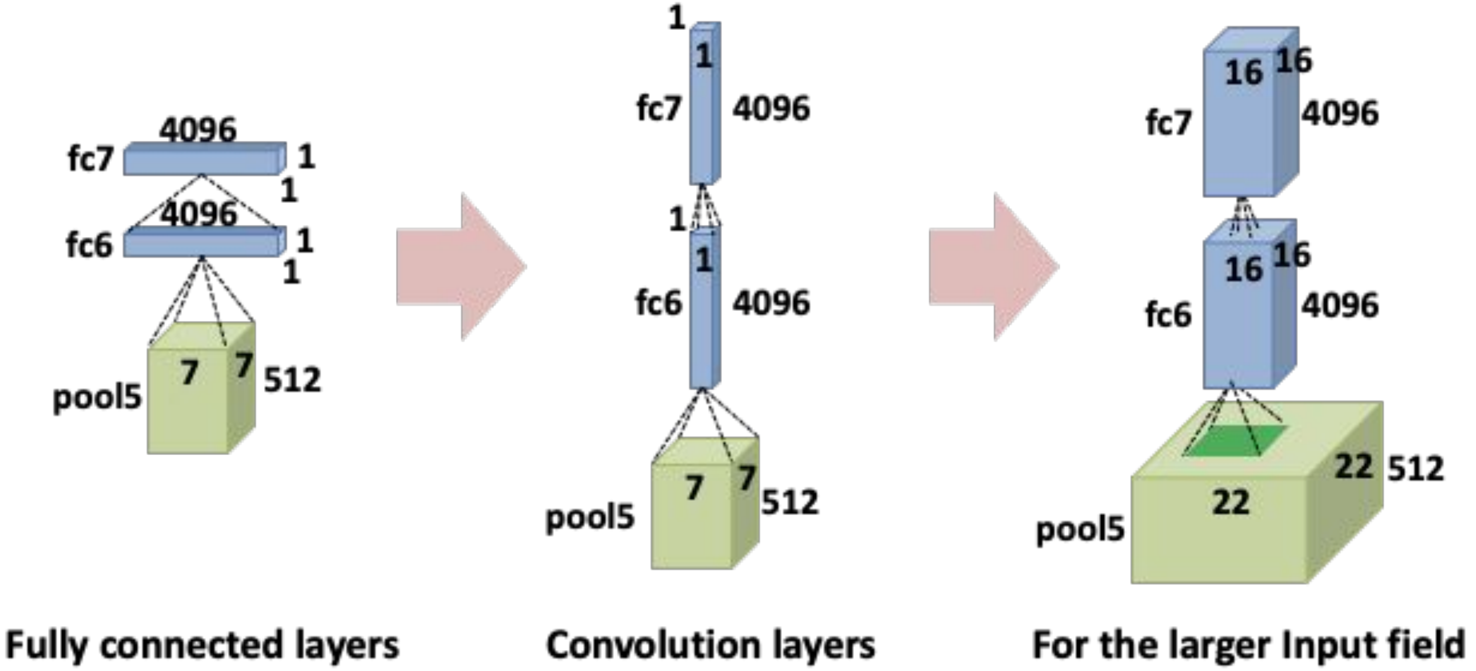


# Fully convolutional network

- Interpreting fully-connected layers by 1x1 convolution.

