Image-to-Image Translation and Semantic Segmentation

Danna Gurari
University of Colorado Boulder
Spring 2024

CC BY

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2024/

Review

- Last lecture:
  - Problem
  - Applications
  - PASCAL VOC detection challenge
  - Faster R-CNN
  - YOLO
  - Programming Tutorial

- Assignments (Canvas)
  - Lab assignment 2 due earlier today
  - Problem set 3 due in 1 week

- Questions?

Today’s Topics

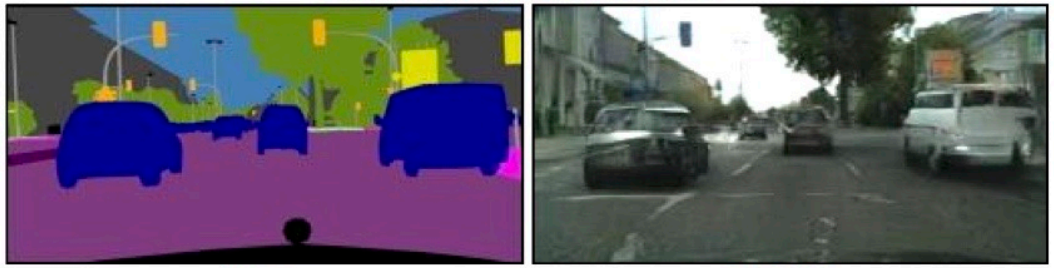
- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

Today’s Topics

- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

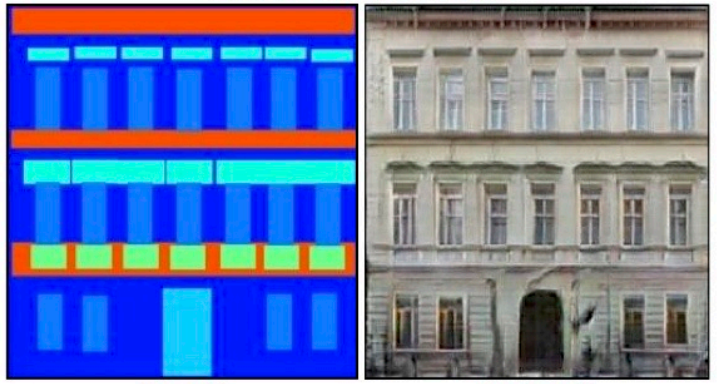
Image-to-Image Translation Examples

Labels to Street Scene



input output

Labels to Facade



input output

BW to Color

input output

Aerial to Map

input output

Day to Night

input output

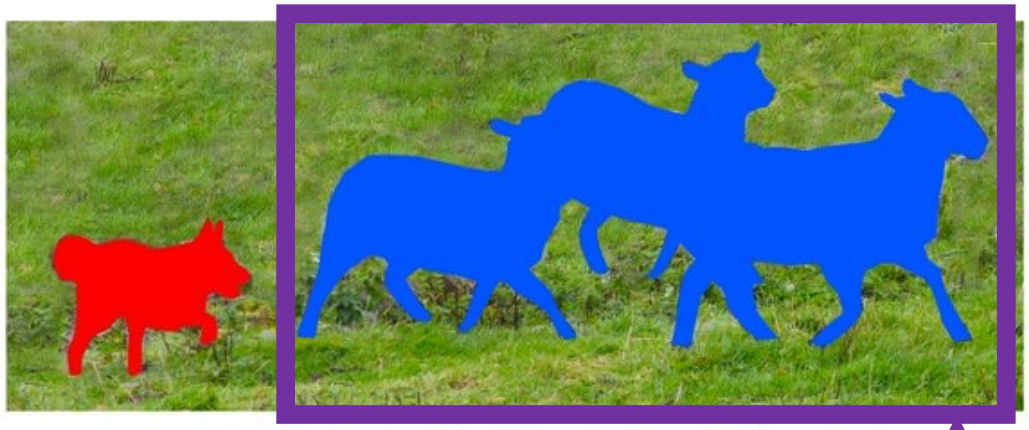
Edges to Photo

input output

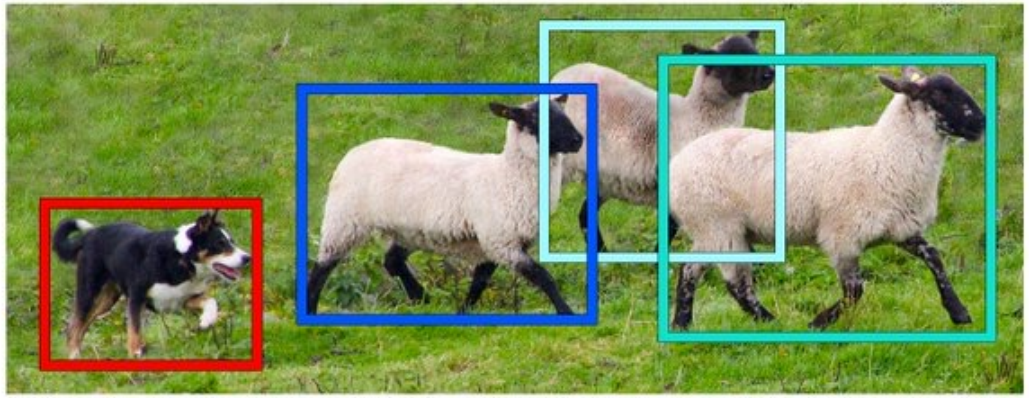
https://phillipi.github.io/pix2pix/

Today's Scope: Semantic Segmentation

Image Recognition



Semantic Segmentation



Object Detection

Locate all pixels that belong to pre-specified categories

Note: instances of the same category are NOT separated

https://ai-pool.com/d/could-you-explain-me-how-instance-segmentation-works

Today's Topics

- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

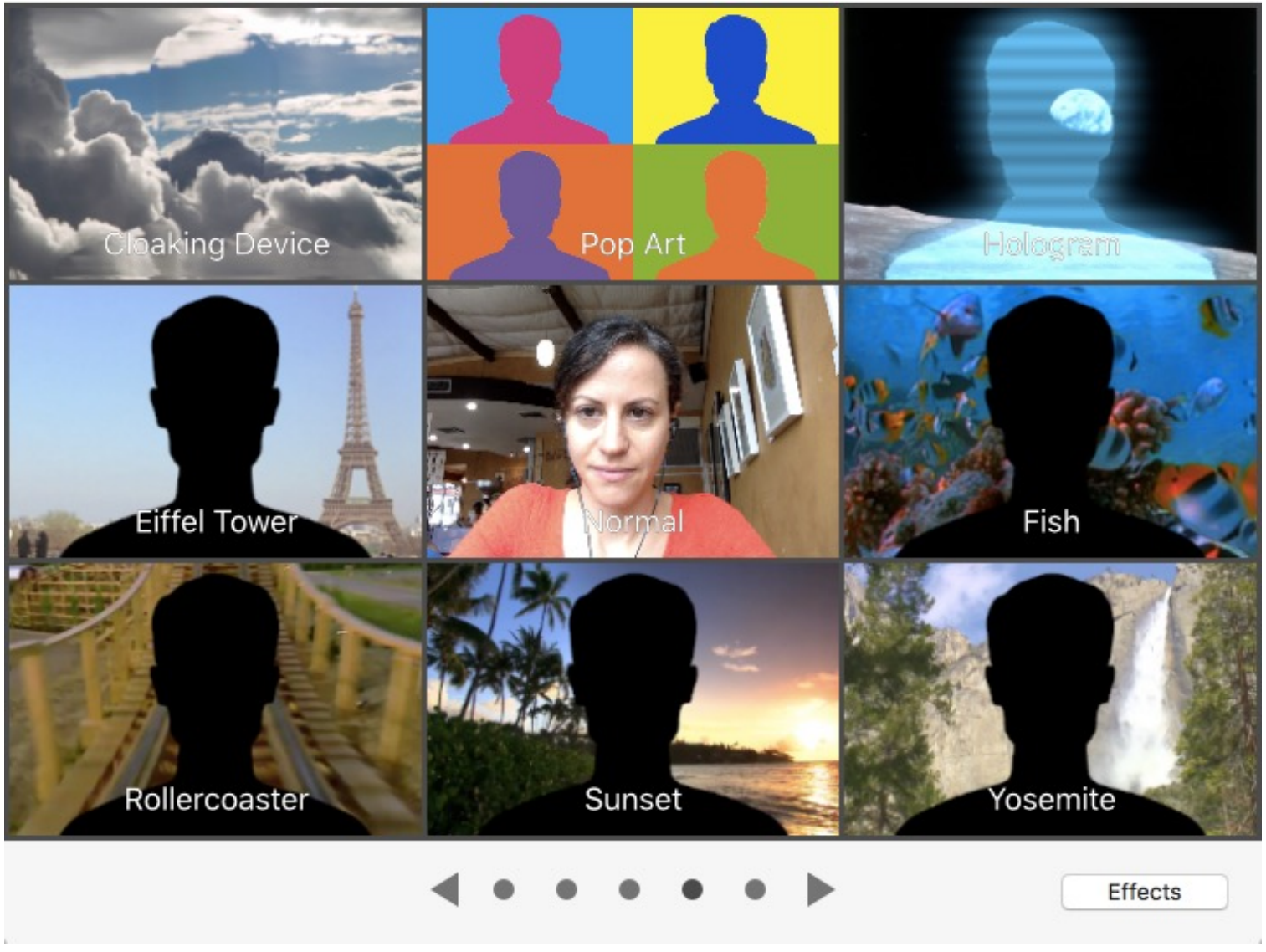
Remodeling Inspiration

(a) Target photo

(b) Retextured

Bell et al; SIGGRAPH; 2013

Rotoscoping (many examples on Wikipedia)



https://www.starnow.co.uk/ahmedmohammed1/photos/4650871/before-and-after-rotoscopinggreen-screening

Disease Diagnosis; e.g., PathAI

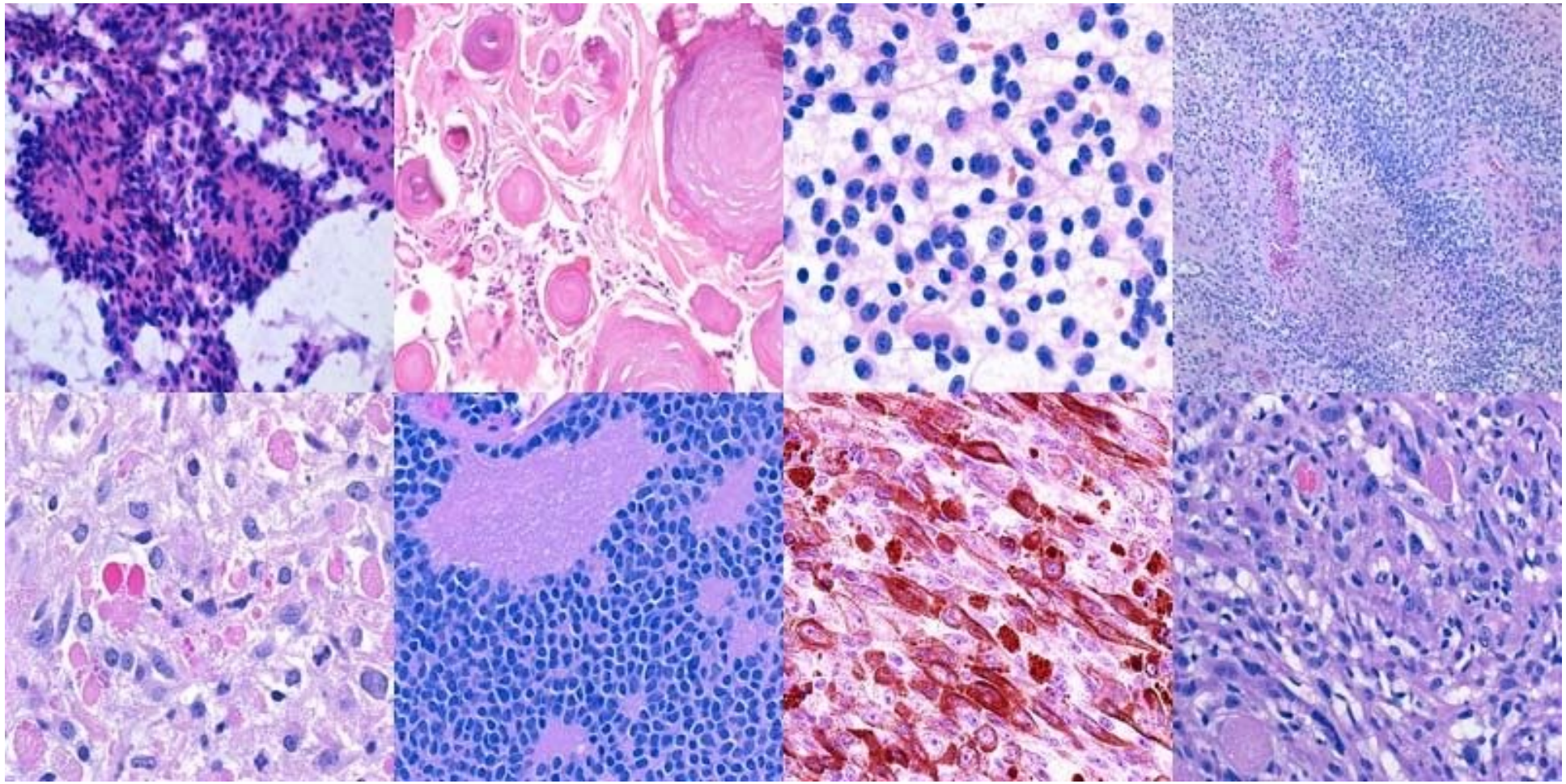
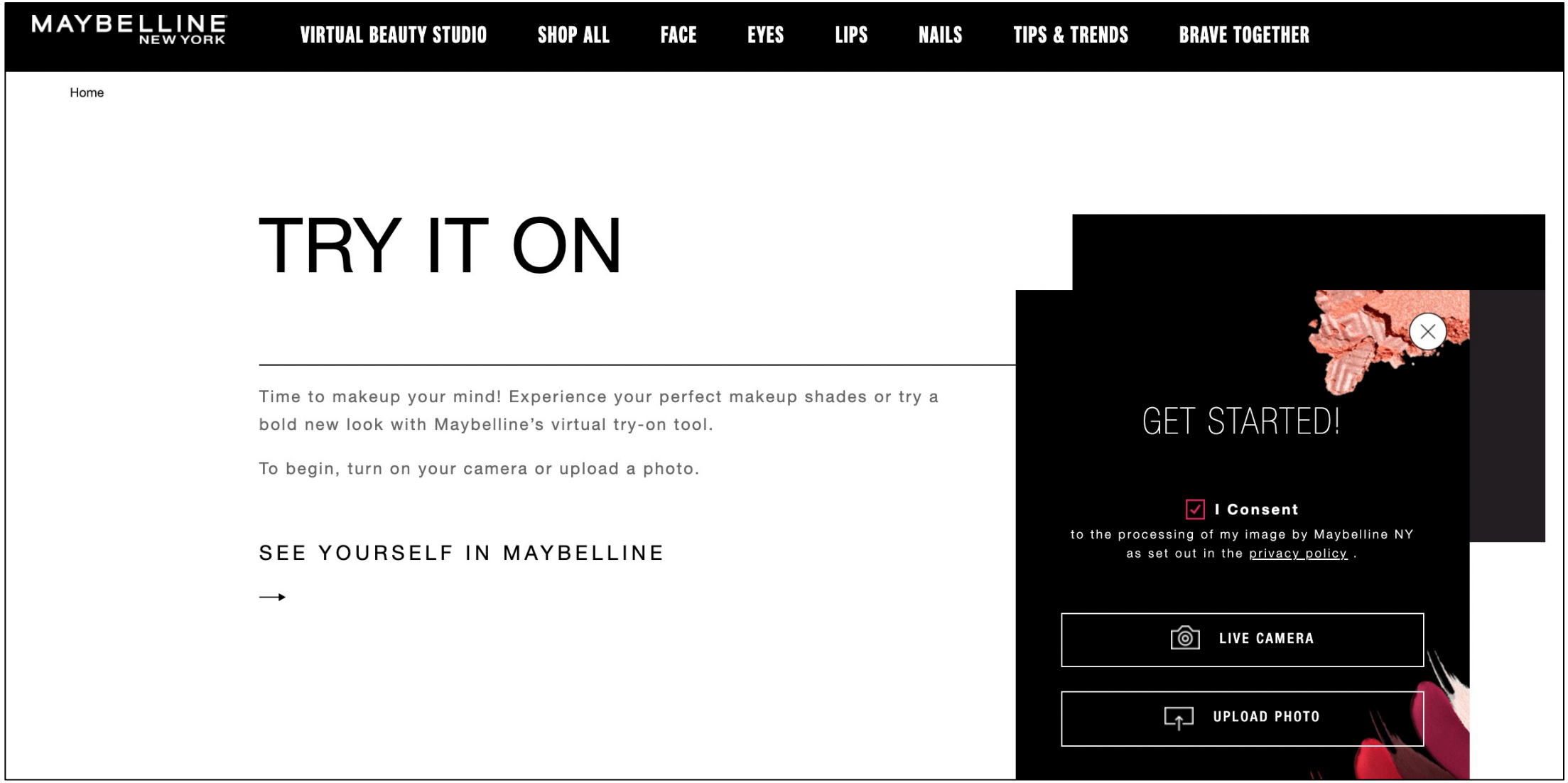