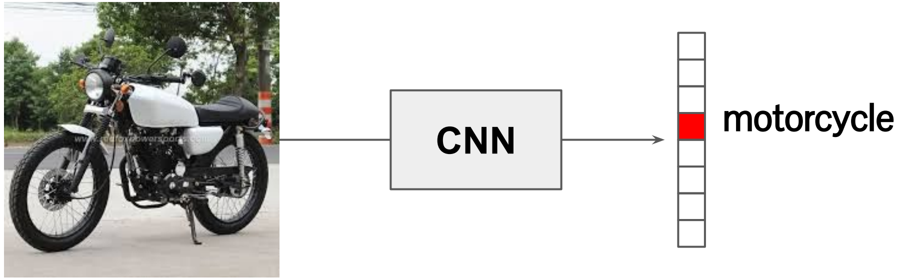


# Fully connected layer as convolution




Slide credit: Bohyung Han

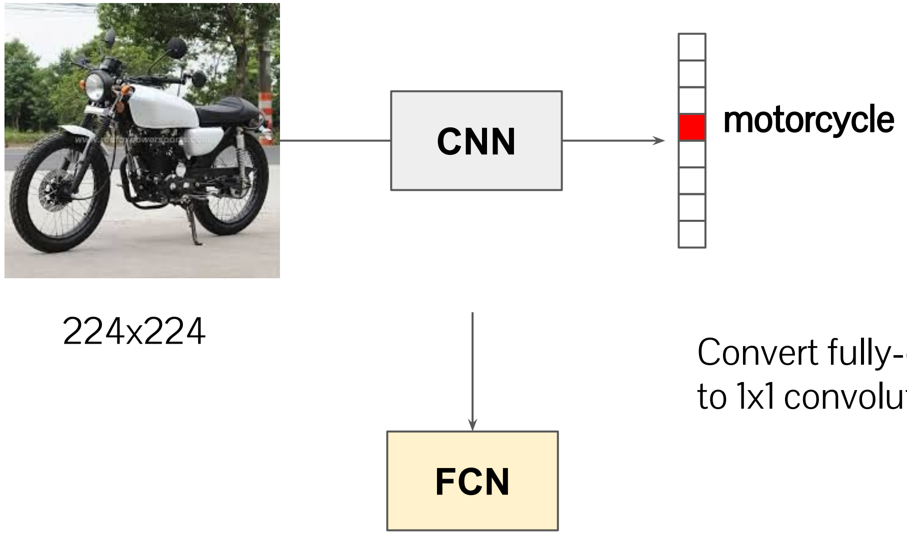
# Fully Convolutional Network




224x224

1. Pre-train a **CNN** for classification

# Fully Convolutional Network

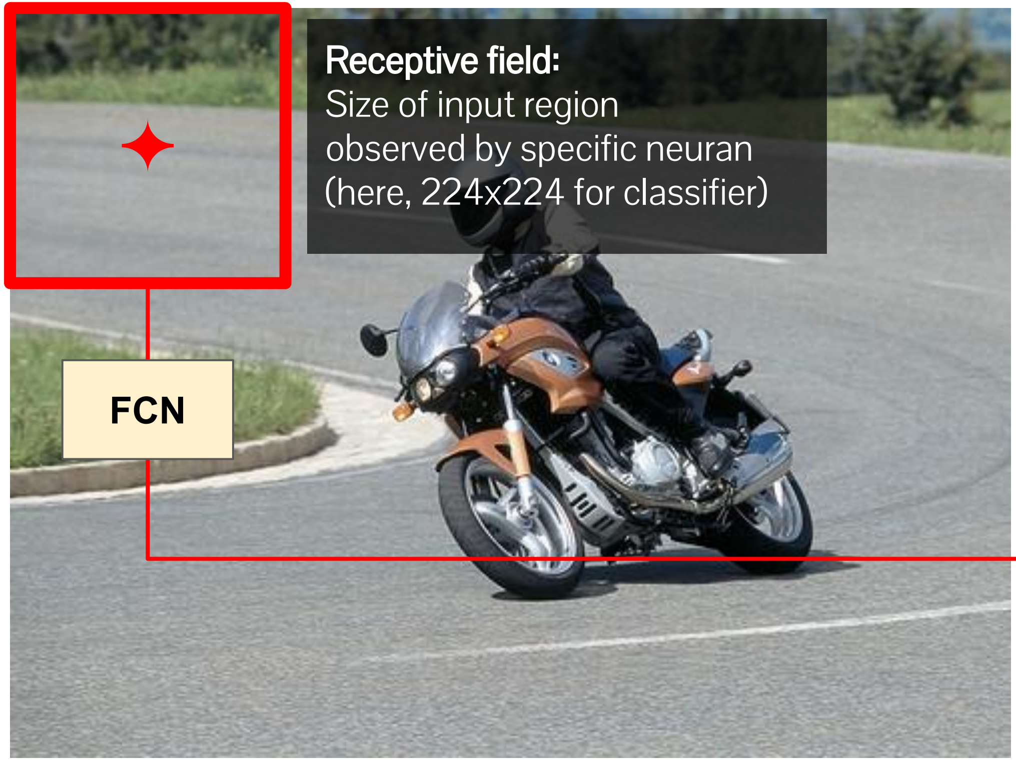



Convert fully-connected layer
to 1x1 convolution

1. Pre-train a CNN for classification

2. Convert CNN to FCN

# Fully Convolutional Network

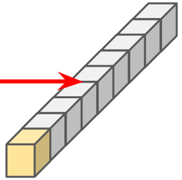



540x540

1. Pre-train a **CNN** for classification

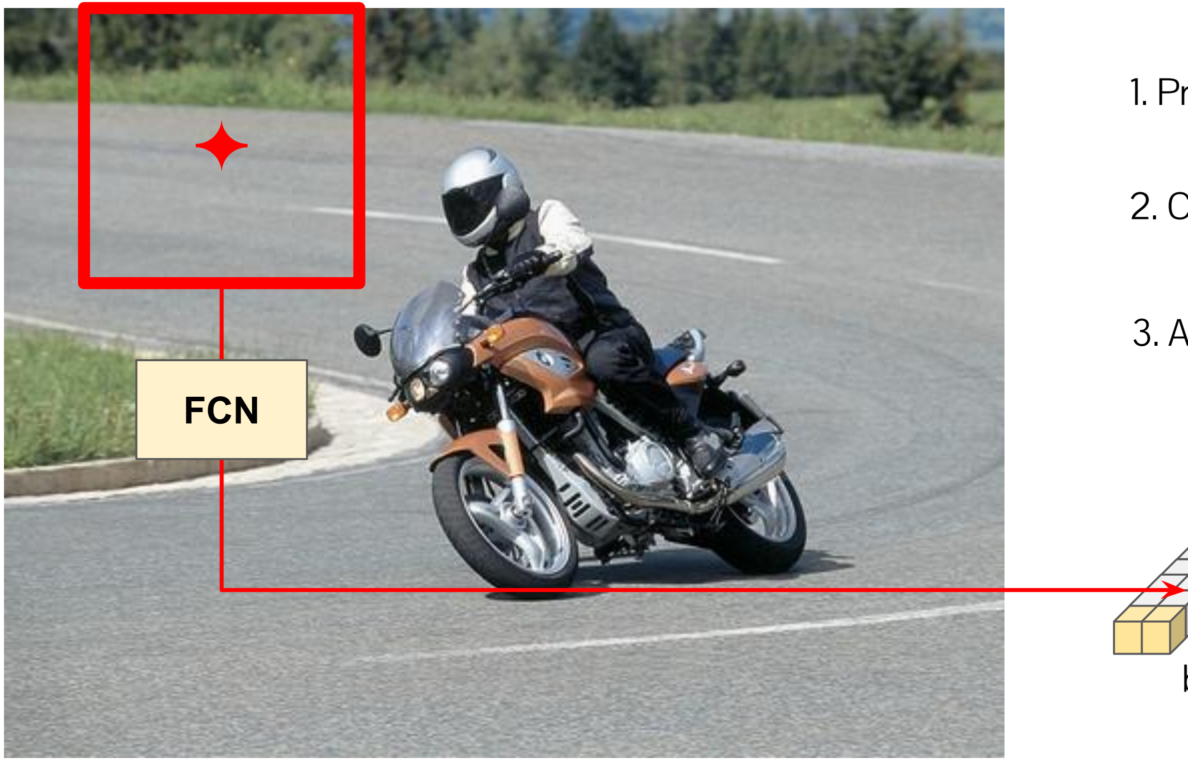
2. Convert **CNN** to **FCN**

3. Apply **FCN** to a larger image



background

# Fully Convolutional Network

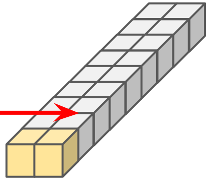



540x540

1. Pre-train a **CNN** for classification

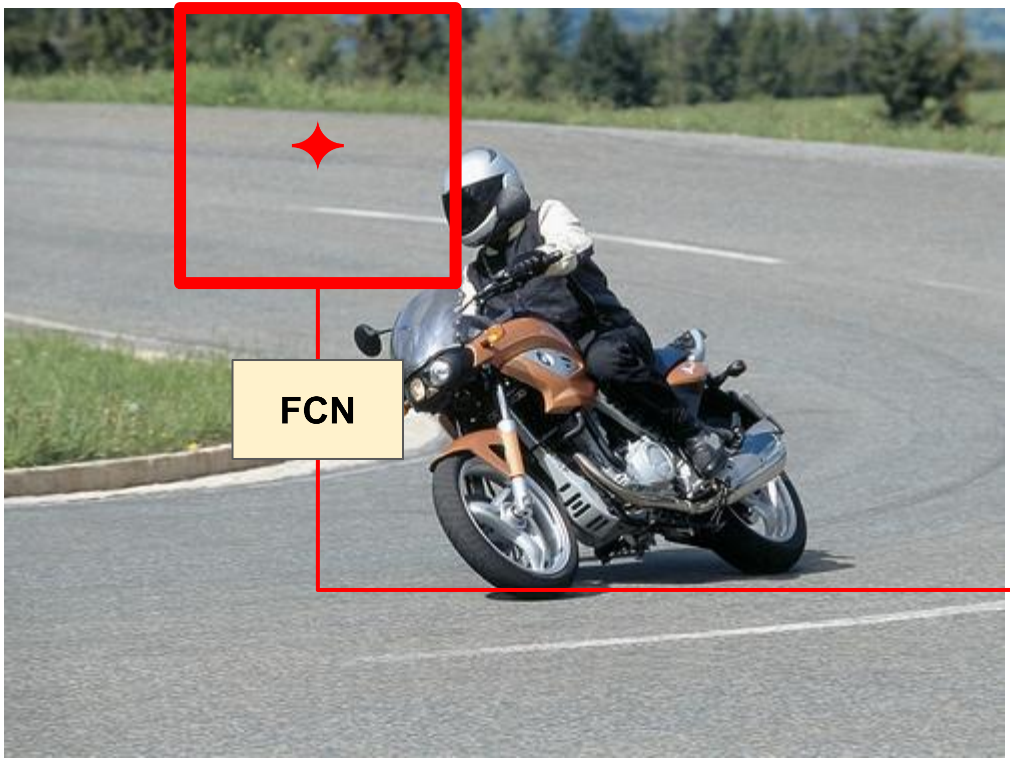
2. Convert **CNN** to **FCN**

3. Apply **FCN** to a larger image



background

# Fully Convolutional Network

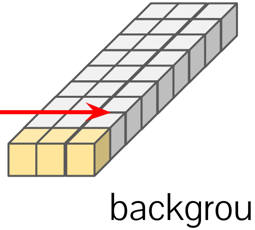



540x540

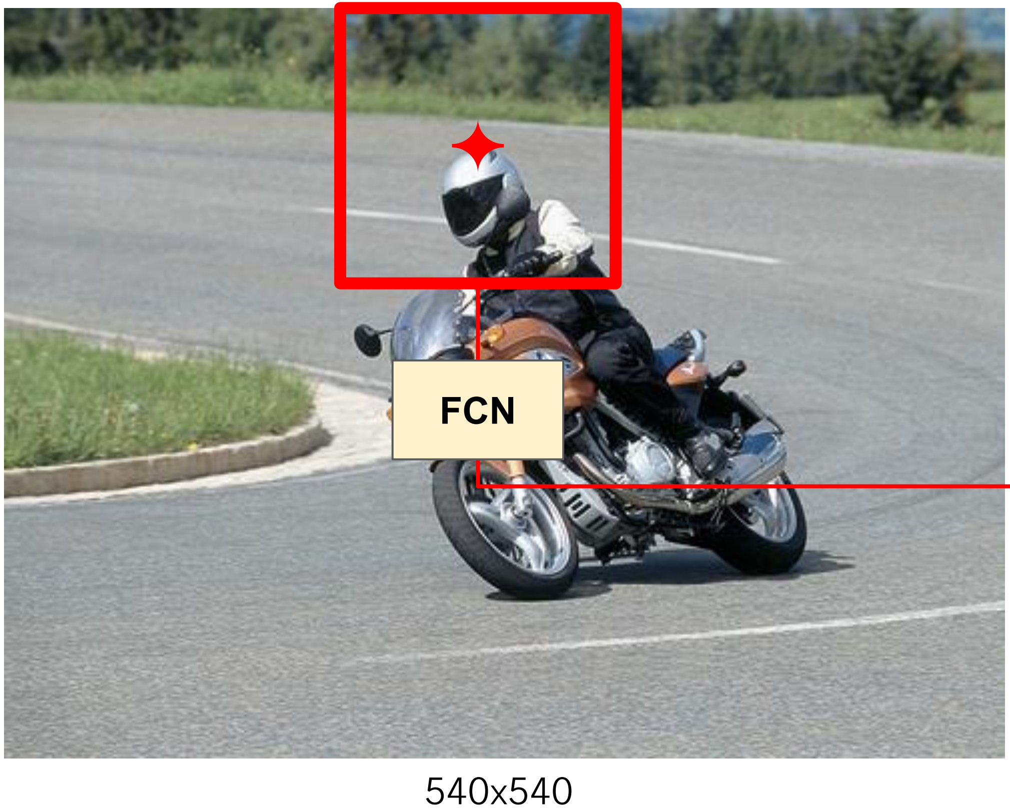
1. Pre-train a **CNN** for classification

2. Convert **CNN** to **FCN**

3. Apply **FCN** to a larger image



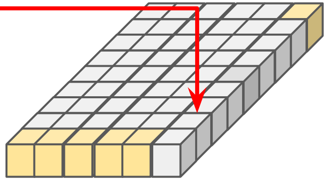
# Fully Convolutional Network



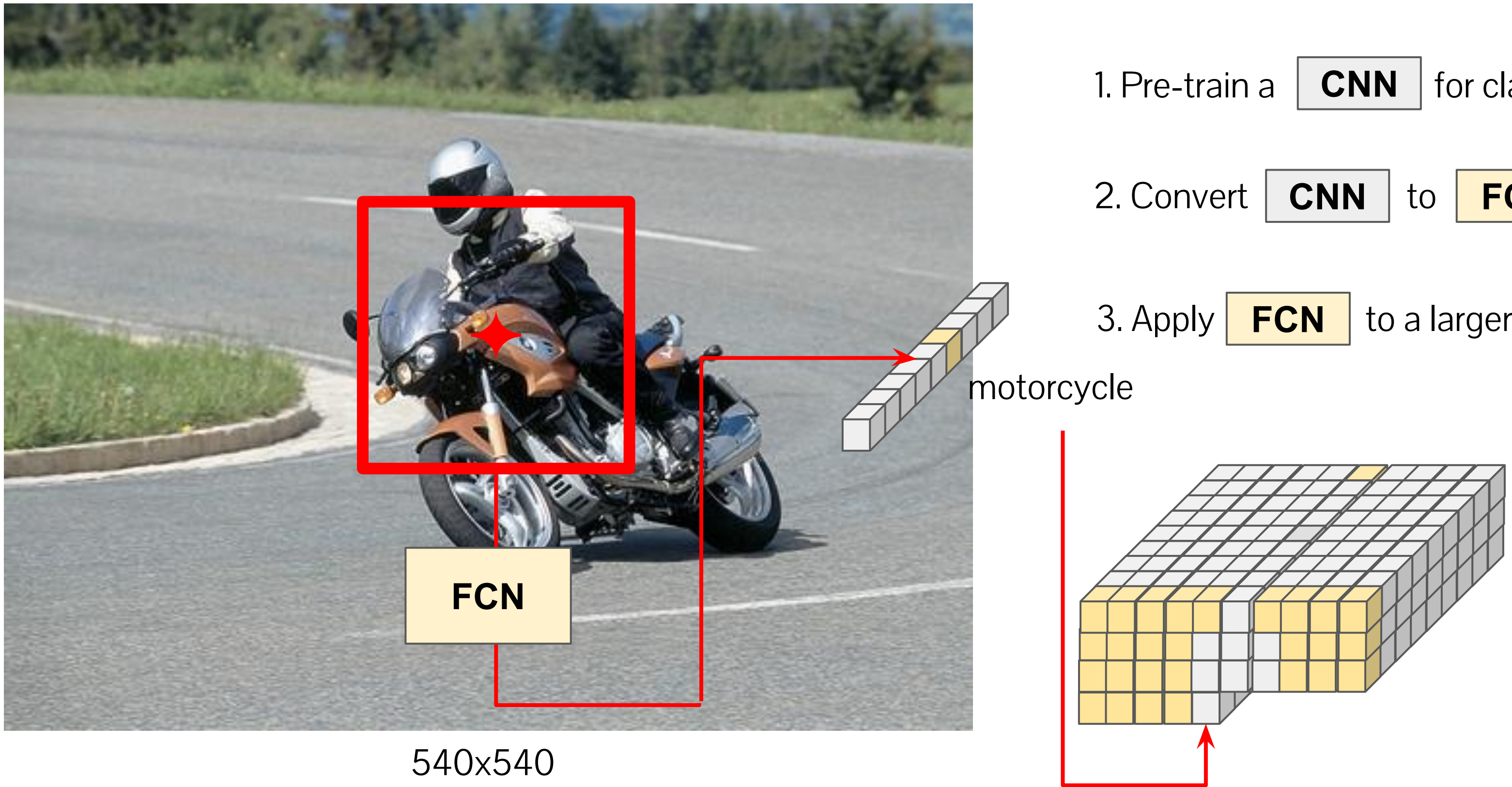

1. Pre-train a **CNN** for classification

2. Convert **CNN** to **FCN**

3. Apply **FCN** to a larger image



# Fully Convolutional Network

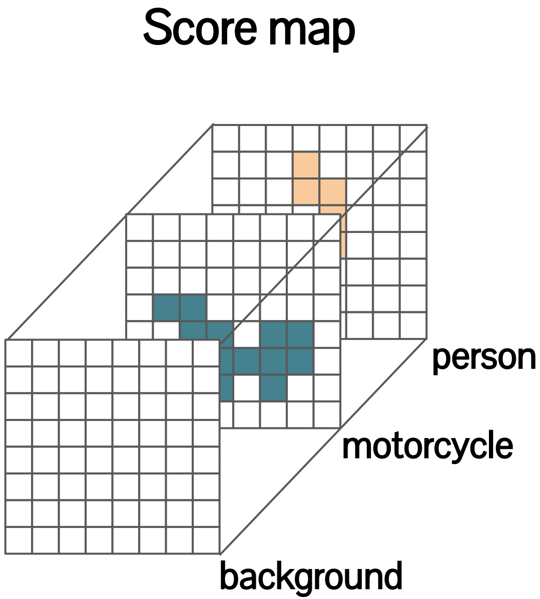



1. Pre-train a **CNN** for classification

2. Convert **CNN** to **FCN**

3. Apply **FCN** to a larger image

# Fully Convolutional Network

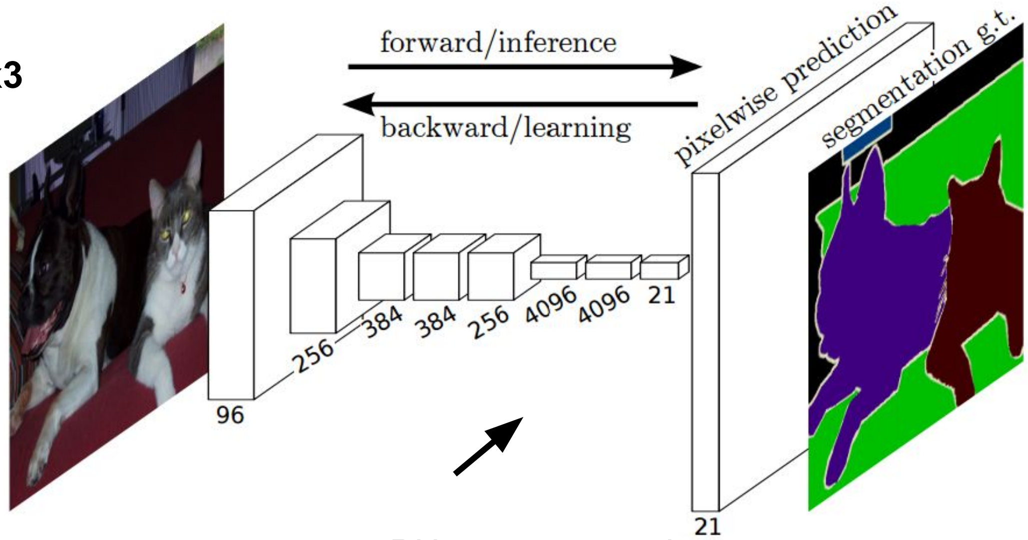



1. Pre-train a **CNN** for classification
2. Convert **CNN** to **FCN**
3. Apply **FCN** to a larger image
4. Get the final label map by taking per-pixel argmax over classes

# Fully Convolutional Network (FCN)

- End-to-end CNN architecture for semantic segmentation
- Interpretation of fully-connected layers to convolutional layers

500x500x3

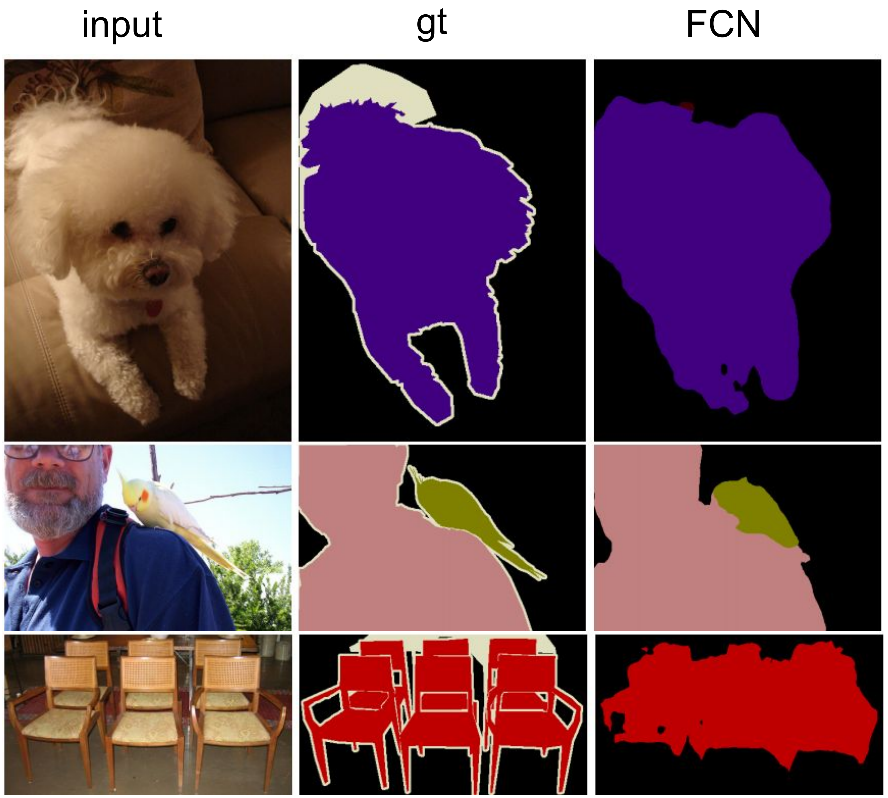



Bilinear upsampling



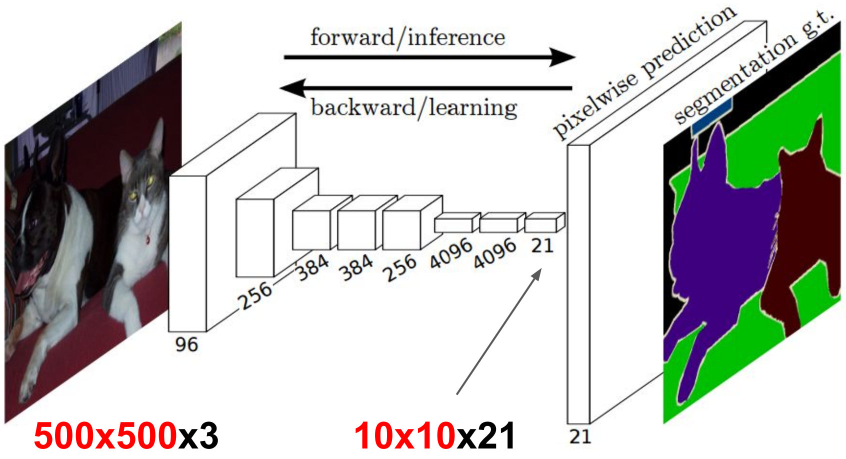
Long et al, “Fully convolutional networks for semantic segmentation,” in CVPR, 2015

# Fully Convolutional Network (FCN)

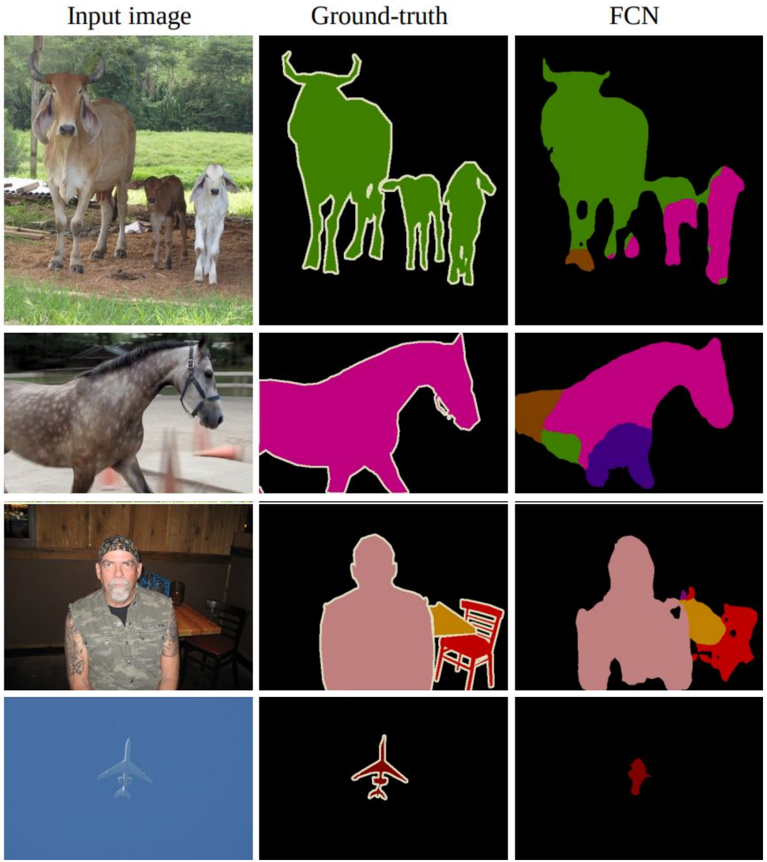


# Limitations in FCN

- Low resolution score map
  - 500x500 input image → 10x10 score map
  - May lost many detailed shape
- Fixed receptive field
  - Cannot handle objects in various size