Figure Source: https://pathology.jhu.edu/brain-tumor/grading-classification

Face Makeover



Demo: https://www.maybelline.com/virtual-try-on-makeup-tools

Self-Driving Vehicles

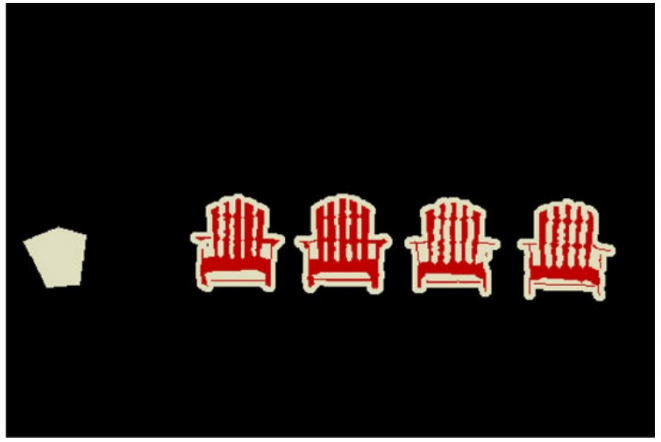
Figure Source: https://www.inc.com/kevin-j-ryan/self-driving-cars-powered-by-people-playing-games-mighty-ai.html

Today's Topics

- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

VOC Challenge

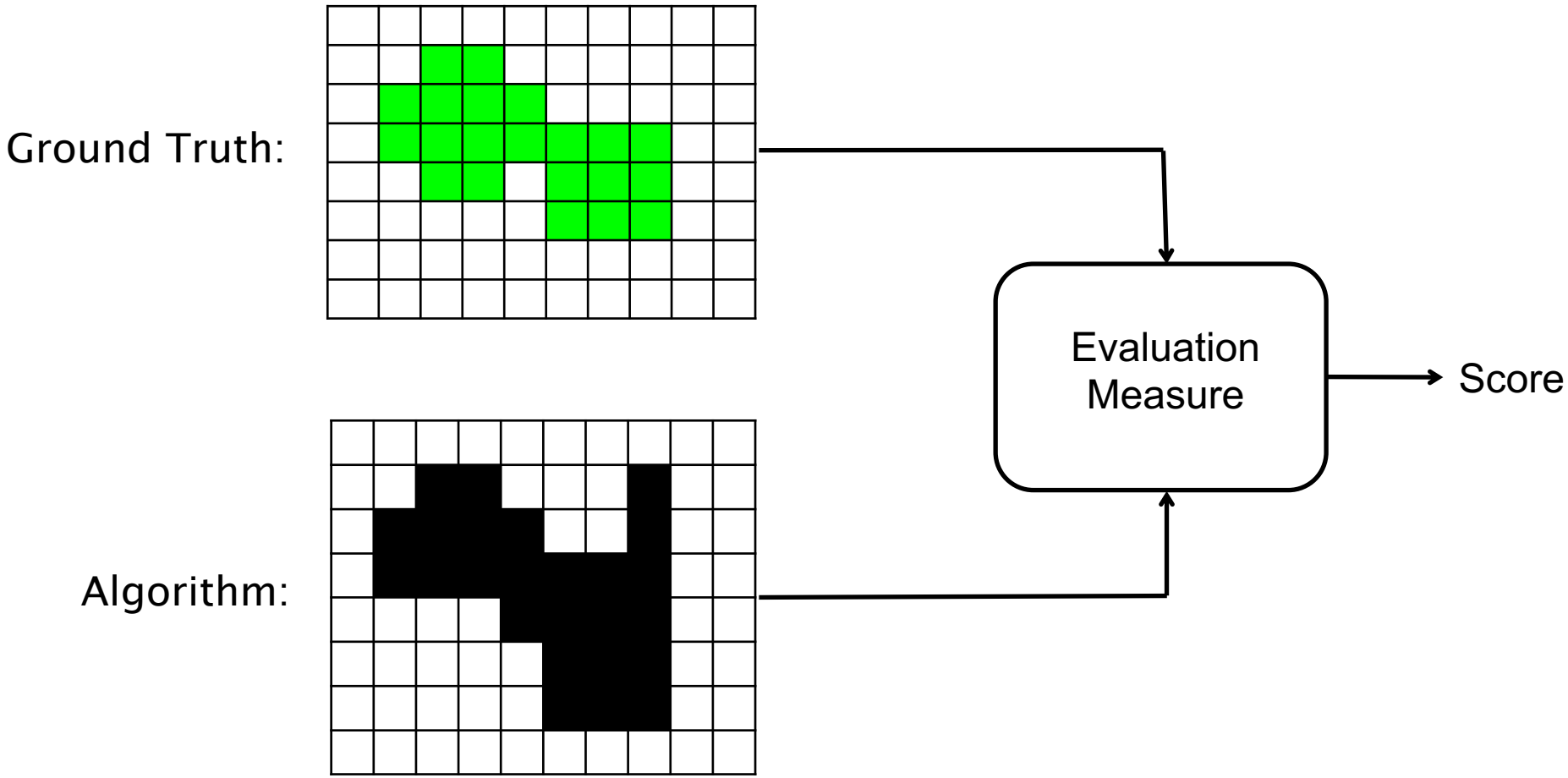
- Goal: locate all pixels belonging to 20 categories (e.g., person, cat, bus, mortorbike, potted plant, bottle) plus background
- Dataset: 11,530 images collected from Flickr and annotated by annotators at University of Leeds



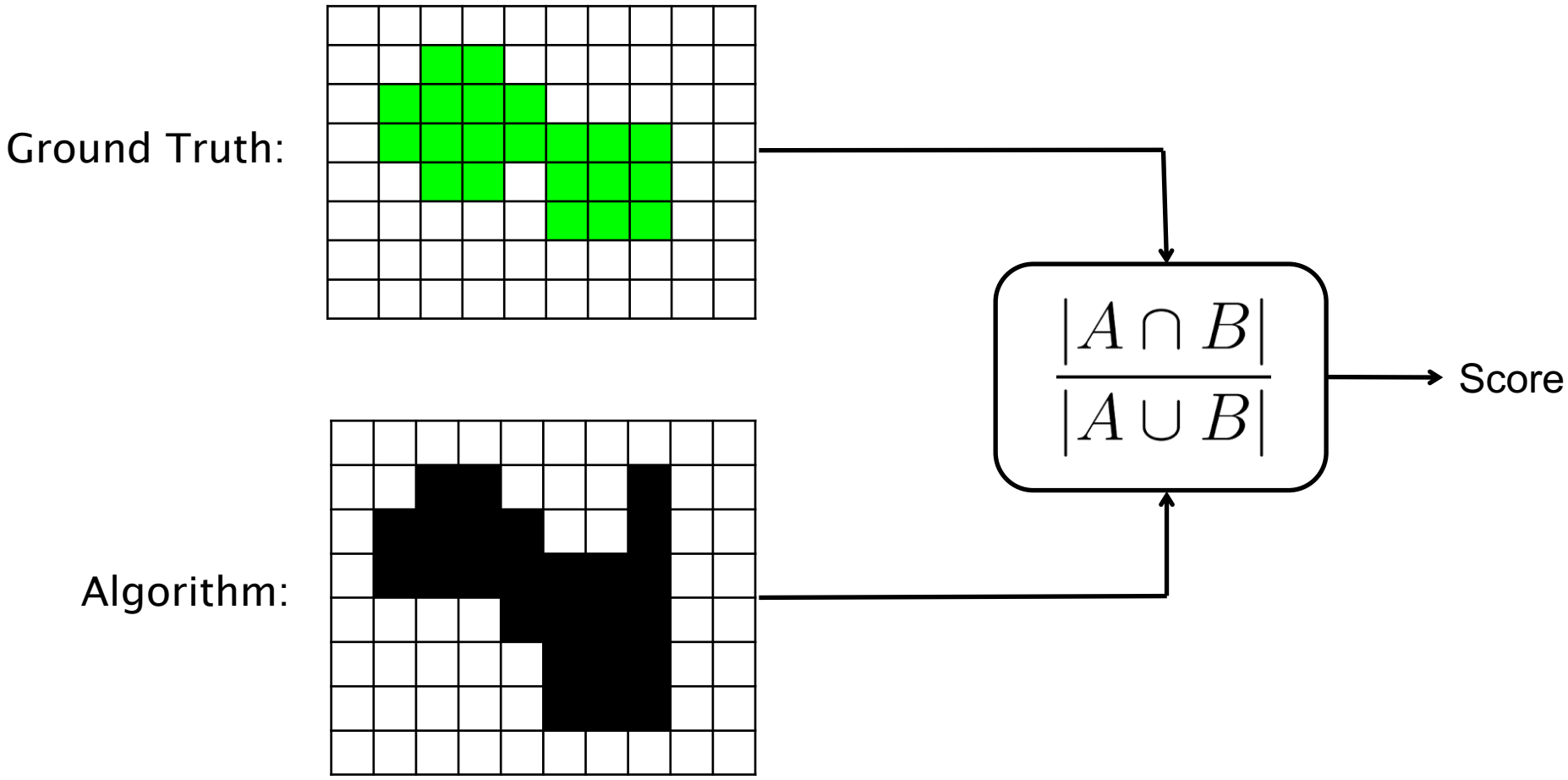
Dataset location: http://host.robots.ox.ac.uk/pascal/VOC/index.html