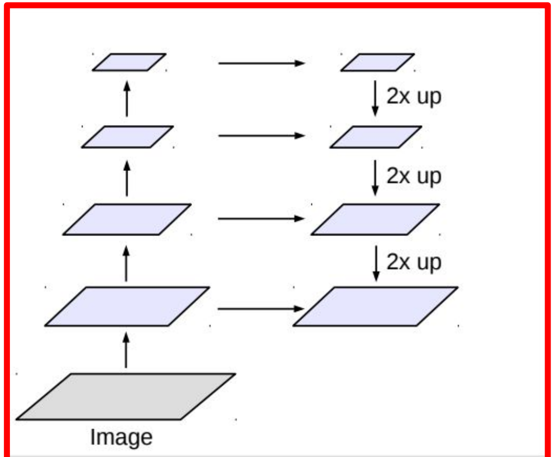
# Limitations in FCN



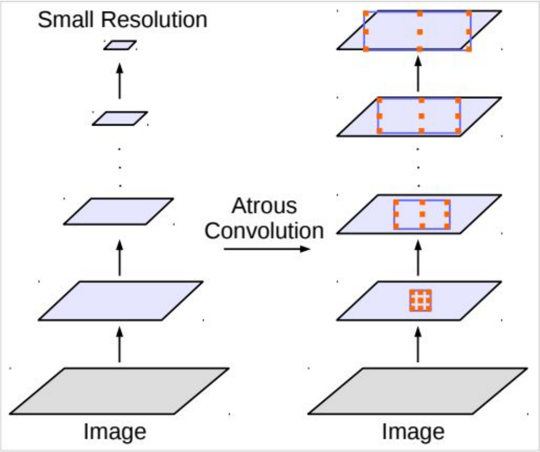

Misprediction due to the fixed receptive field size

Lost in detailed shape

# How to improve semantic segmentation




Encoder-decoder network

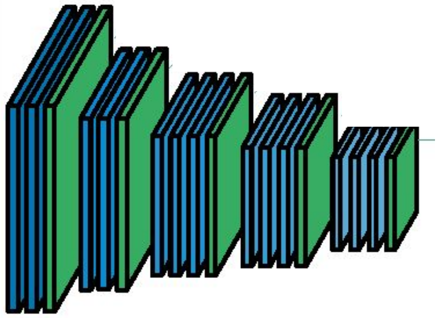


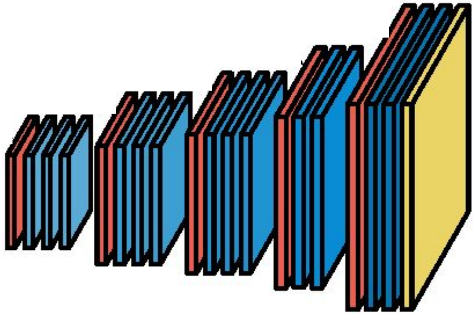

Atrous convolution

Figure credit: Liang-Chieh Chen

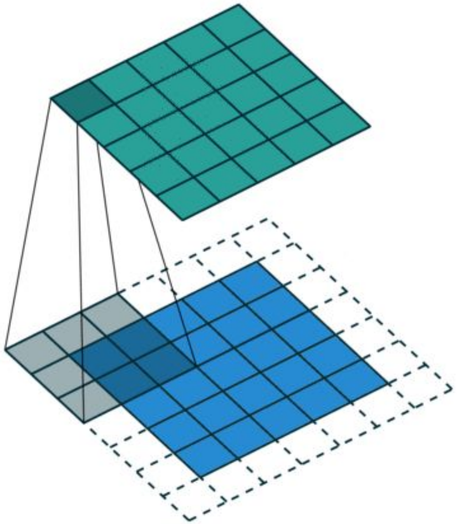
# Encoder & decoder networks

- Encoder:
  - **Compress** the information in the original data (e.g. CNN for image classification)
  - Abstract the original information in data and extract higher-order information
- Decoder:
  - **Reconstruct** the information from the representation (e.g. DCNN)
  - Extract original information in data lost during the encoding process

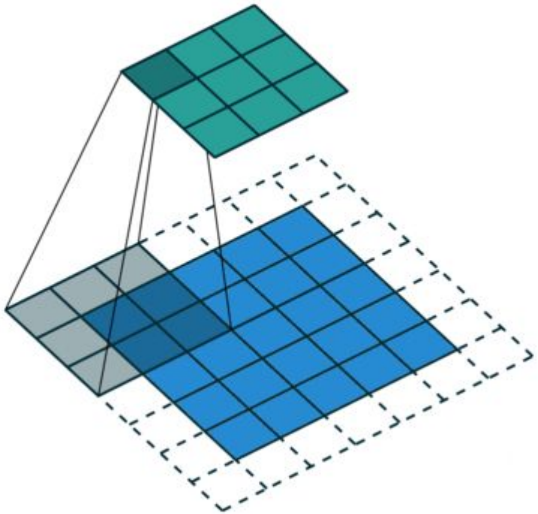




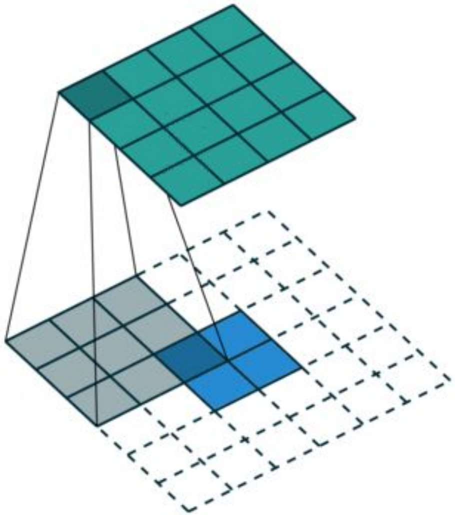
# Deconvolution for upsampling



**Convolution**



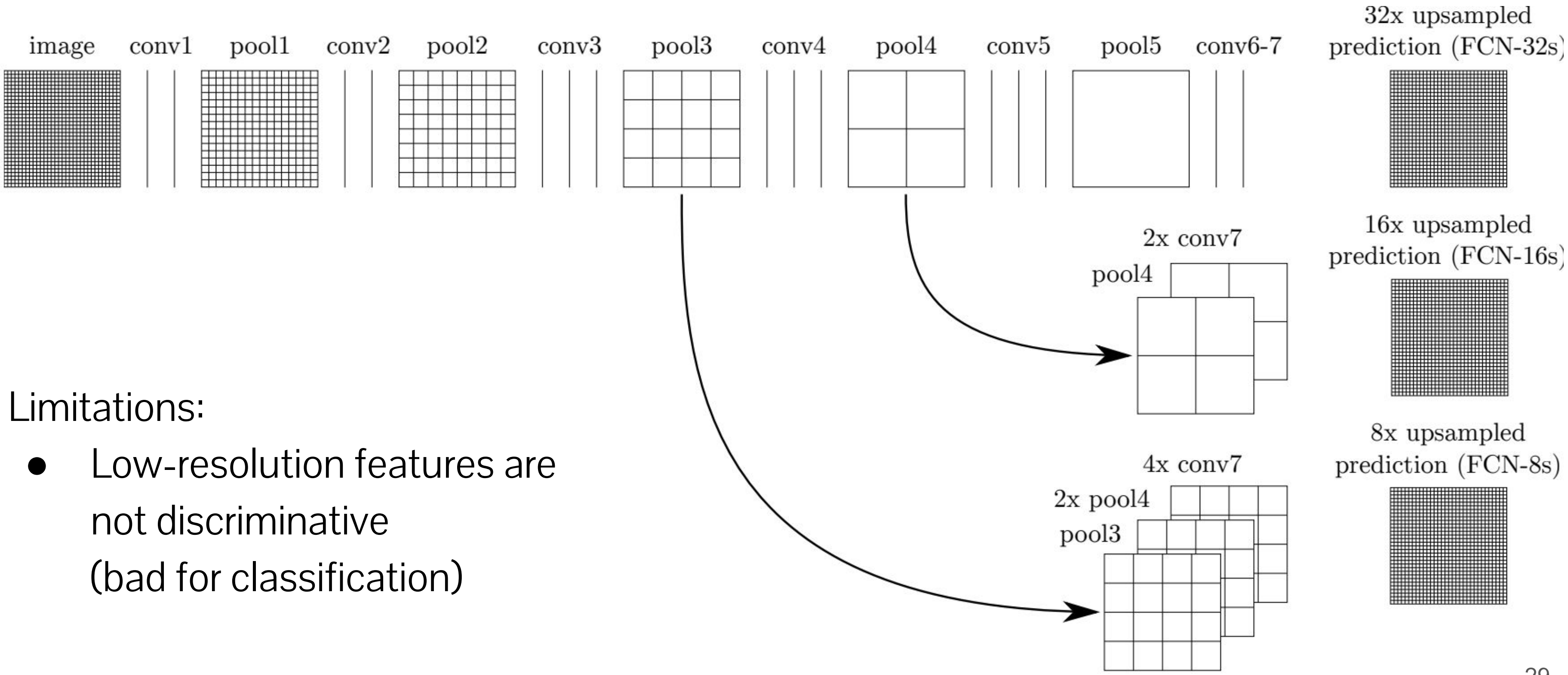
**Convolution with stride**
[downsampling]



**Deconvolution**
(transposed convolution)
[upsampling]

Figure source: https://github.com/vdumoulin/conv_arithmeti

# Skip connection for capturing detailed shapes




Limitations:

- Low-resolution features are not discriminative (bad for classification)



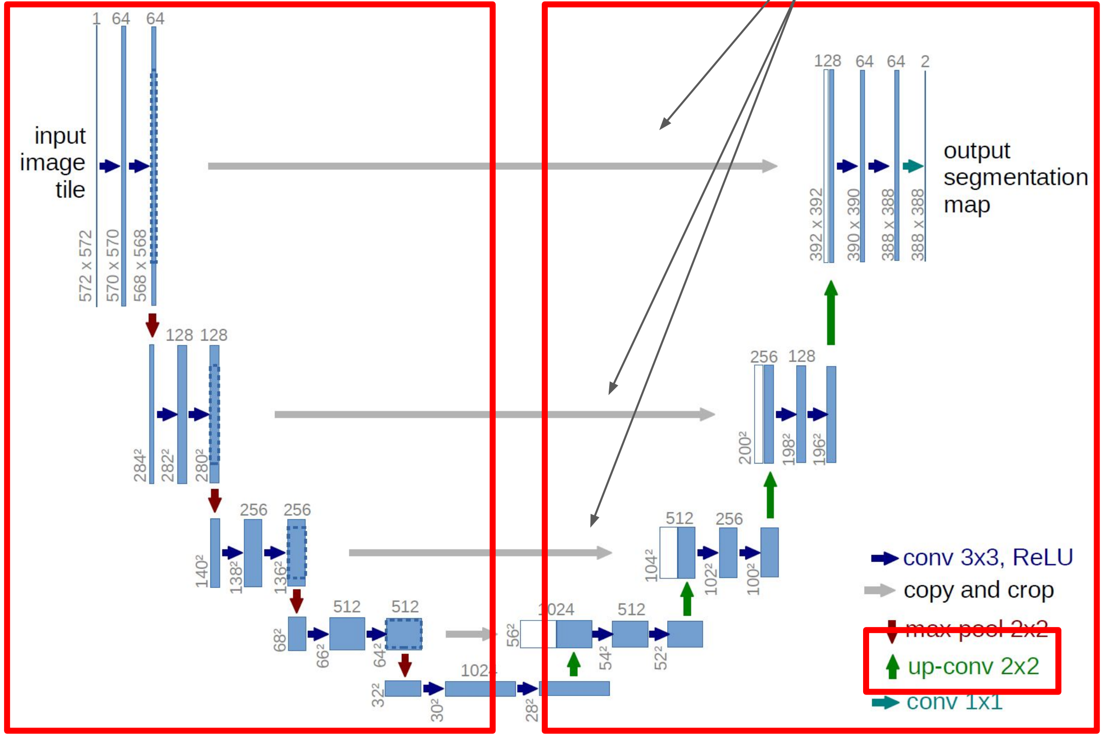
Long et al, "Fully convolutional networks for semantic segmentation," in CVPR, 2015

# Unet

**Skip connection**:
Deliver fine details of input image to higher layers by concatenating features

- Encoder-decoder network with **skip-connection**

**Encoding**:
- Downsampling to lower resolution
- Abstract from low pixels to higher semantics



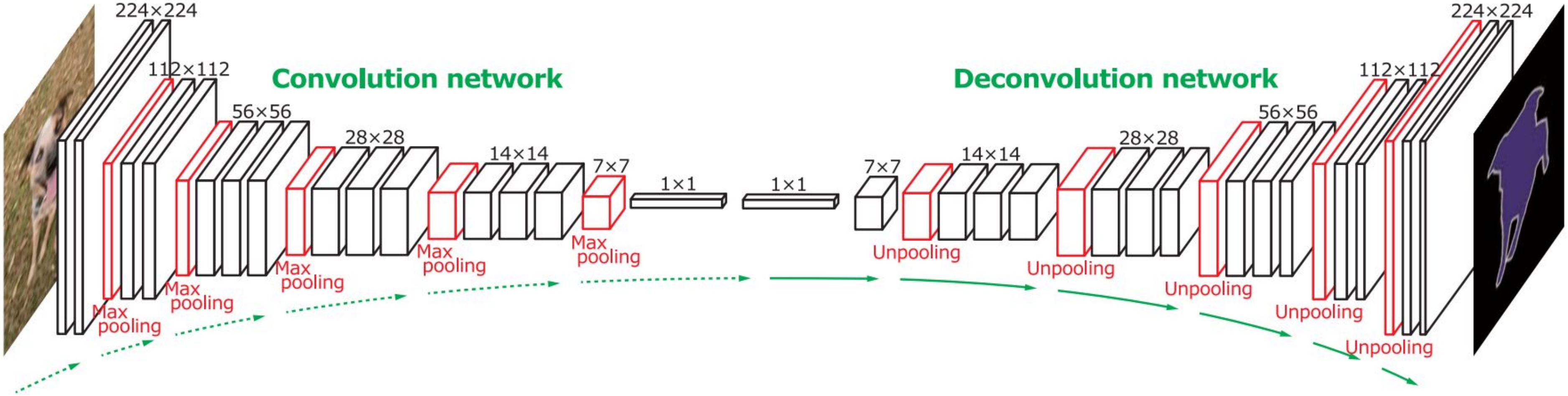

**Decoding**:
- Upsampling to higher resolution
- Reconstruct shape information

What is this upsampling?

Ronneberger et al., U-Net: Convolutional Networks for Biomedical Image Segmentation, In MICCAI, 2015

# Deconvolution network

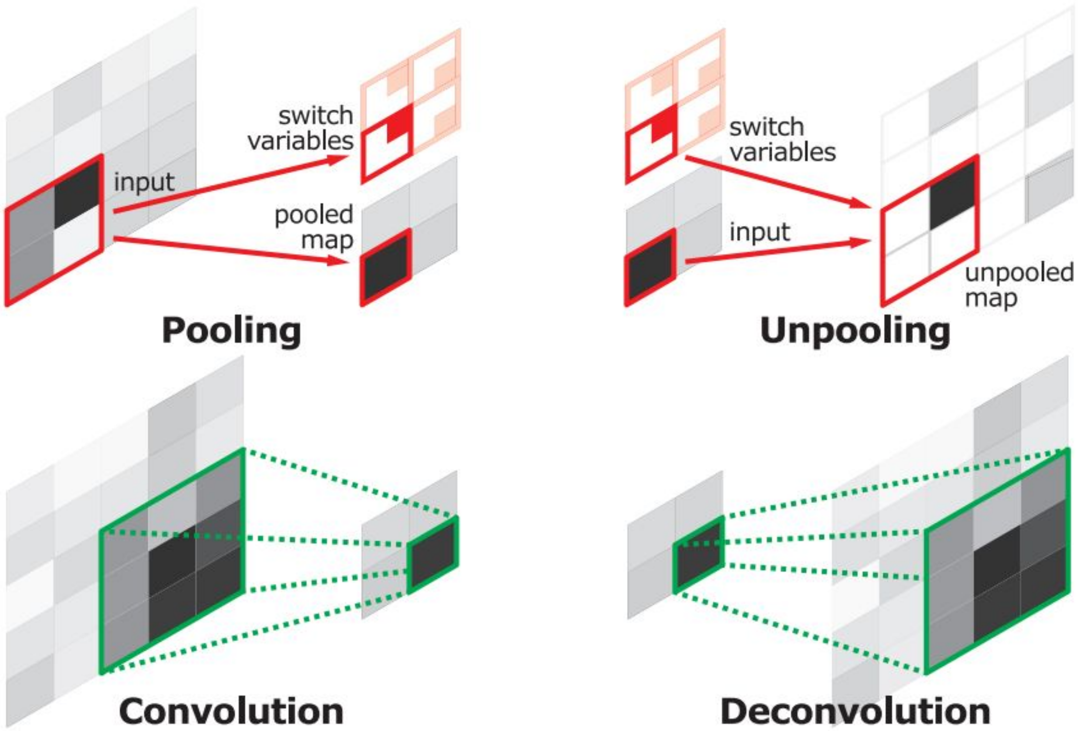
- Encoder-decoder network with **shared pooling switches**






Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Operations in deconvolution network

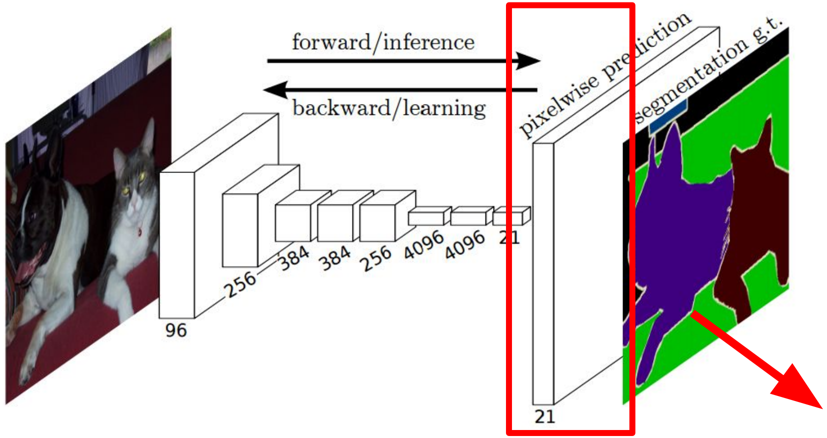



- Unpooling
  Increase the resolution
  using pooling switches

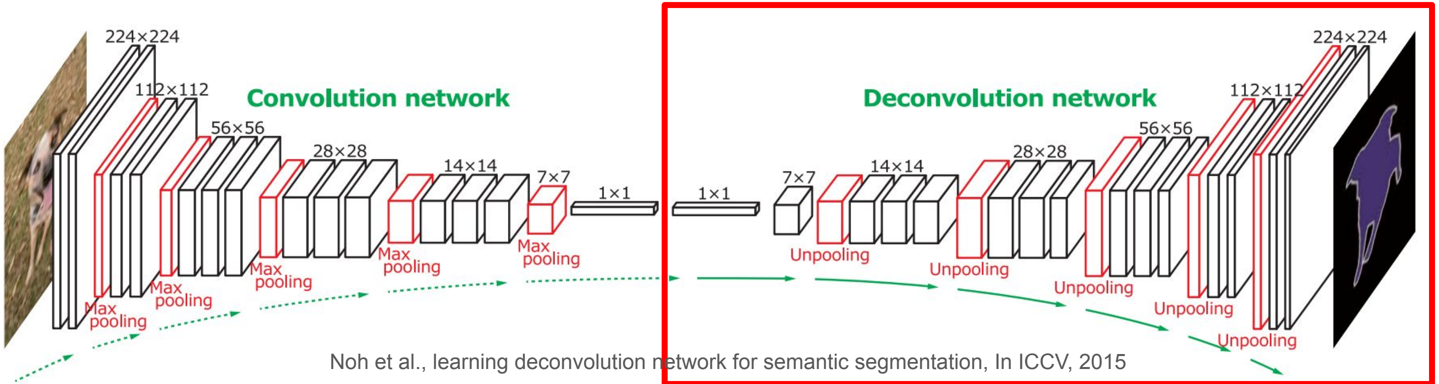
- Deconvolution
  Reconstruct the shapes
  missing in unpooling

Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Comparisons to FCN

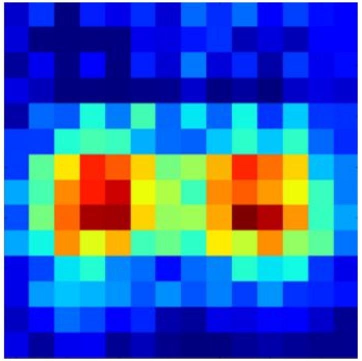



Replacing the upsampling by learnable parameters
→ trainable upsampling!




Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Visualization of deconvolution network

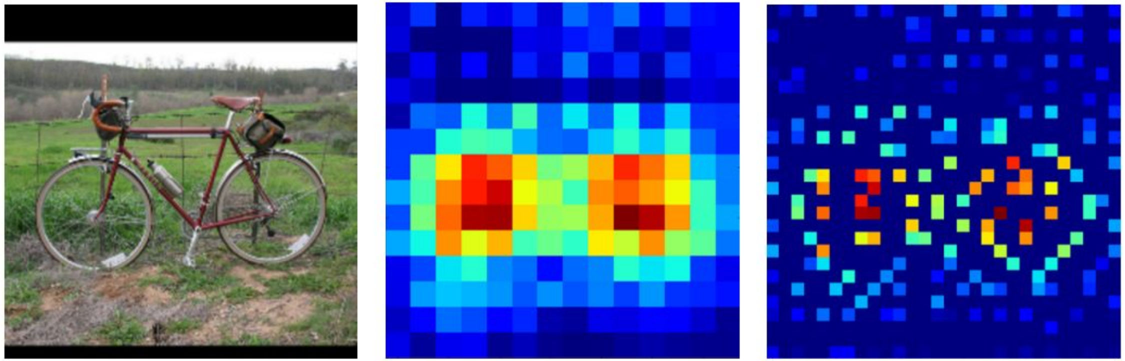


Coarse activation map obtained from
the output of the encoder network



Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Visualization of deconvolution network



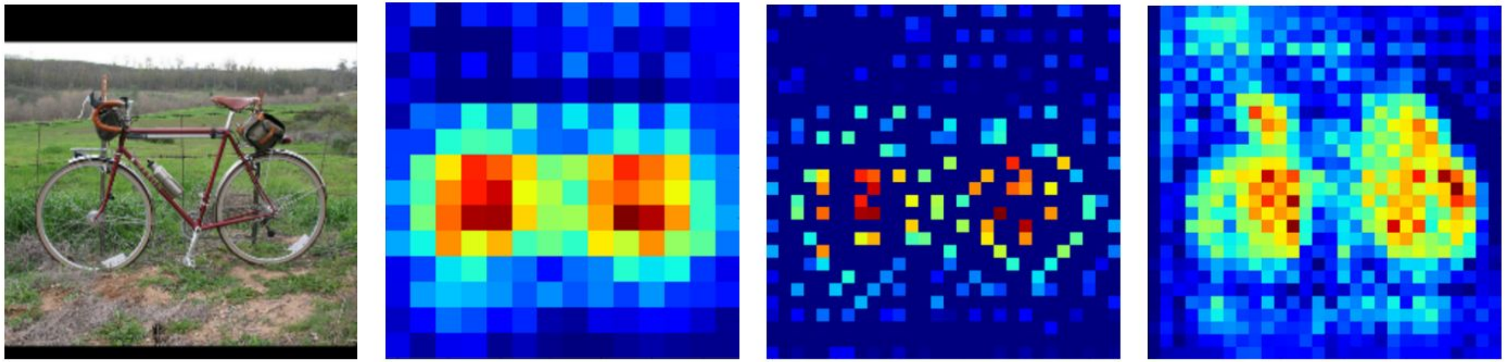
**Unpooling:**

- Double the resolution
- sparse activation with reconstructed shape information



Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Visualization of deconvolution network



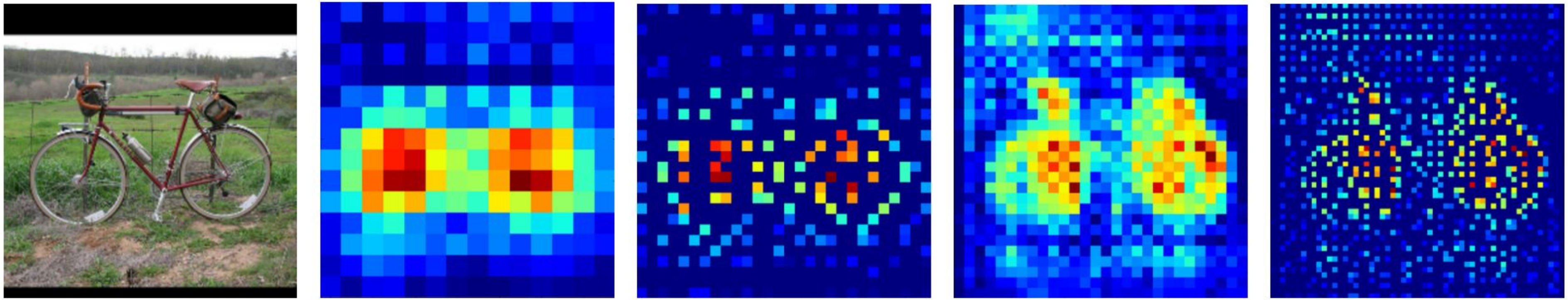
**Deconvolution:**

- Densify the activation from the sparse unpooled feature map
- Reconstruct more detailed shapes



Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Visualization of deconvolution network

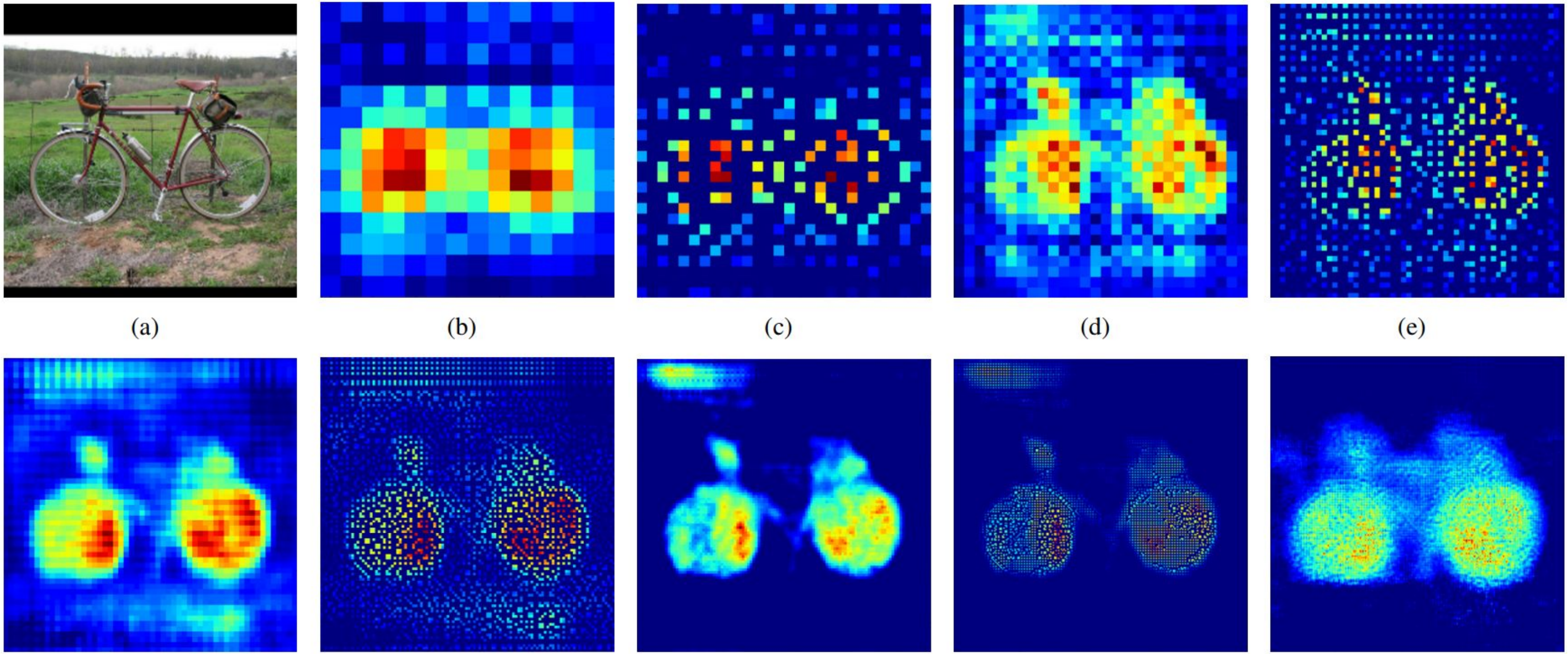


2nd Unpooling



Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Visualization of deconvolution network