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

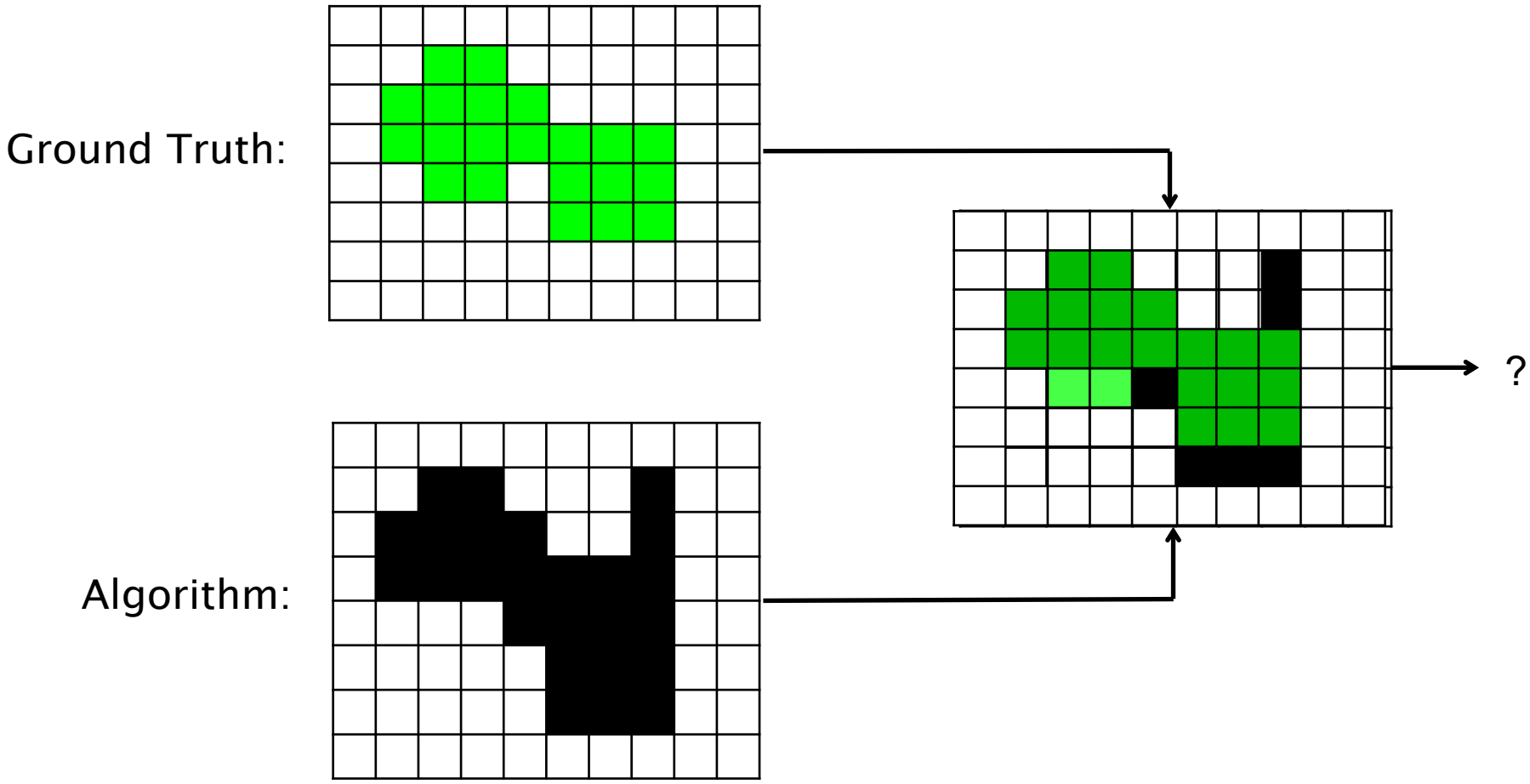
Evaluation Metric: Recall IoU



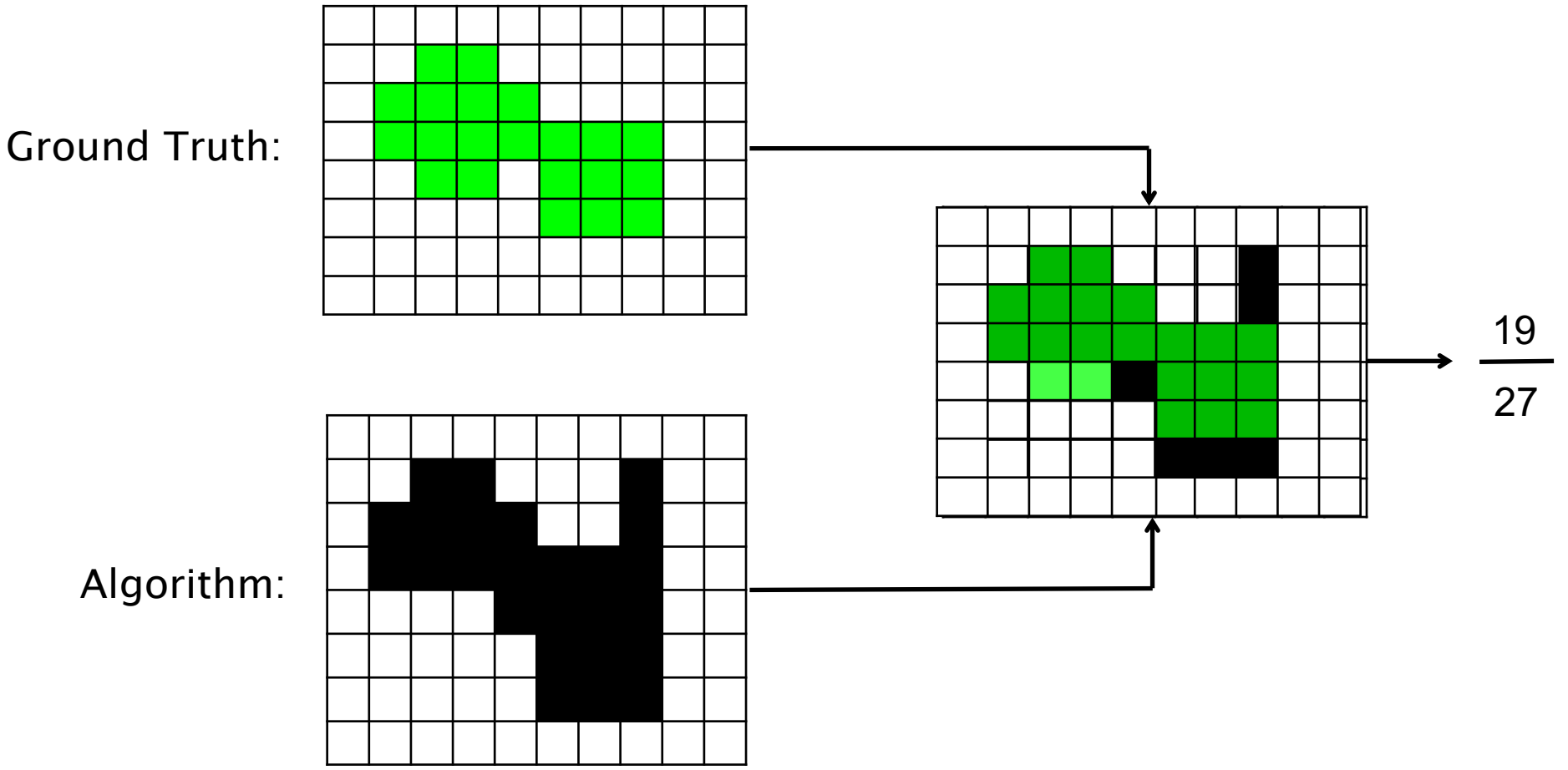
Evaluation Metric: Recall IoU



Evaluation Metric: Recall IoU



Evaluation Metric: Recall IoU



Evaluation Metric: Mean IoU (mIoU)

- Mean IoU score over all categories

Today's Topics

- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

Why Fully Convolutional Network?

Named after the proposed technique that excludes fully connected layers:

> Jonathon Long, Evan Shelhamer, and Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation.” CVPR 2015.

Key Novelties of Fully Convolutional Networks

First work for pixelwise prediction to:

1. Train fully convolutional networks end-to-end

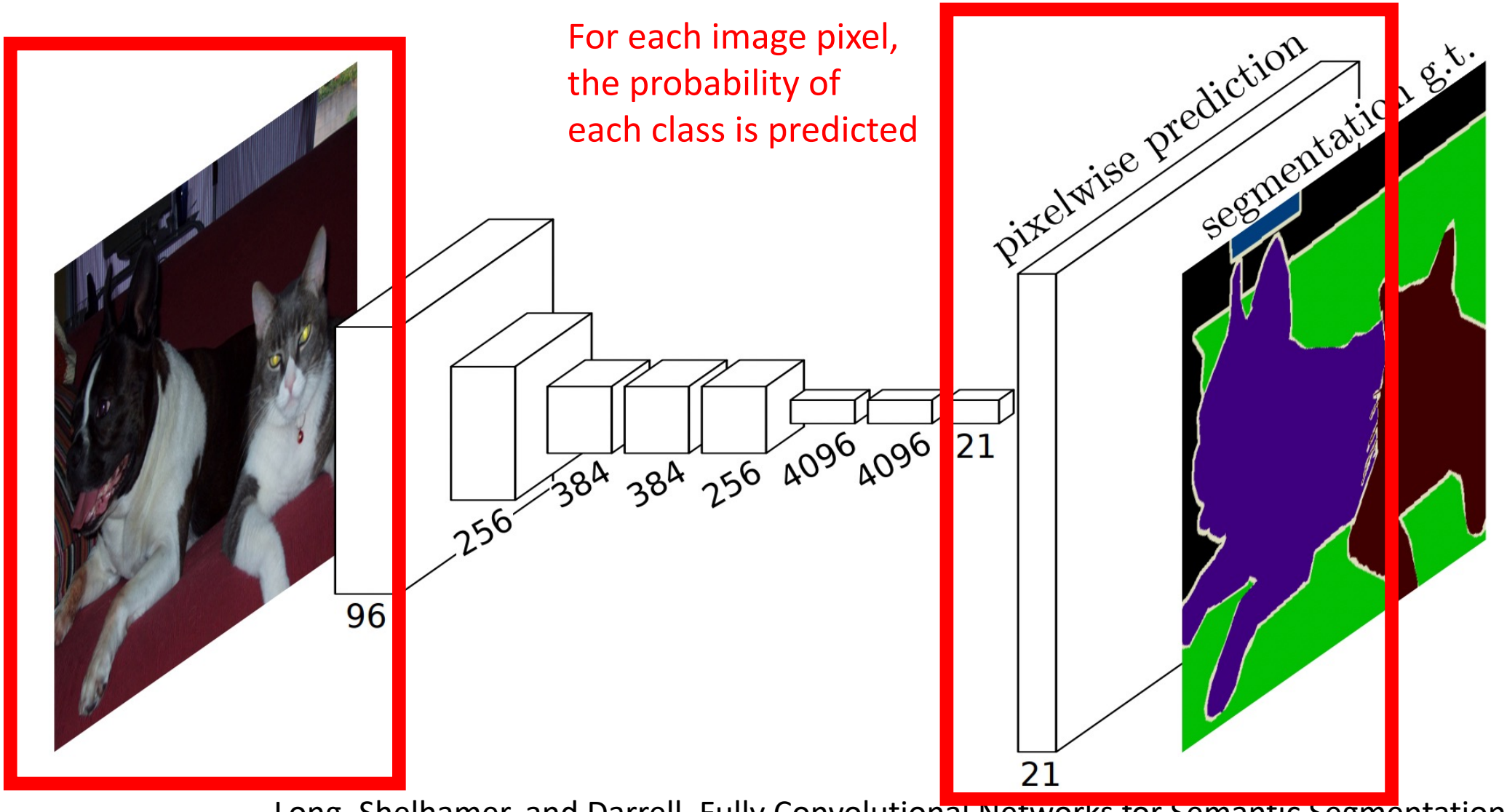
2. Use supervised pre-training (recall, R-CNN paper showed this can be a great idea when there is a scarce amount of annotated data)

Input: RGB image of ANY size

Output: Image of same size as input

Architecture

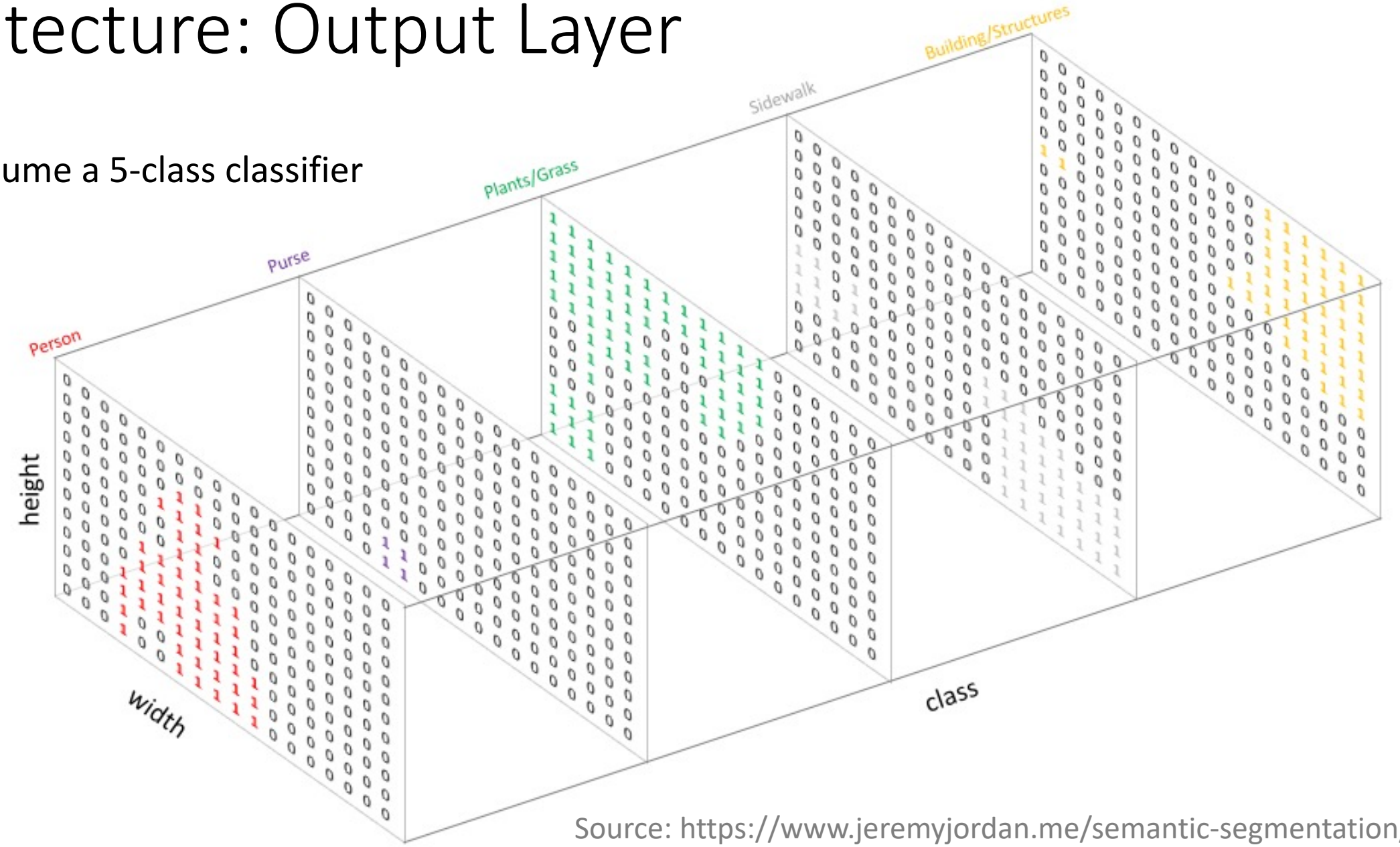
For each image pixel, the probability of each class is predicted



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Output Layer

- e.g., assume a 5-class classifier



Source: https://www.jeremyjordan.me/semantic-segmentation/

Architecture: Output Layer

- e.g., assume a 5-class classifier; output 1-hot encoding collapsed into single mask image

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

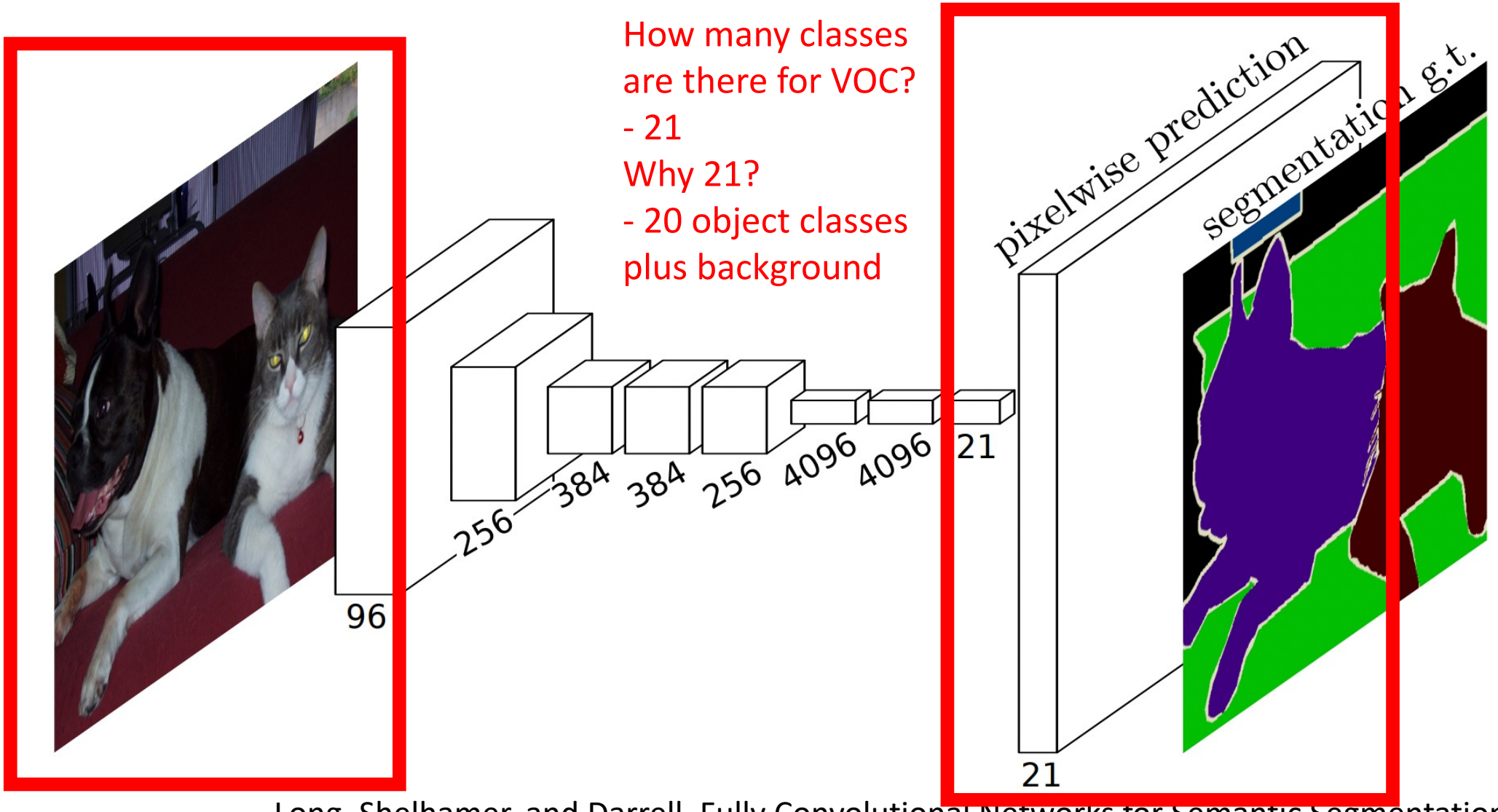
Source: https://www.jeremyjordan.me/semantic-segmentation/

Input: RGB image of ANY size

Output: Image of same size as input

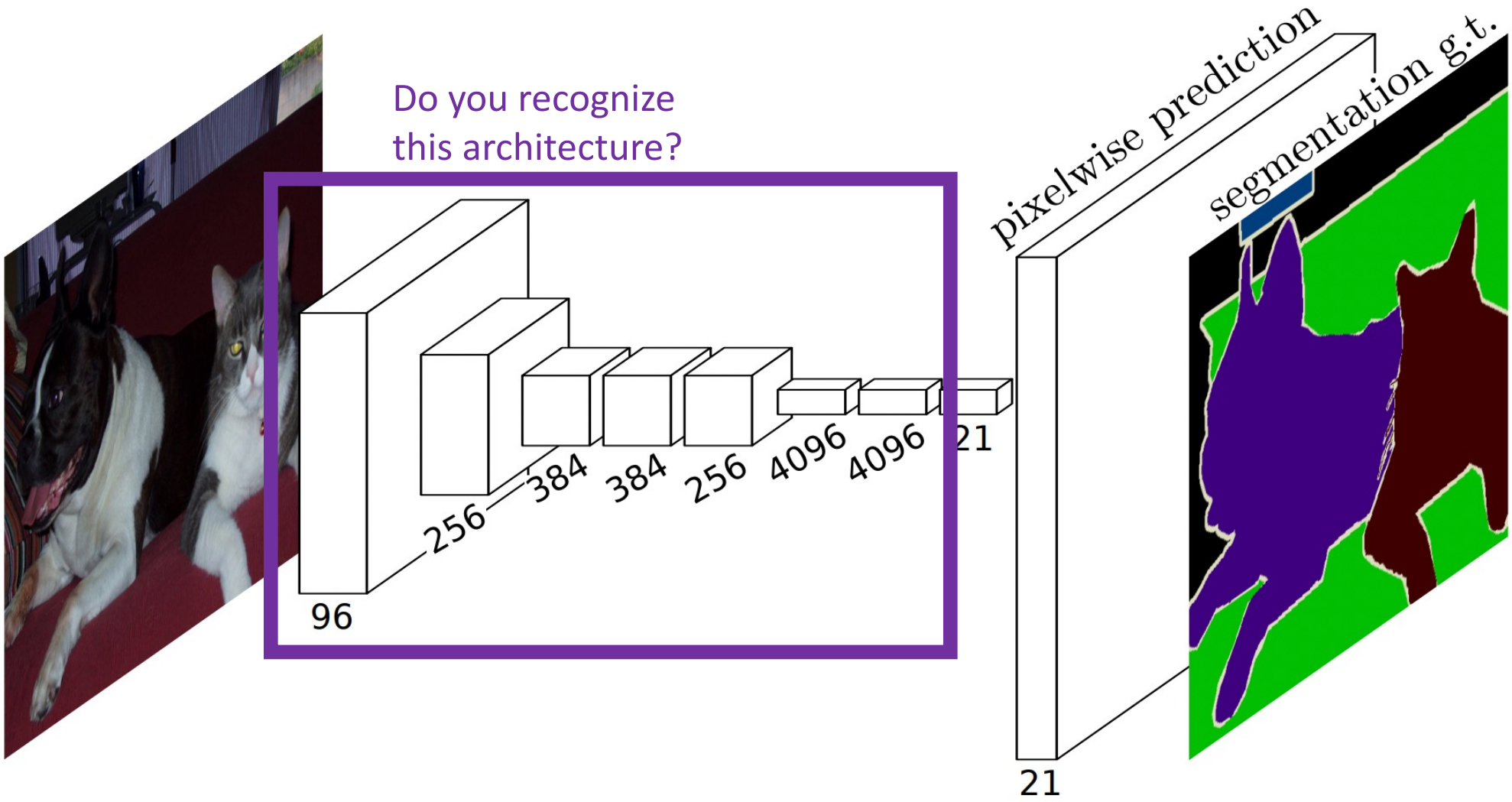
Architecture

How many classes are there for VOC?
- 21
Why 21?
- 20 object classes plus background



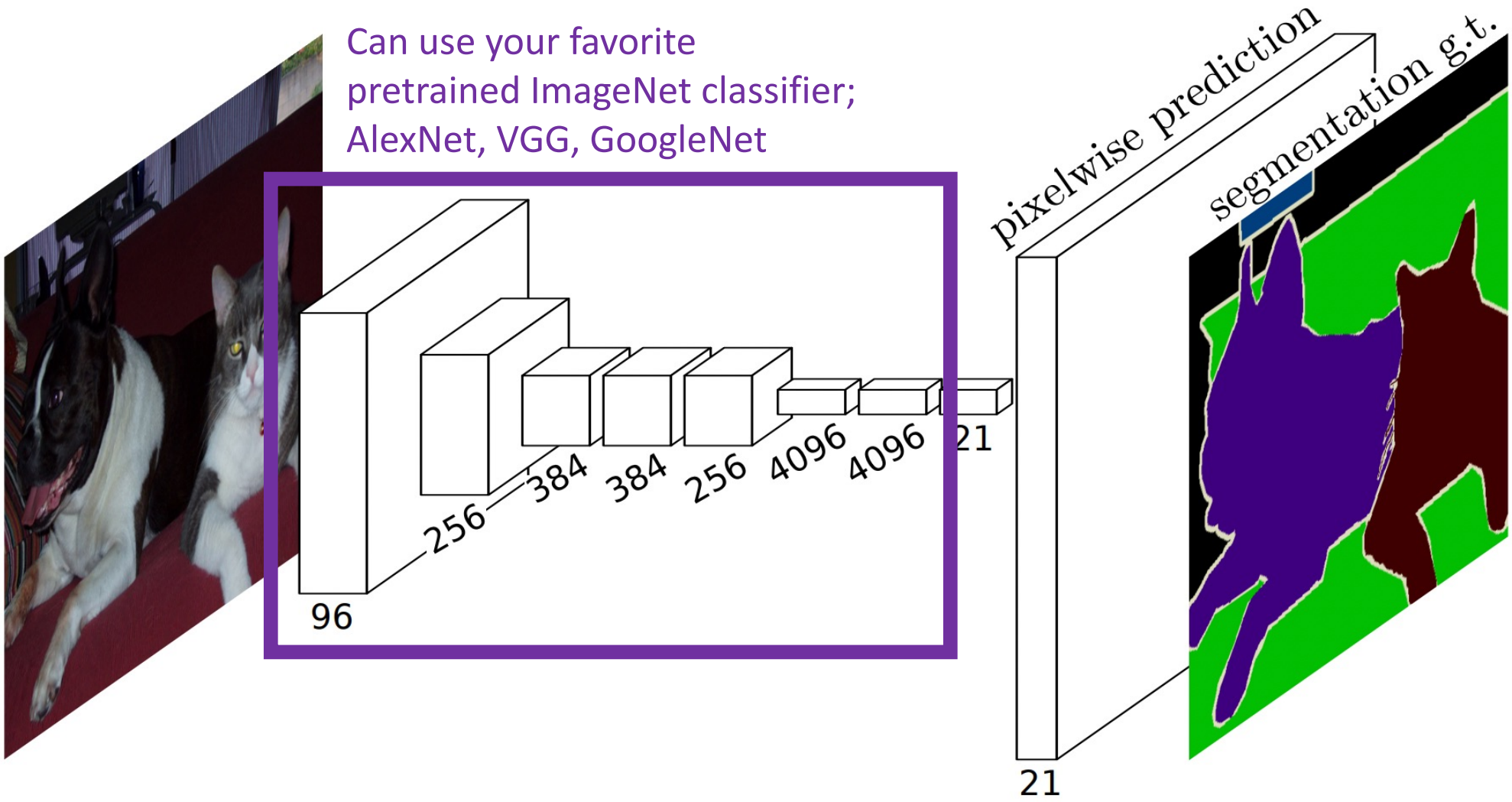
Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture

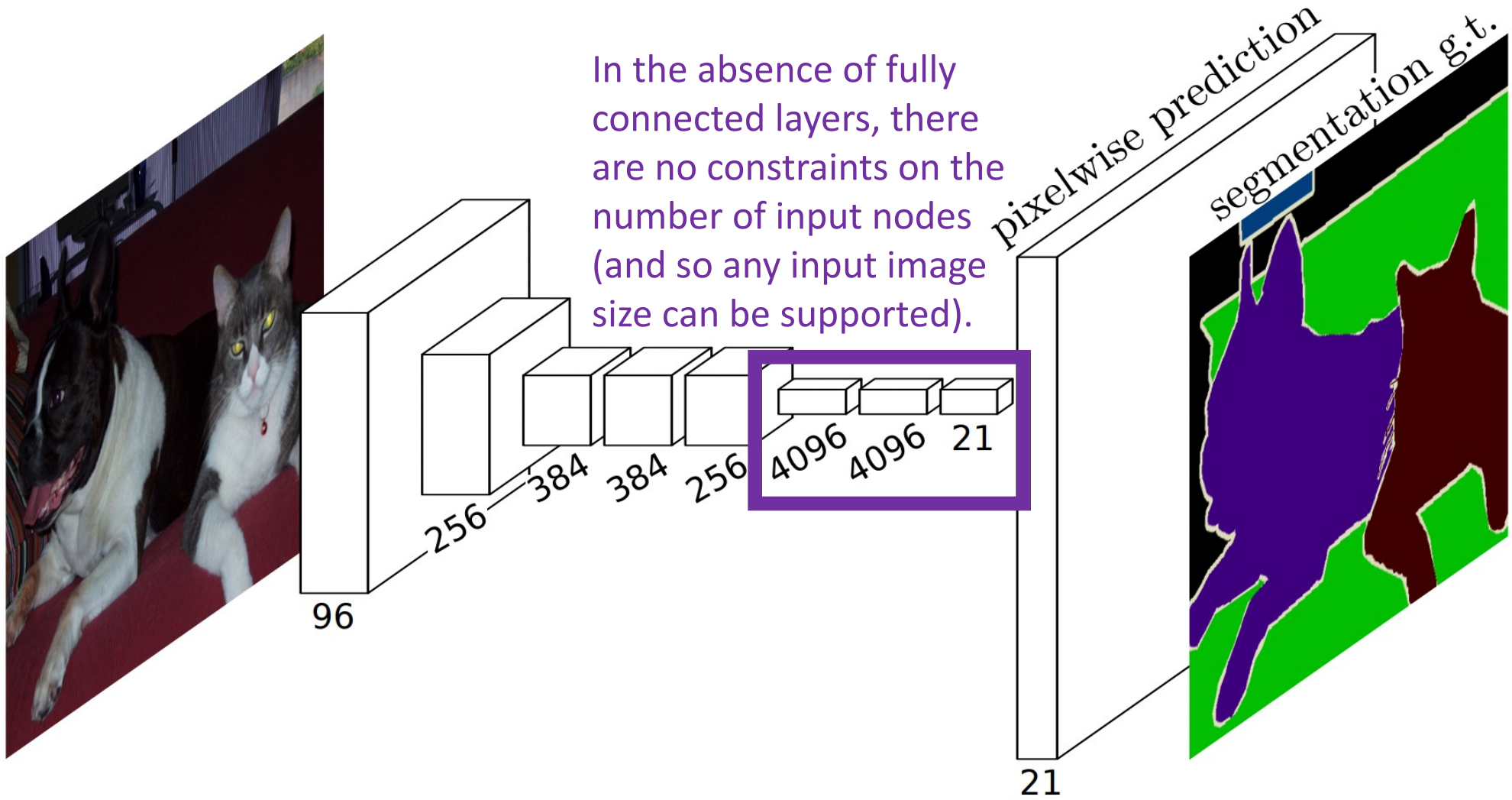


Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture

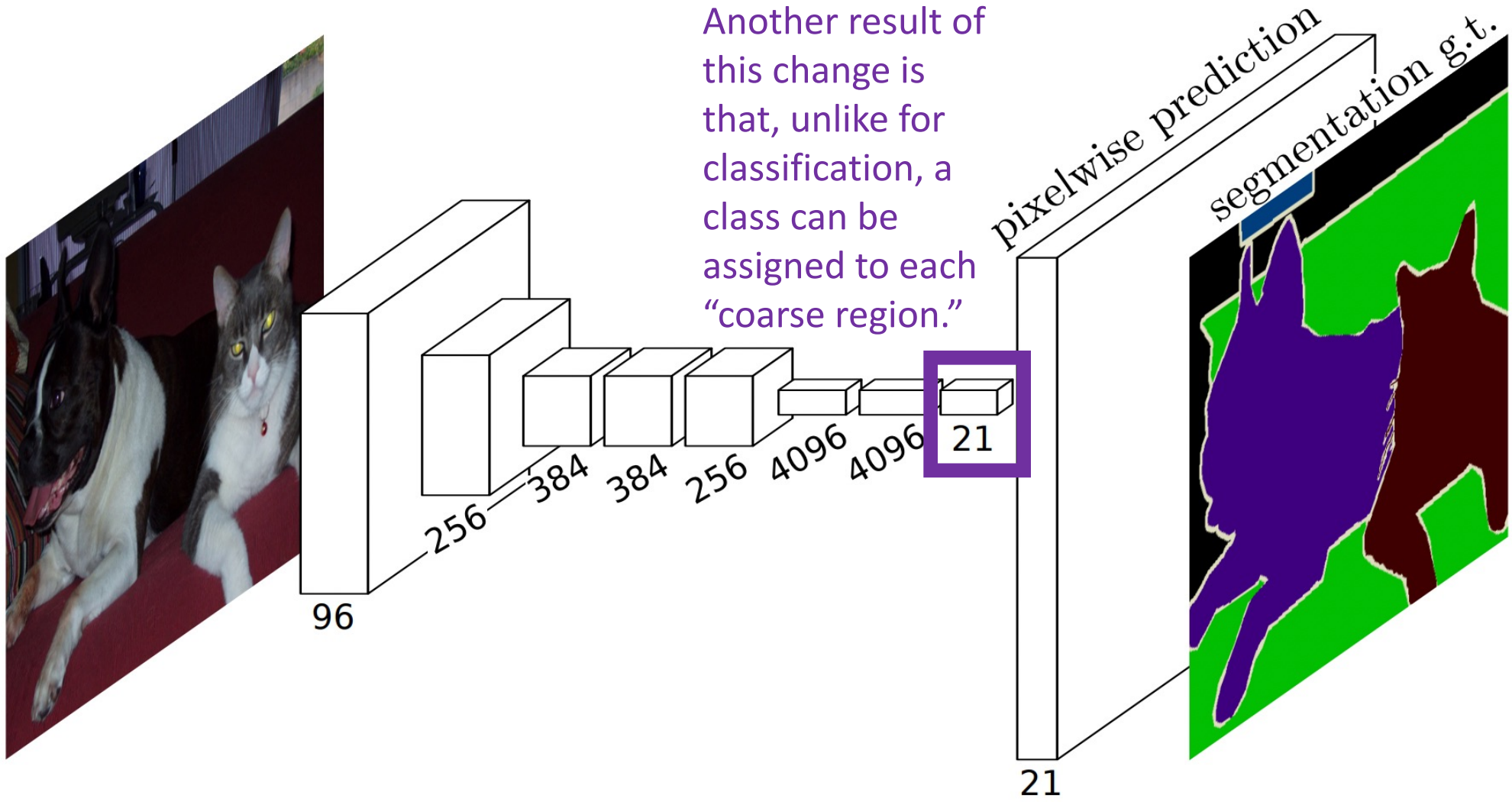
To make the architecture fully convolutional, fully connected layers are converted to convolutional layers.