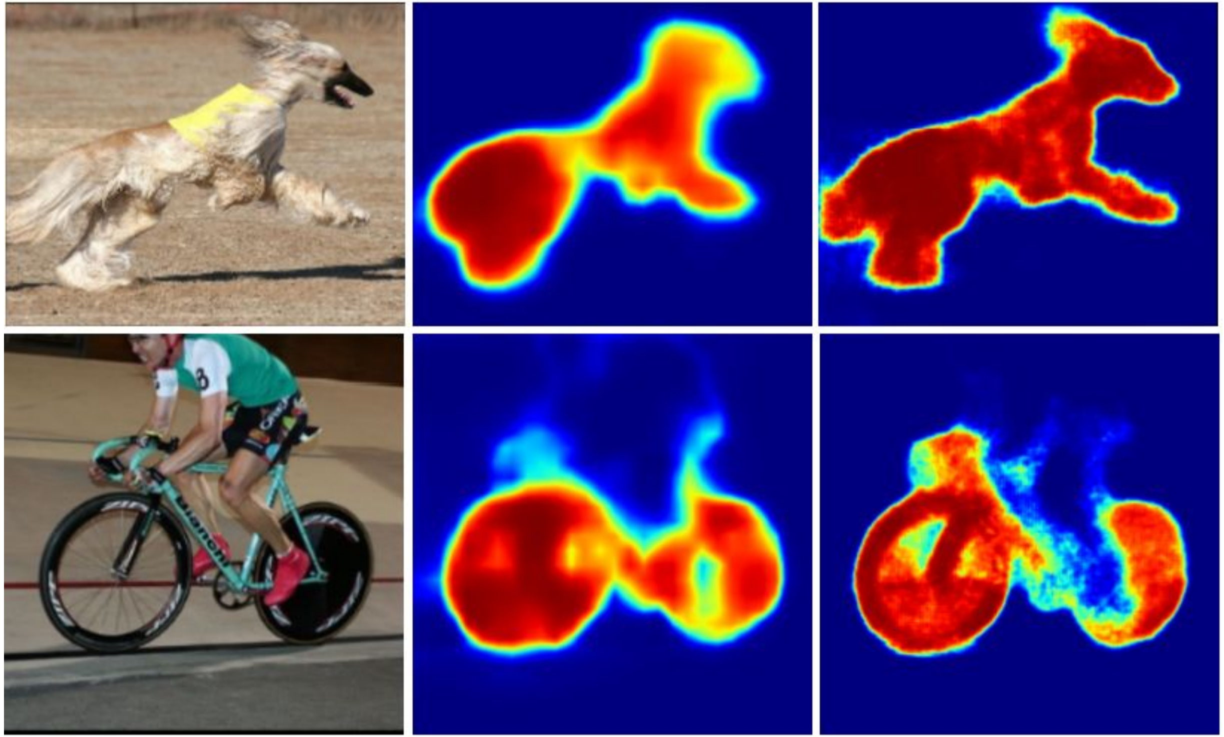


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

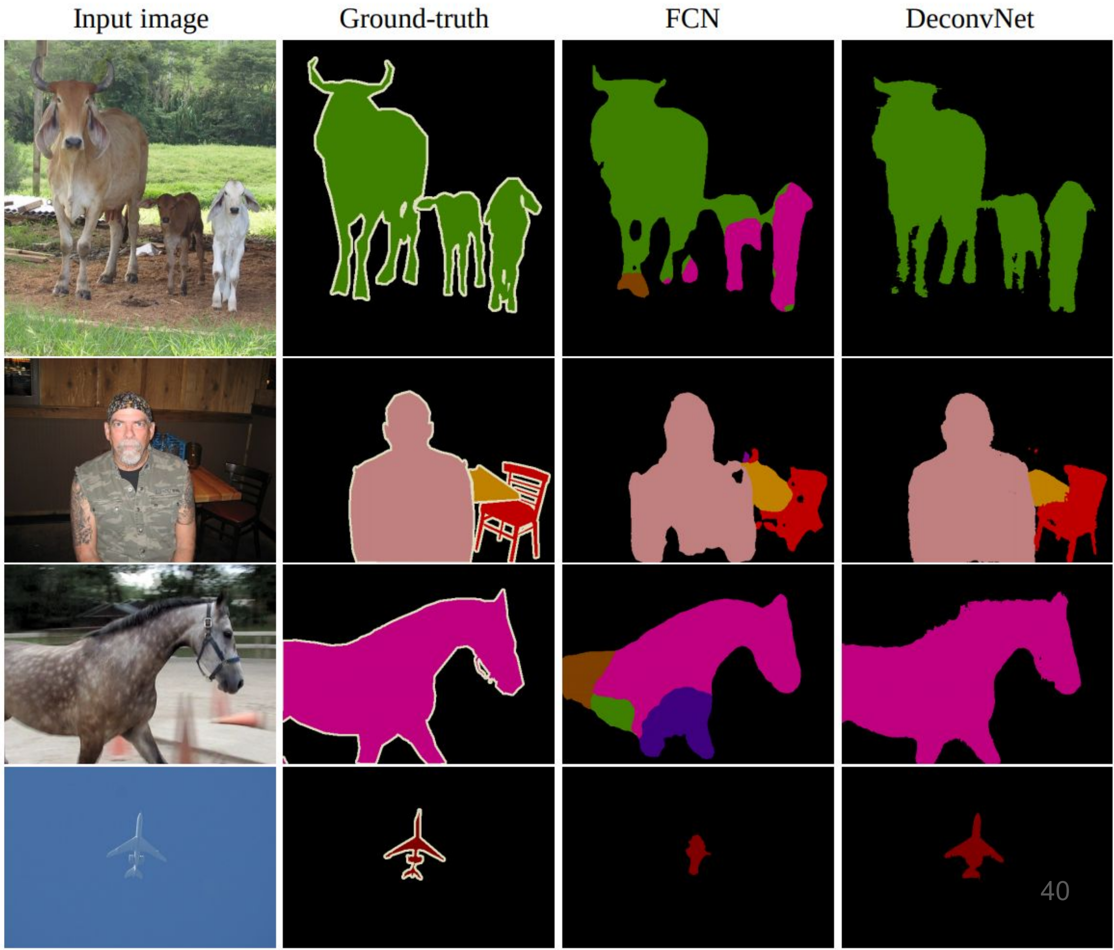
# Comparisons to FCN



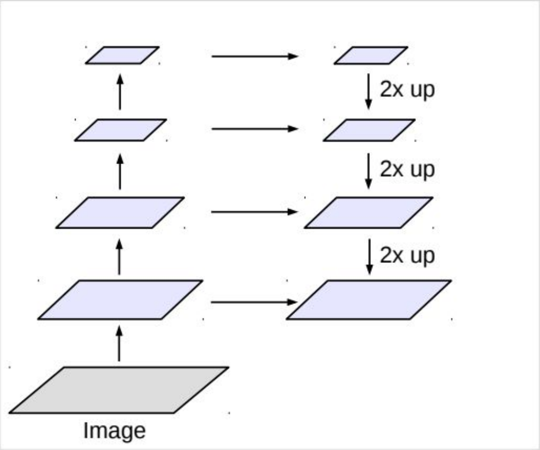
(a) Input image (b) FCN-8s (c) Ours

Noh et al., learning deconvolution network for semantic segmentation, In ICCV, 2015

# Comparisons to FCN



# Summary: encoder-decoder network



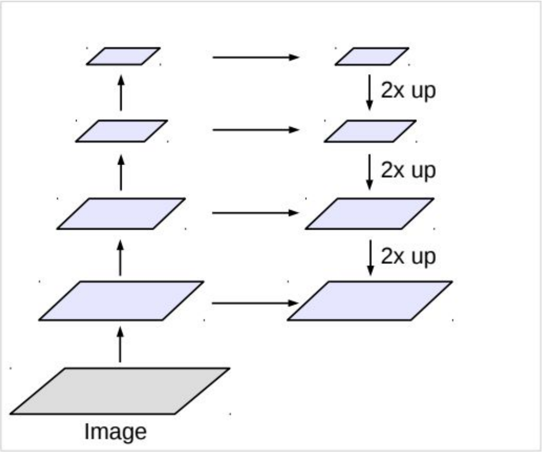

Encoder-decoder network

- Reconstruct spatial information lost in encoding network

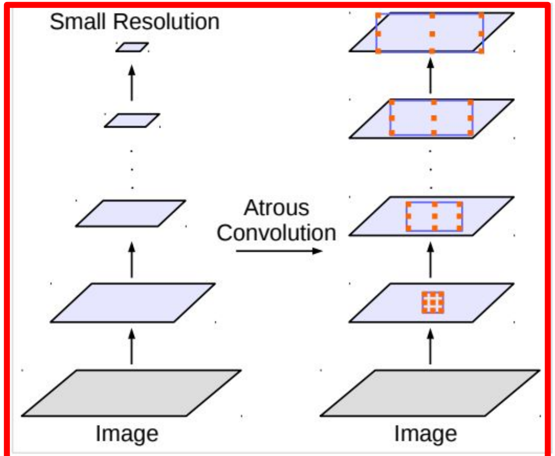
- Three approaches
  - Skip connection
  - Deconvolution for learnable upsampling
  - Using pooling switch to reconstruct spatial information

- Encoder-decoder is a popular architecture outside the segmentation domain too!
(also appears in following lectures)