In the absence of fully connected layers, there are no constraints on the number of input nodes (and so any input image size can be supported).



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

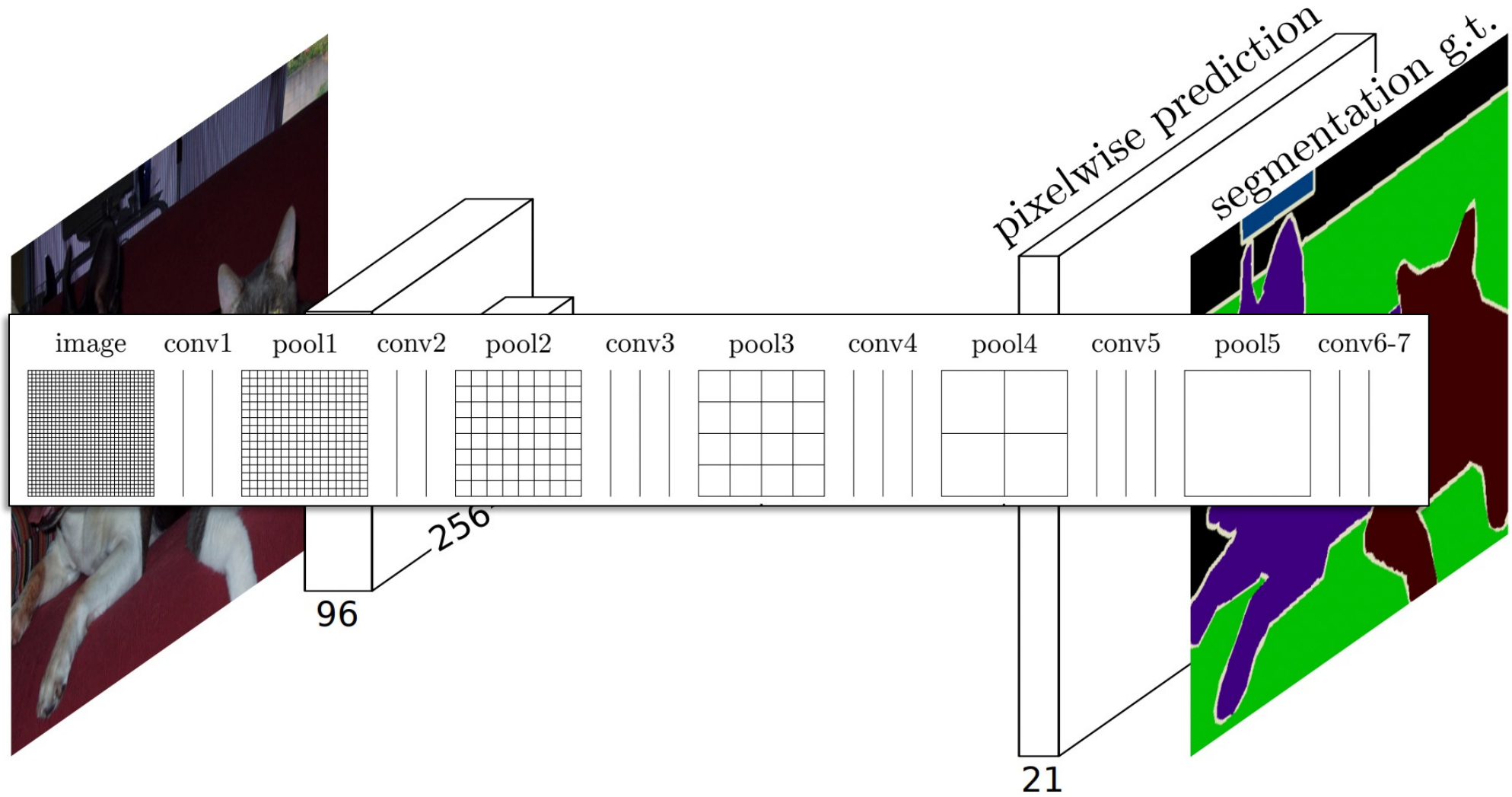
Architecture



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

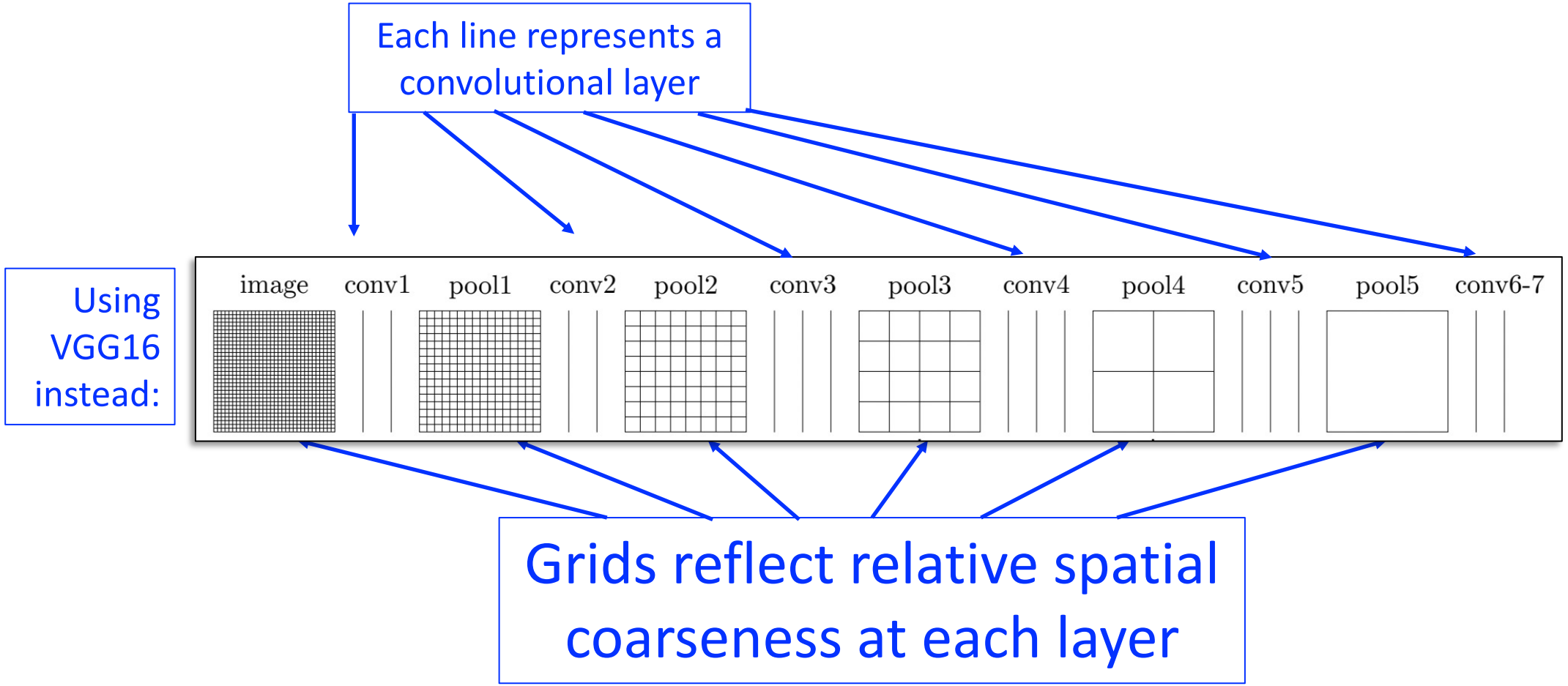
Architecture: Coarse Region Classification (Recall Intuition)

Using VGG16 instead:



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

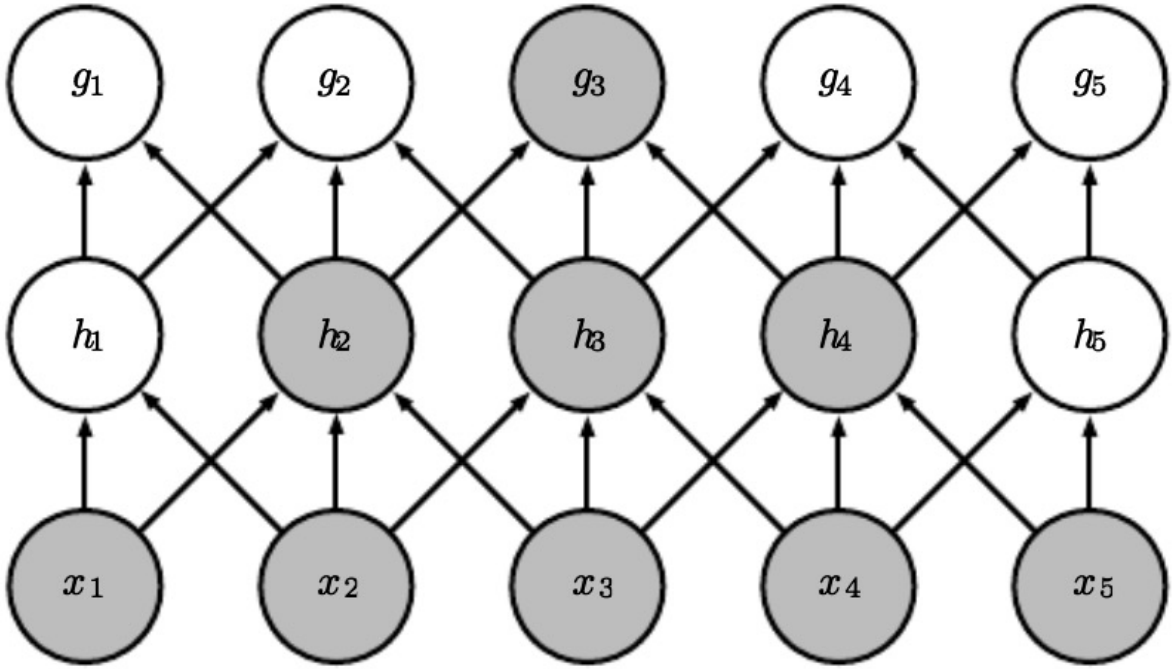
Architecture: Coarse Region Classification (Recall Intuition)



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

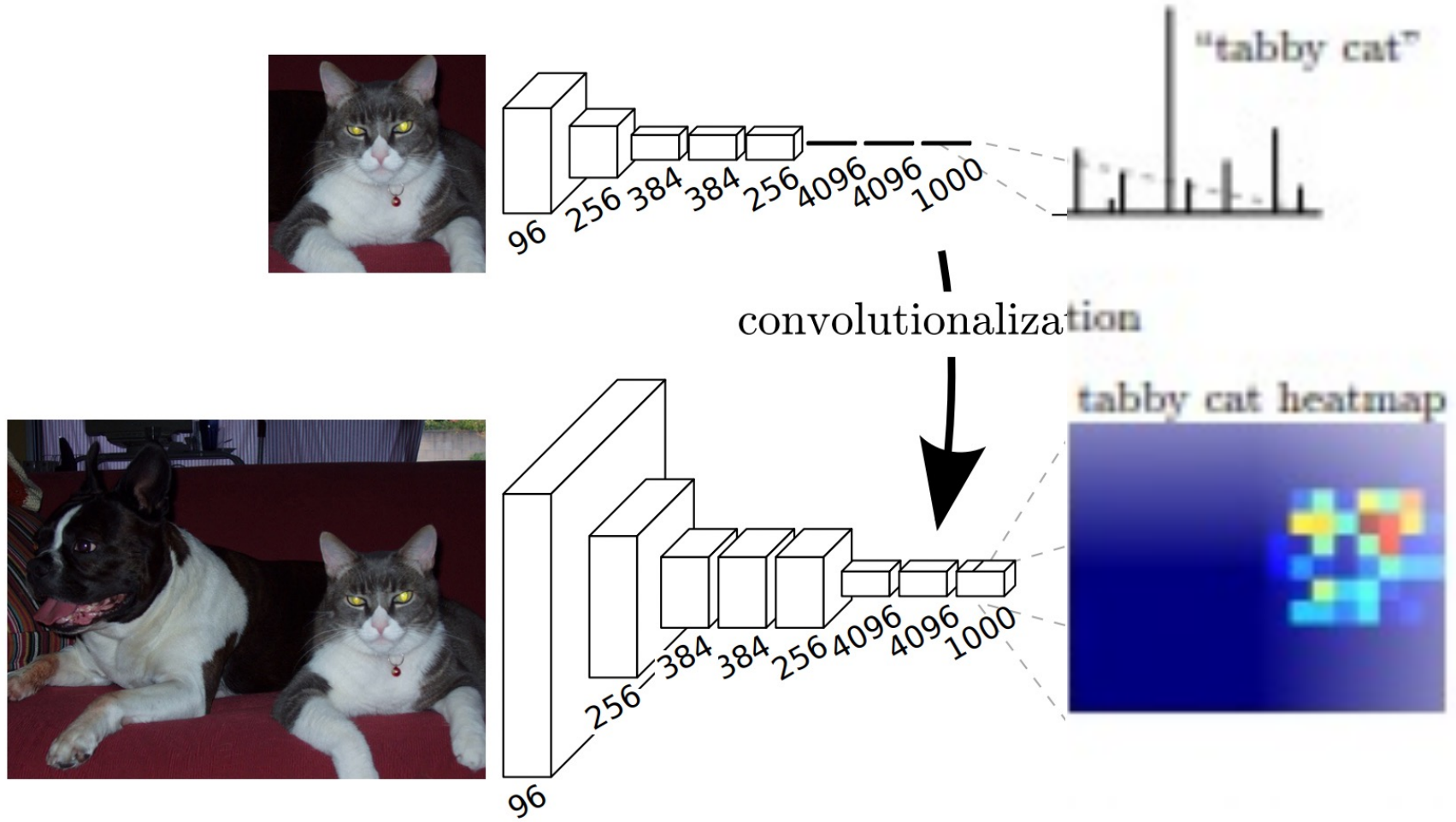
Architecture: Coarse Region Classification (Recall Intuition)

Stacking many convolutional layers leads to learning patterns in increasingly **larger regions of the input (e.g., pixel) space.**



https://www.deeplearningbook.org/contents/convnets.html

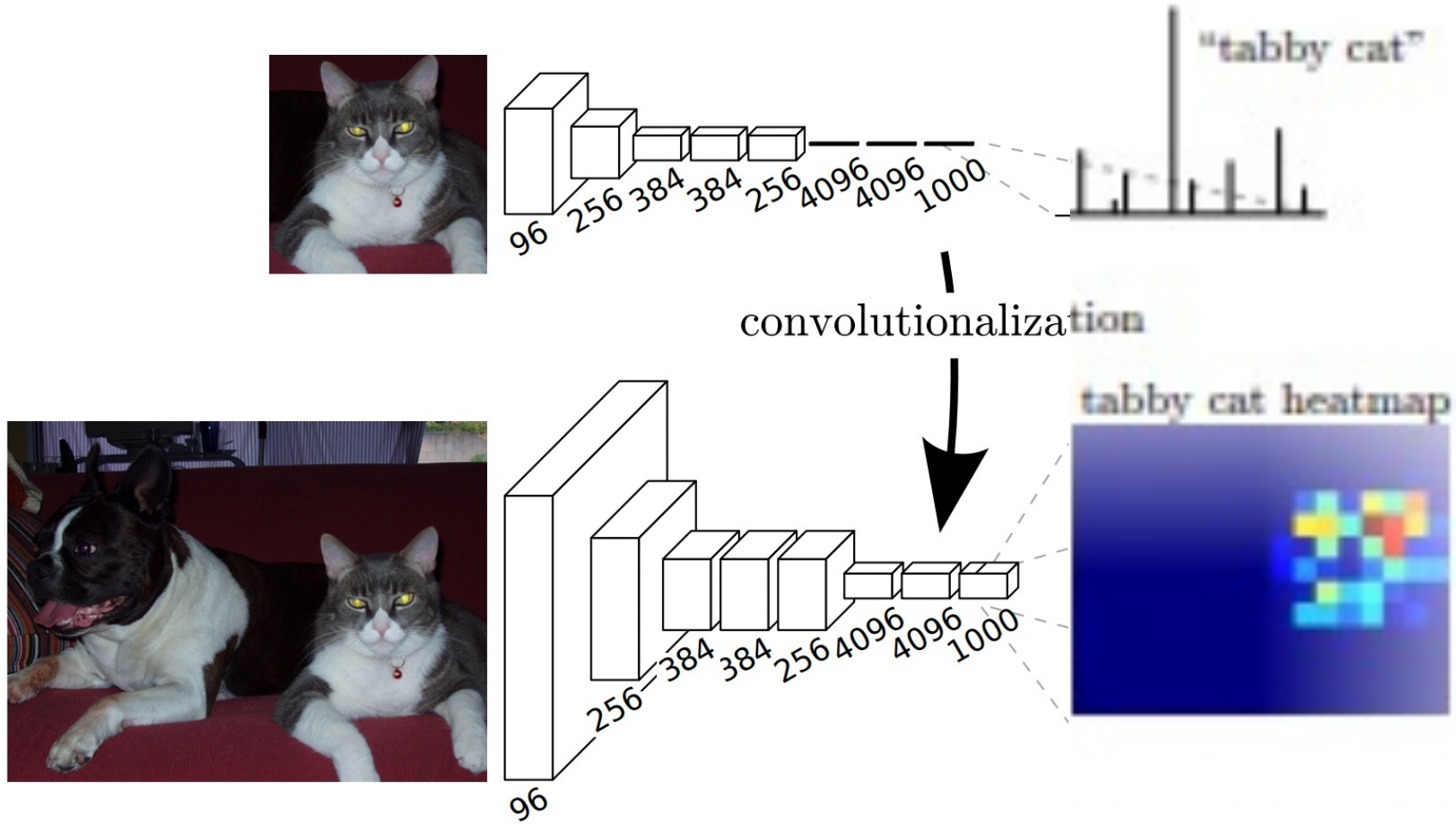
Architecture: Fully vs Convolution Layers



Each slice indicates the likelihood each pixel in the coarse region belongs to the class identified by the filter

Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

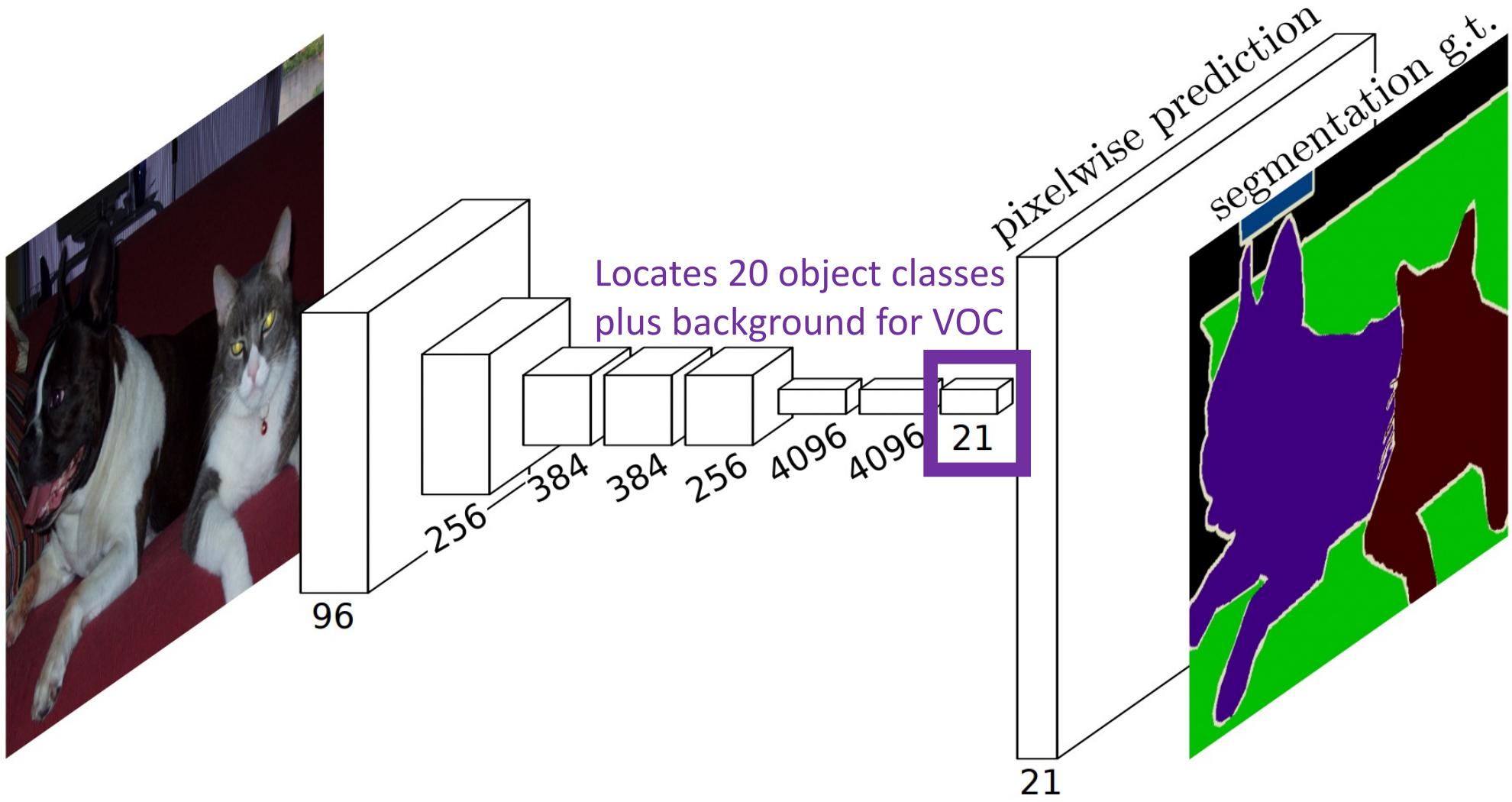
Architecture: Fully vs Convolution Layers



If convolutionizing ImageNet trained classifiers, how many classes would be predicted for each coarse region?

Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

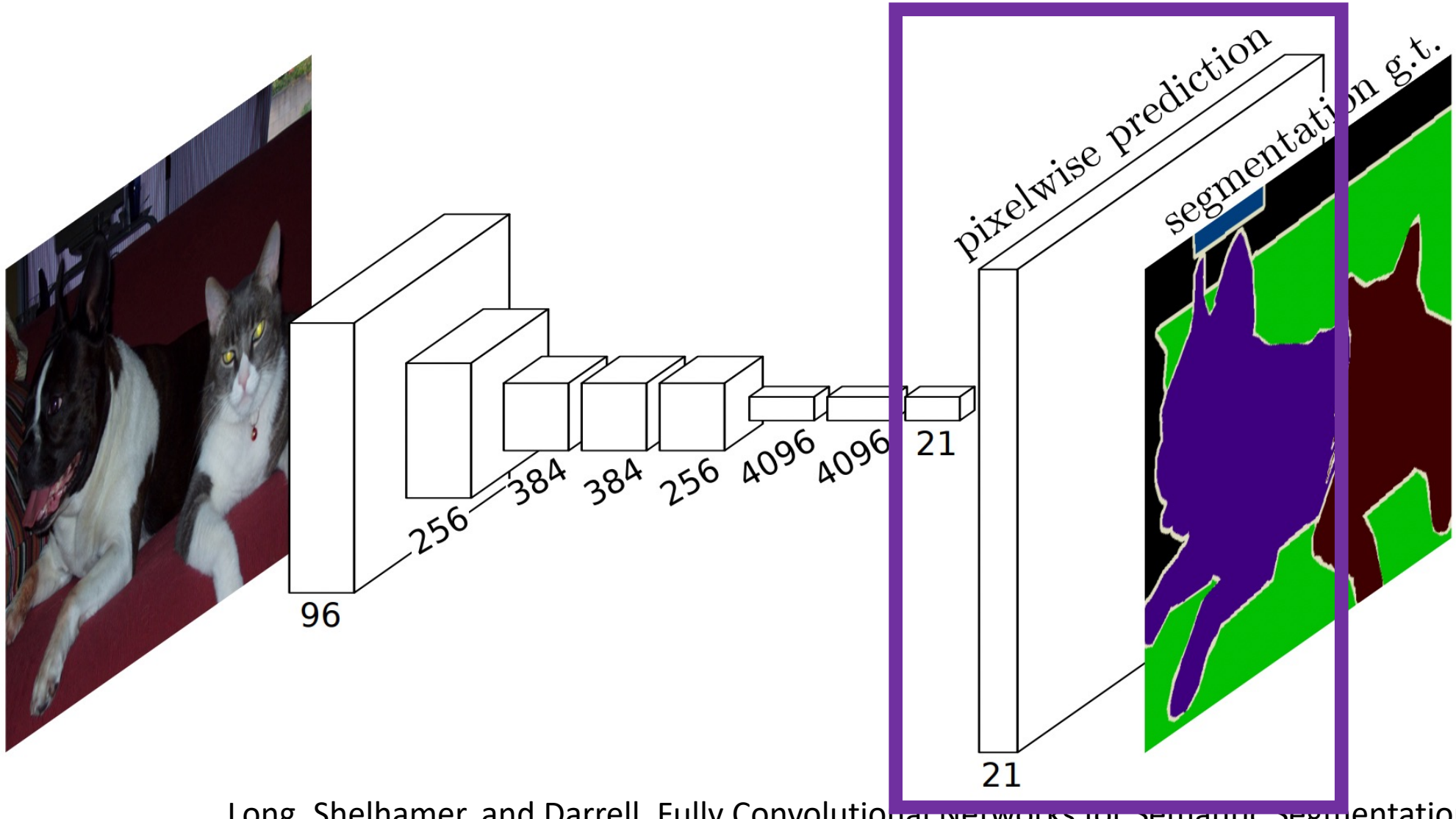
Architecture: Coarse Region Classification



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

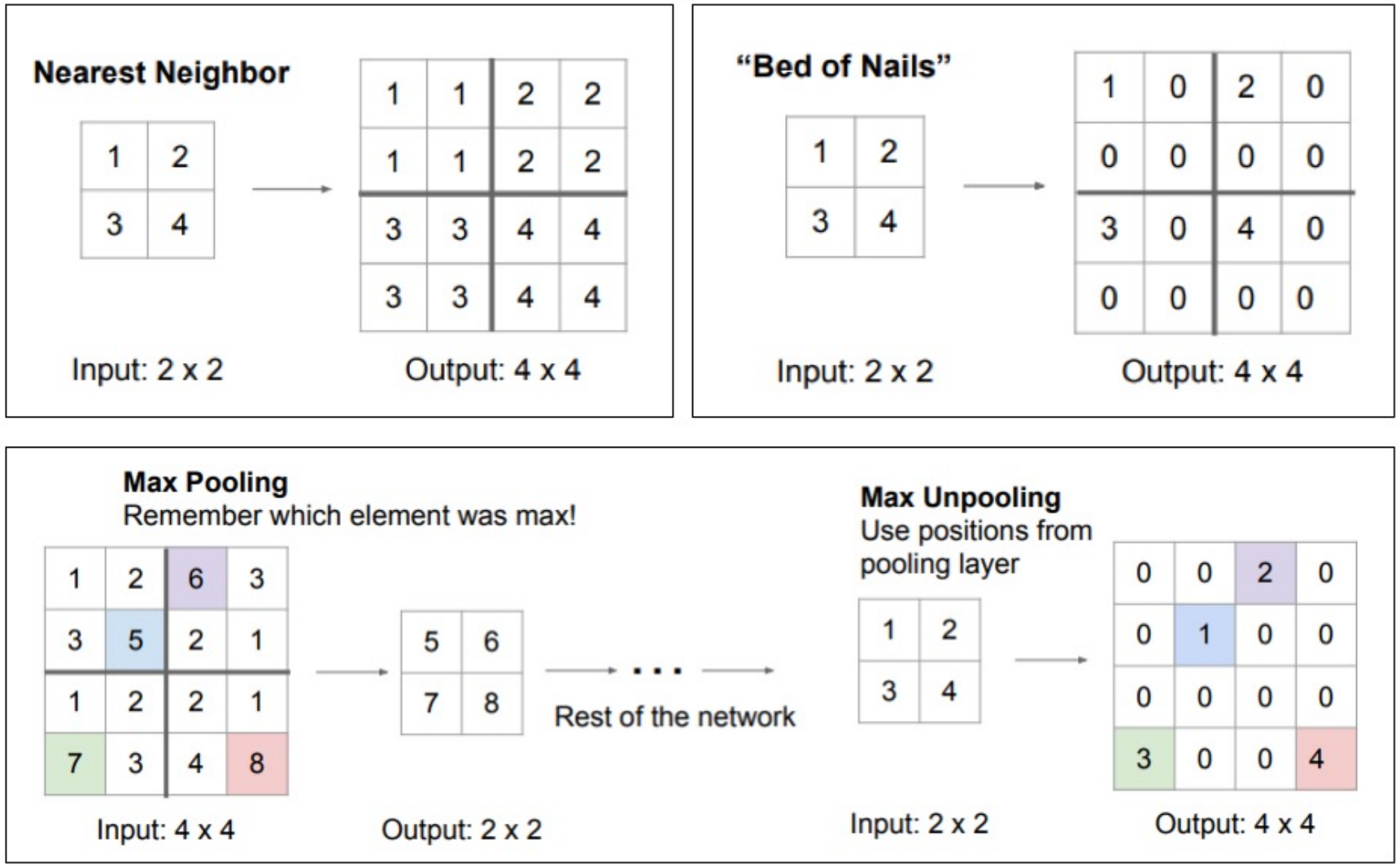
Architecture

Challenge: how to decode from coarse region classifications to per pixel classification?