# How to improve semantic segmentation




Encoder-decoder network

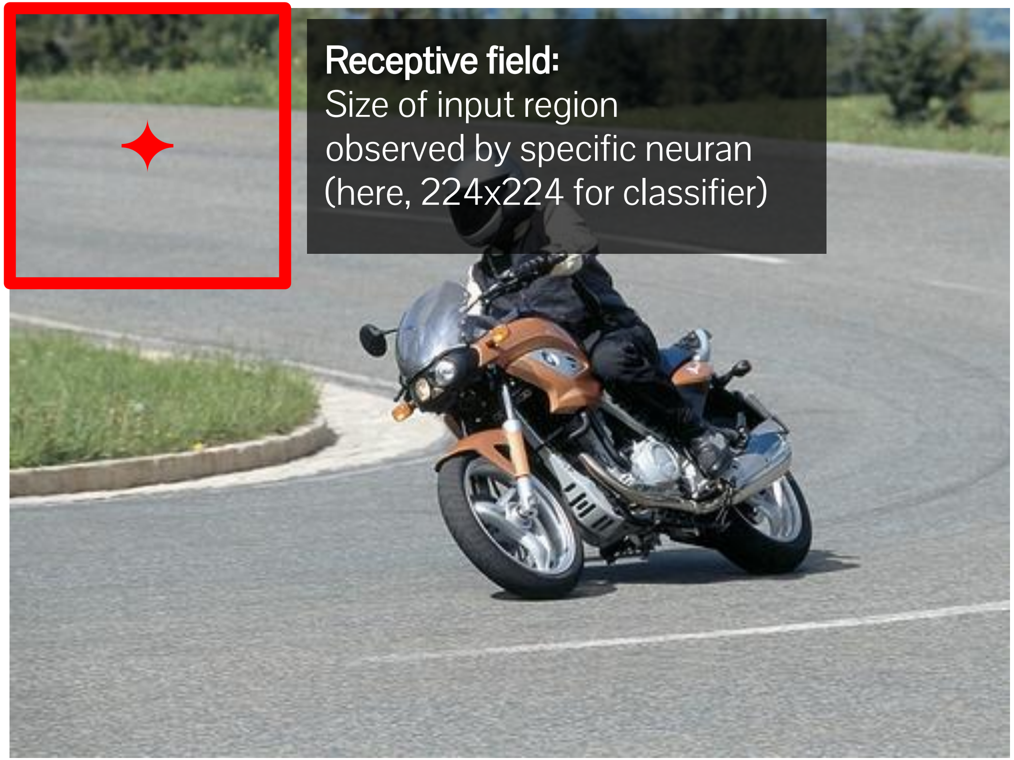



Atrous convolution

Figure credit: Liang-Chieh Chen

# Field-of-View (FoV) in segmentation

- Receptive field / FoV

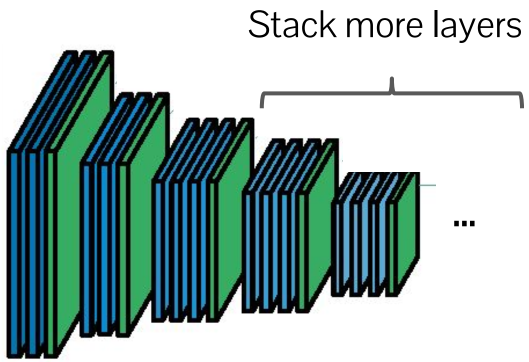


# Field-of-View (FoV) in segmentation

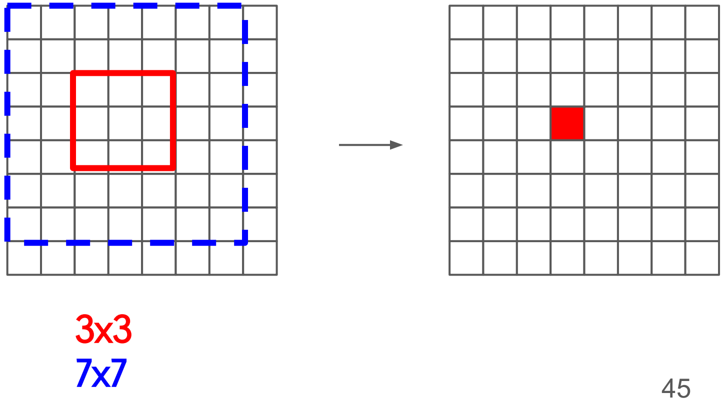
- Too small FoV
  - Increase ambiguity of classification due to local observations
  - Cannot consider the rich context around the objects

# Increasing FoV

- Increase subsampling ratio
  - Lose spatial information




- Increase the convolutional filter size
  - Increase the parameters of the model
  - Increase the computational cost / prone to overfitting



# Atrous convolution

- Convolution with **holes**
- Increase the FoV using the same parameters

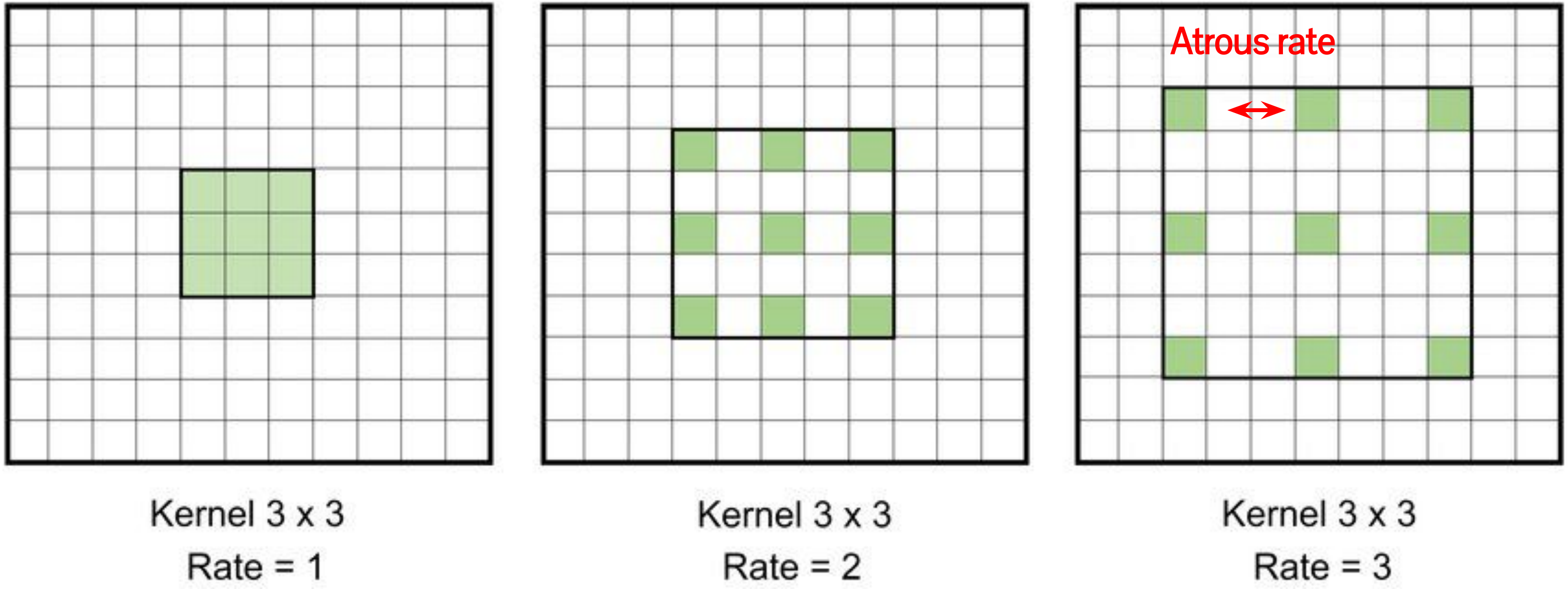



Figure source: Morales et al., Automatic Segmentation of Mauritia flexuosa in Unmanned Aerial Vehicle (UAV) Imagery Using Deep Learning

# Atrous convolution

- Convolution with **holes**
- Increase the FoV using the same parameters

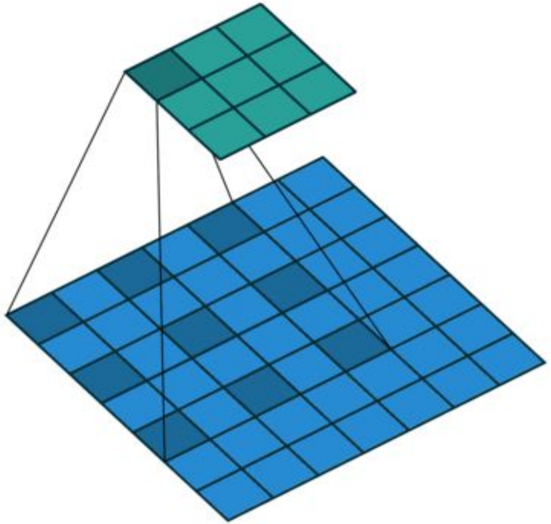
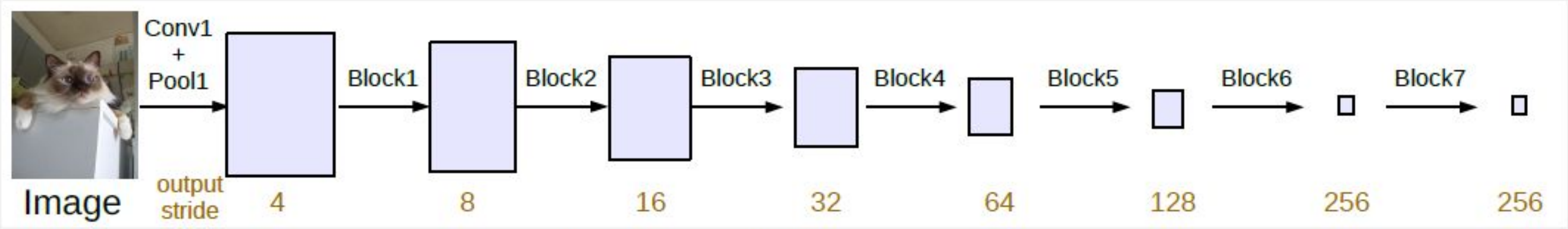
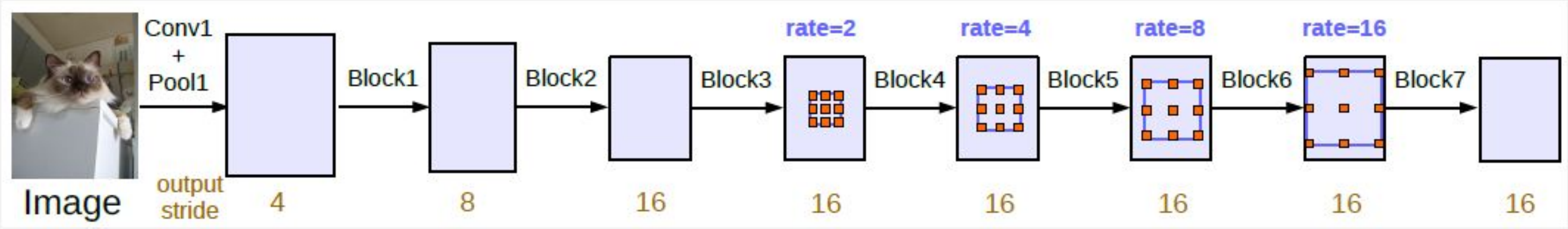


Figure source: https://github.com/vdumoulin/conv_arithmeti

# DeepLab: FCN with atrous convolution




(a) Going deeper without atrous convolution.




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



Chen et al., Rethinking atrous convolution for semantic image segmentation

# DeepLab: FCN with atrous convolution

| MSC | COCO | Aug | LargeFOV | ASPP | CRF | mIOU |
|---|---|---|---|---|---|---|
| | | | | | | 68.72 |
| ✓ | | | | | | 71.27 |
| ✓ | ✓ | | | | | 73.28 |
| ✓ | ✓ | ✓ | | | | 74.87 |
| ✓ | ✓ | ✓ | ✓ | | | 75.54 |
| ✓ | ✓ | ✓ | | ✓ | | 76.35 |
| ✓ | ✓ | ✓ | | ✓ | ✓ | 77.69 |