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Upsampling (Many Approaches)



Source: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

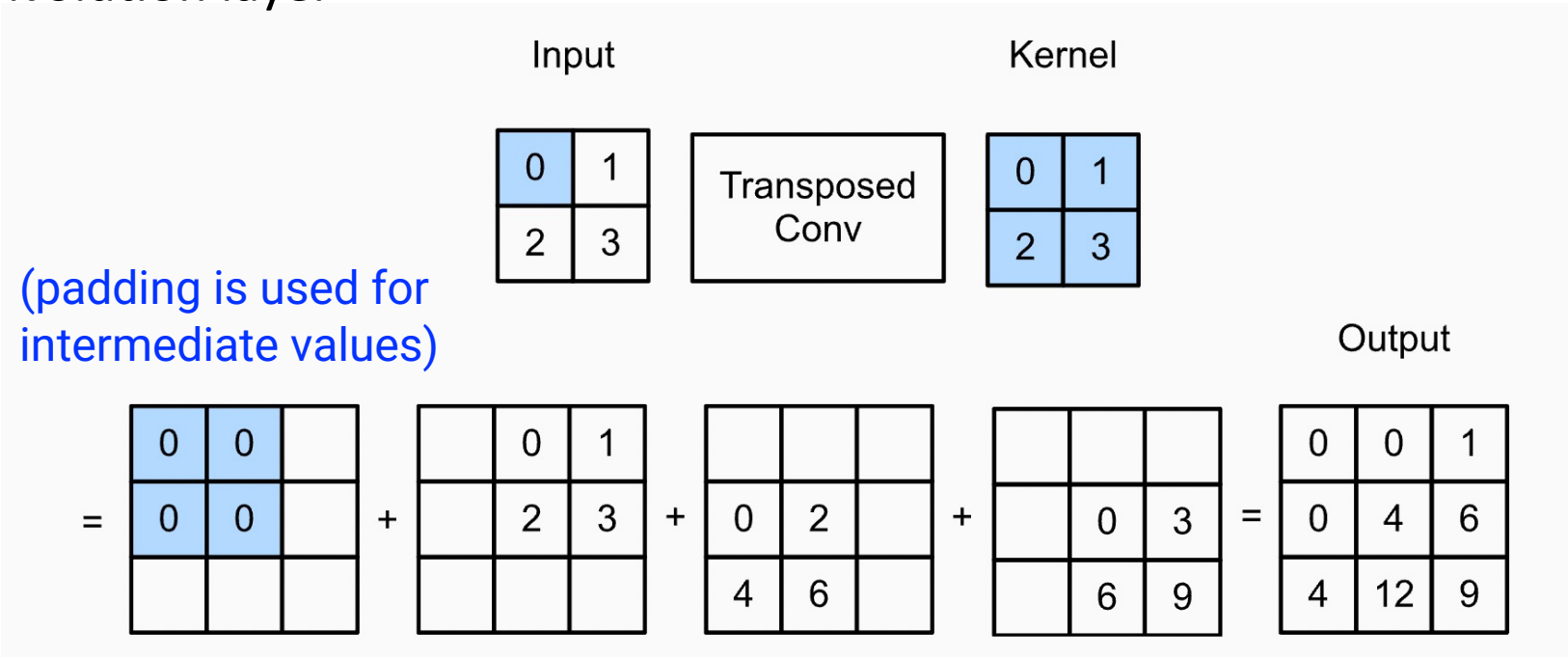
Architecture: Upsampling (Transposed Convolutional Layer)

- **Idea**: learn convolutional filters with a fractional sized stride to upsample the coarse image while refining it; e.g., 1/2 stride
- Also called “fractional convolutional layer”, “backward convolution”, and, incorrectly, ”deconvolution layer”

https://www.machinecurve.com/index.php/2019/09/29/understanding-transposed-convolutions/#the-goal-reconstructing-the-original-input

Architecture: Upsampling (Transposed Convolutional Layer)

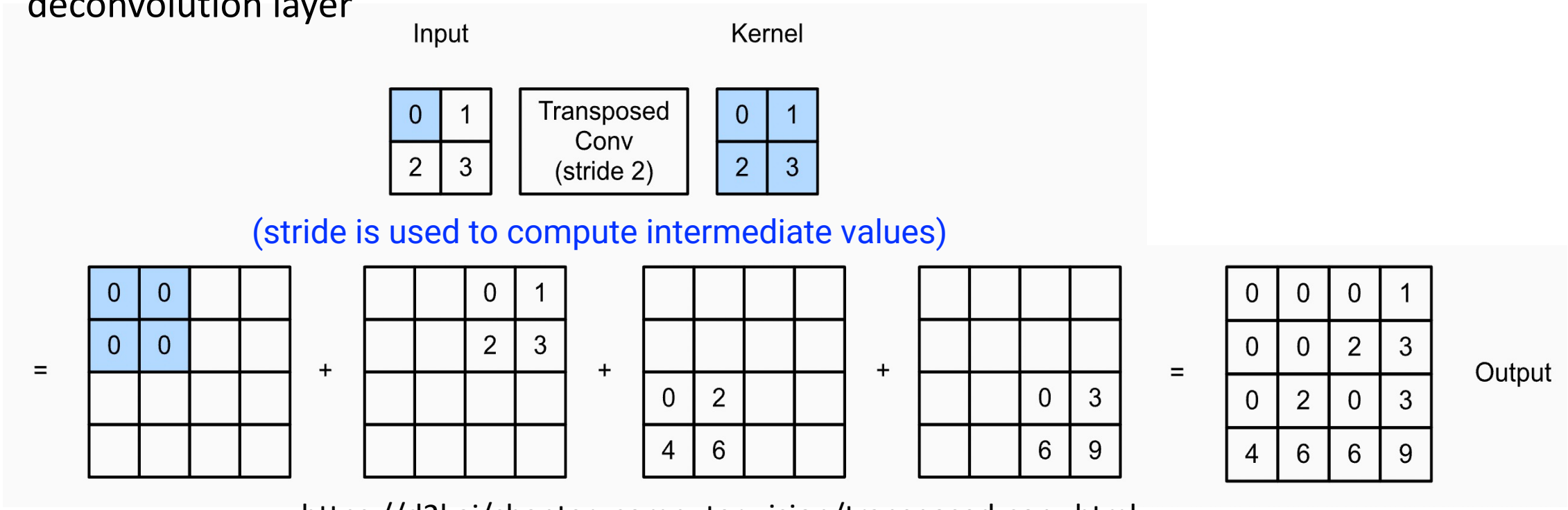
- **Idea**: learn convolutional filters with a fractional sized stride to upsample the coarse image while refining it; e.g., 1/2 stride
- Also called “fractional convolutional layer”, “backward convolution”, and, incorrectly, ”deconvolution layer”



https://d2l.ai/chapter_computer-vision/transposed-conv.html

Architecture: Upsampling (Transposed Convolutional Layer)

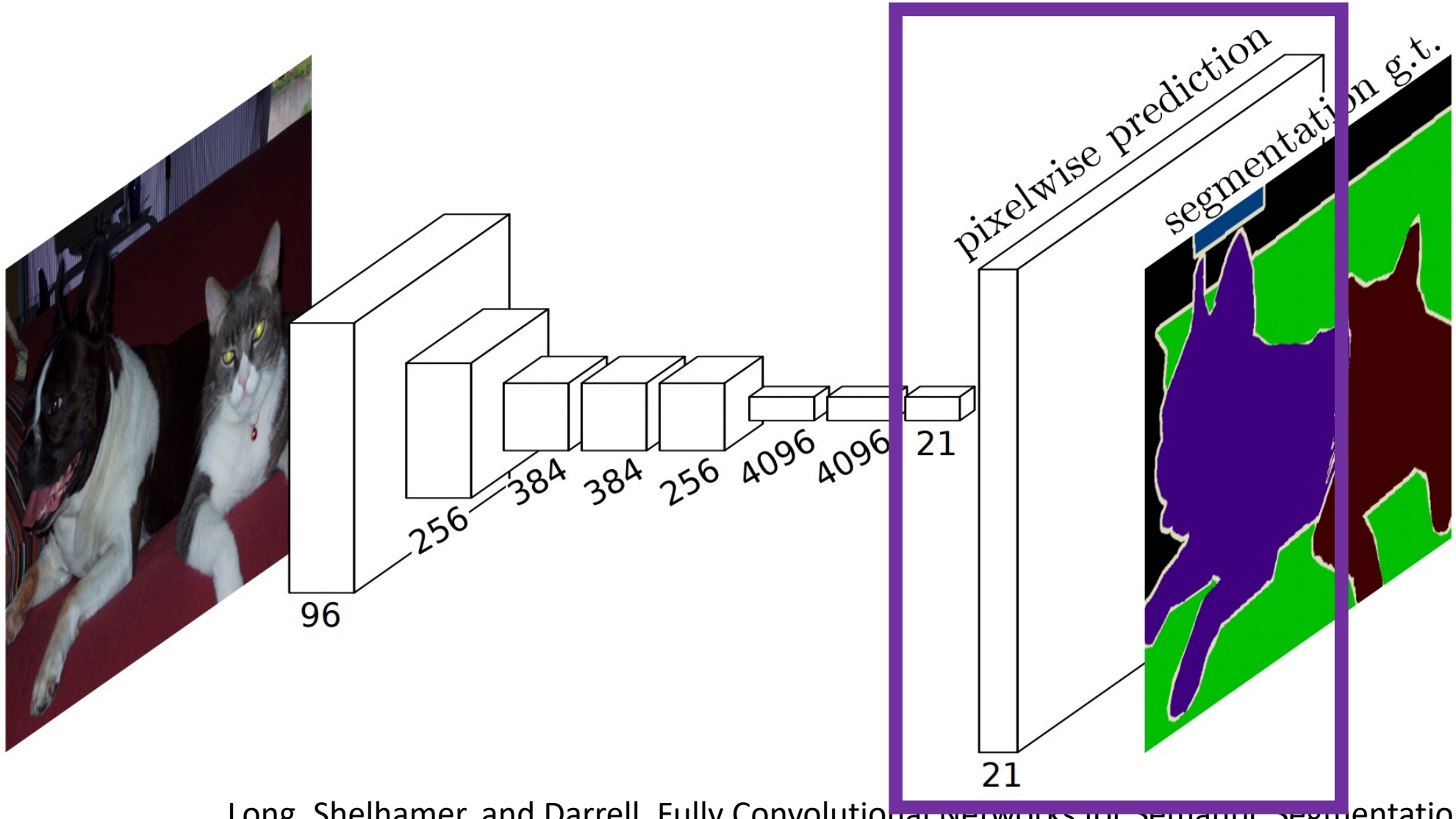
- **Idea**: learn convolutional filters with a fractional sized stride to upsample the coarse image while refining it; e.g., 1/2 stride
- Also called “fractional convolutional layer”, “backward convolution”, and, incorrectly, ”deconvolution layer”



https://d2l.ai/chapter_computer-vision/transposed-conv.html

Architecture

Next challenge: how to decode a **highly detailed** per pixel classification from the coarse region classifications?



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Results

Next challenge: how to decode a **highly detailed** per pixel classification from the coarse region classifications?

Ground truth target

Predicted segmentation

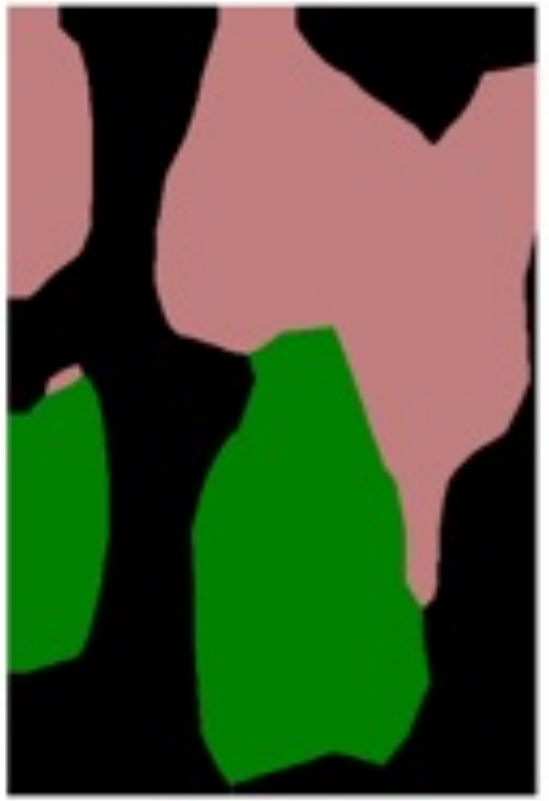
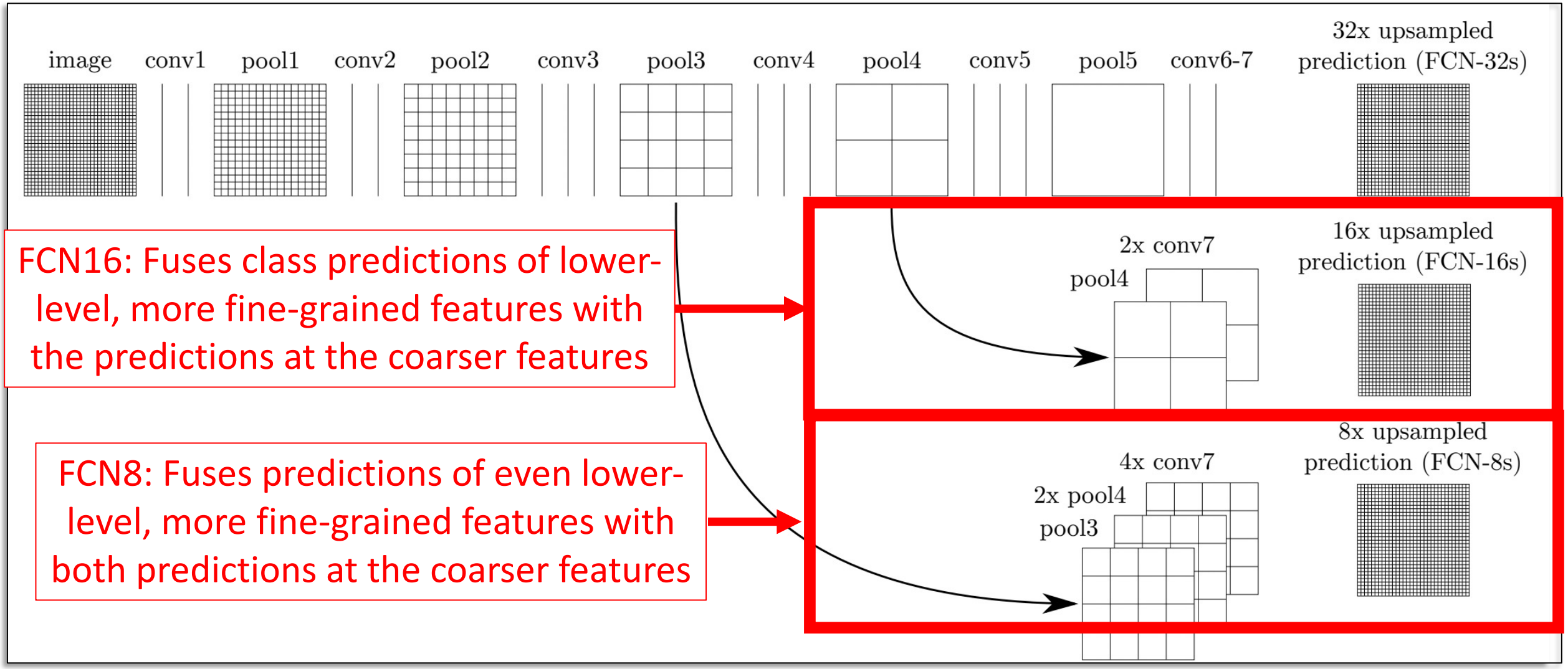
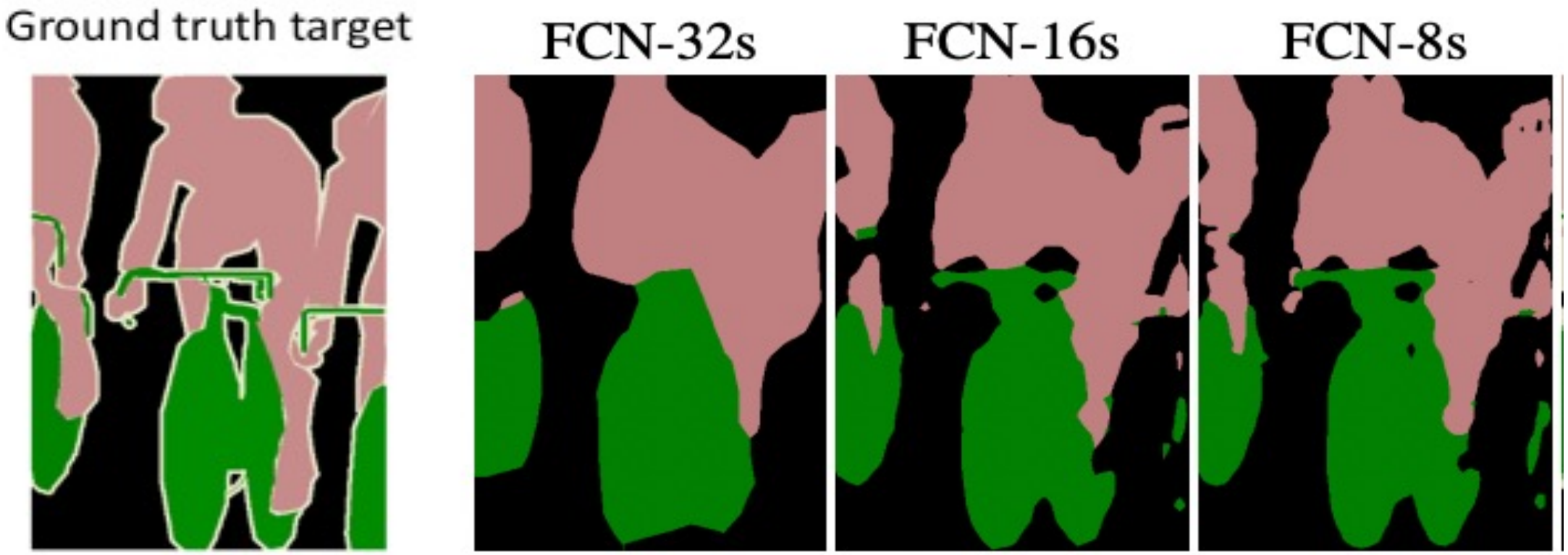


Figure source: https://www.jeremyjordan.me/semantic-segmentation/

Architecture: Update to Use Skip Connections



Architecture: Results

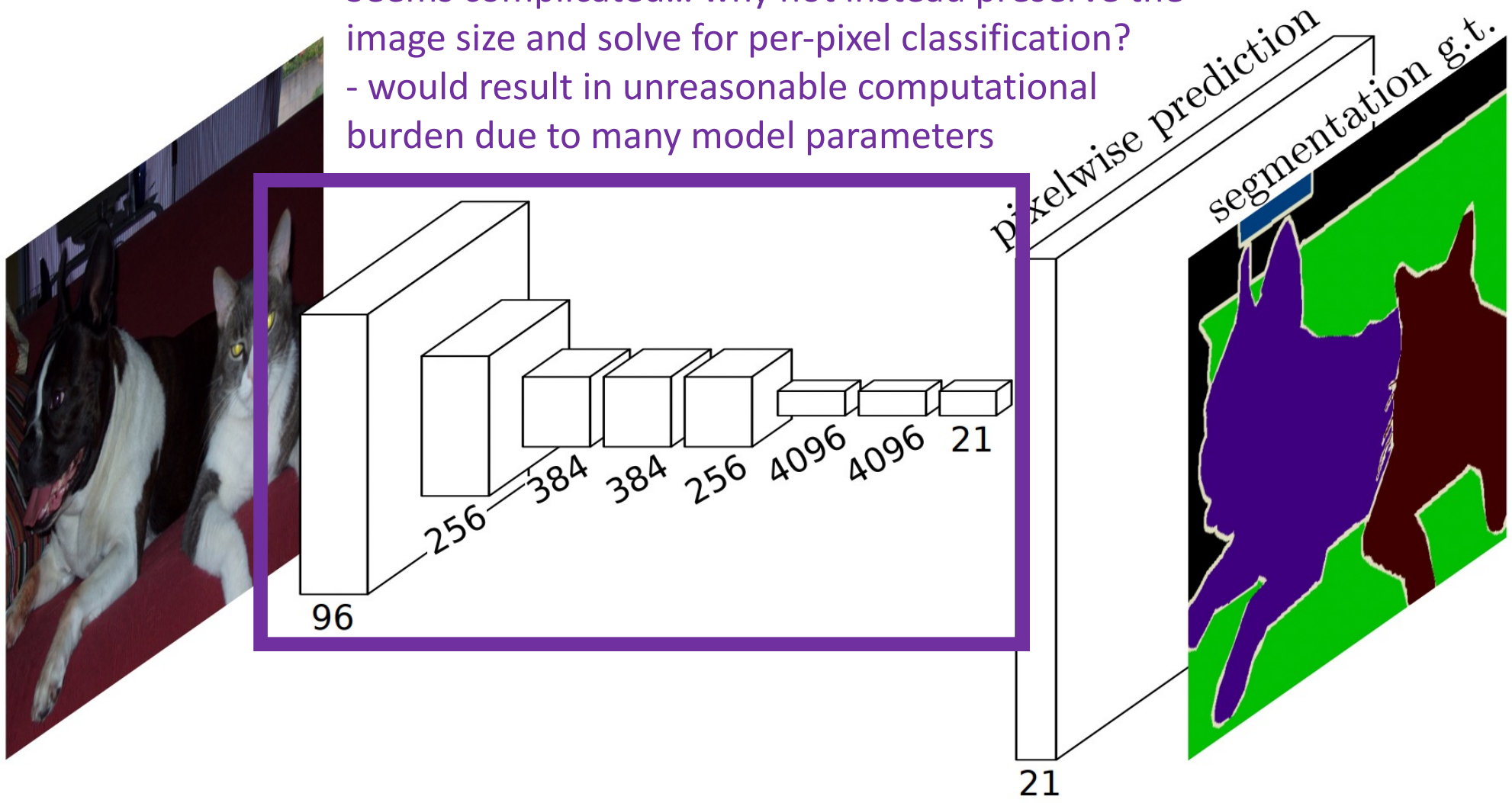


Skip connections support capturing finer-grained details while retaining correct semantic information!

Figure source: https://www.jeremyjordan.me/semantic-segmentation/

Architecture: Upsampling + Skip Connections

Seems complicated… why not instead preserve the image size and solve for per-pixel classification?
- would result in unreasonable computational burden due to many model parameters

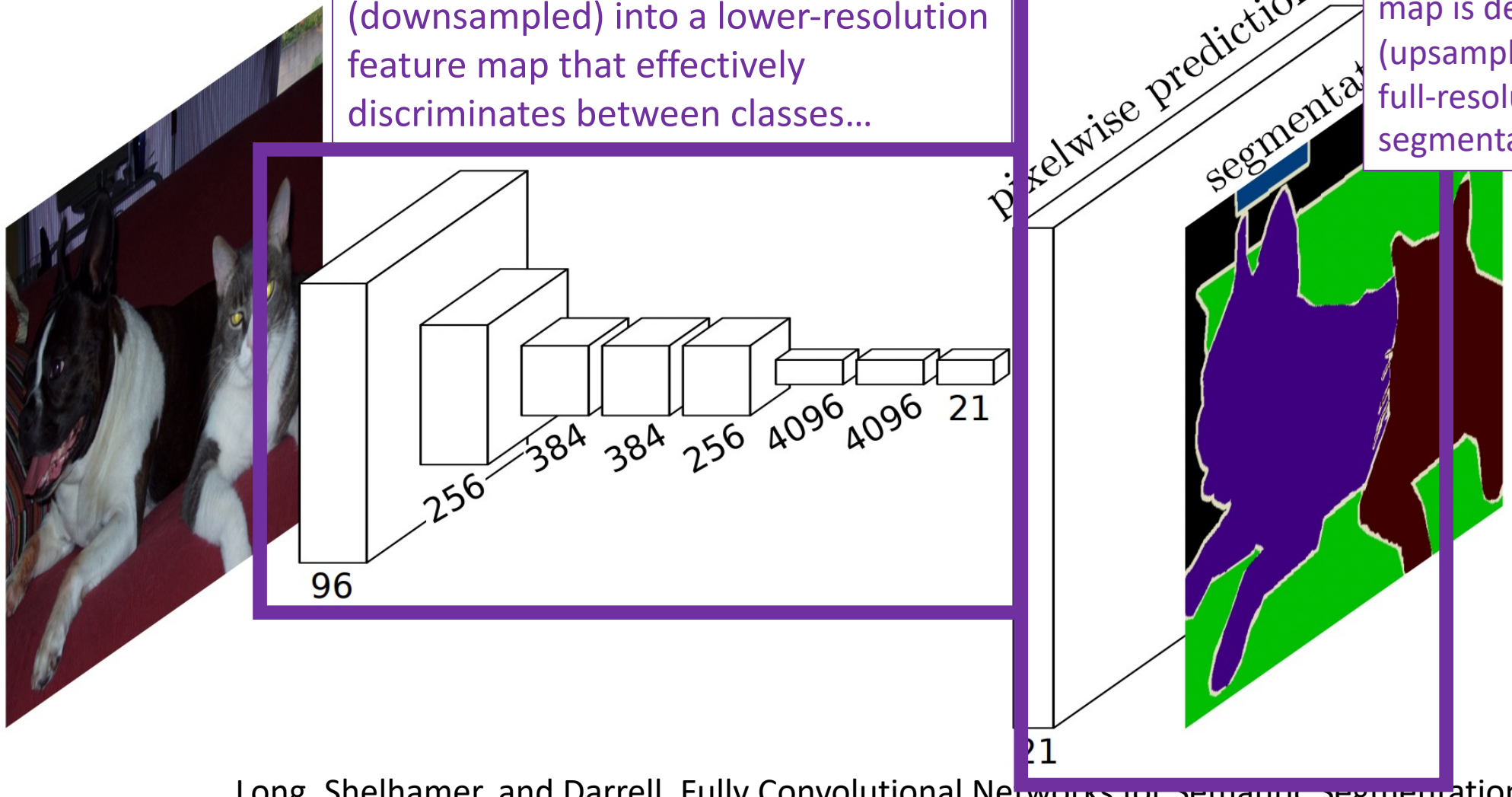


Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Encoder Decoder Architecture

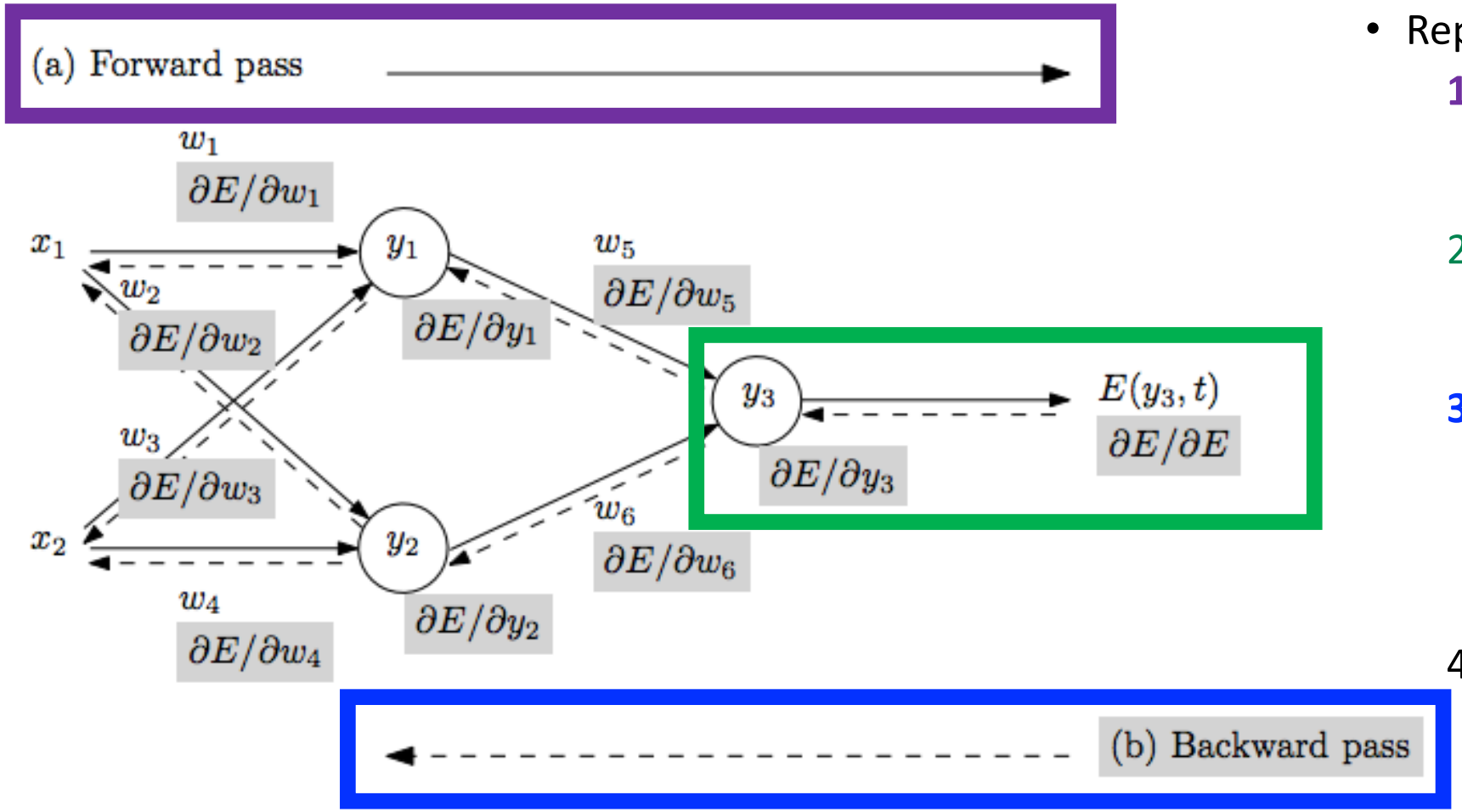
For efficiency, the image is encoded (downsampled) into a lower-resolution feature map that effectively discriminates between classes…

Then, the feature map is decoded (upsampled) into a full-resolution segmentation map.



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Training: Took 3 days on 1 GPU



- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

Figure from: Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, Jeffrey Mark Siskind; Automatic Differentiation in Machine Learning: a Survey; 2018

Training: How Neural Networks Learn

- Repeat until stopping criterion met:
  1. **Forward pass**: propagate training data through model to make prediction
  2. Quantify the dissatisfaction with a model's results on the training data
  3. **Backward pass**: using predicted output, calculate gradients backward to assign blame to each model parameter
  4. Update each parameter using calculated gradients

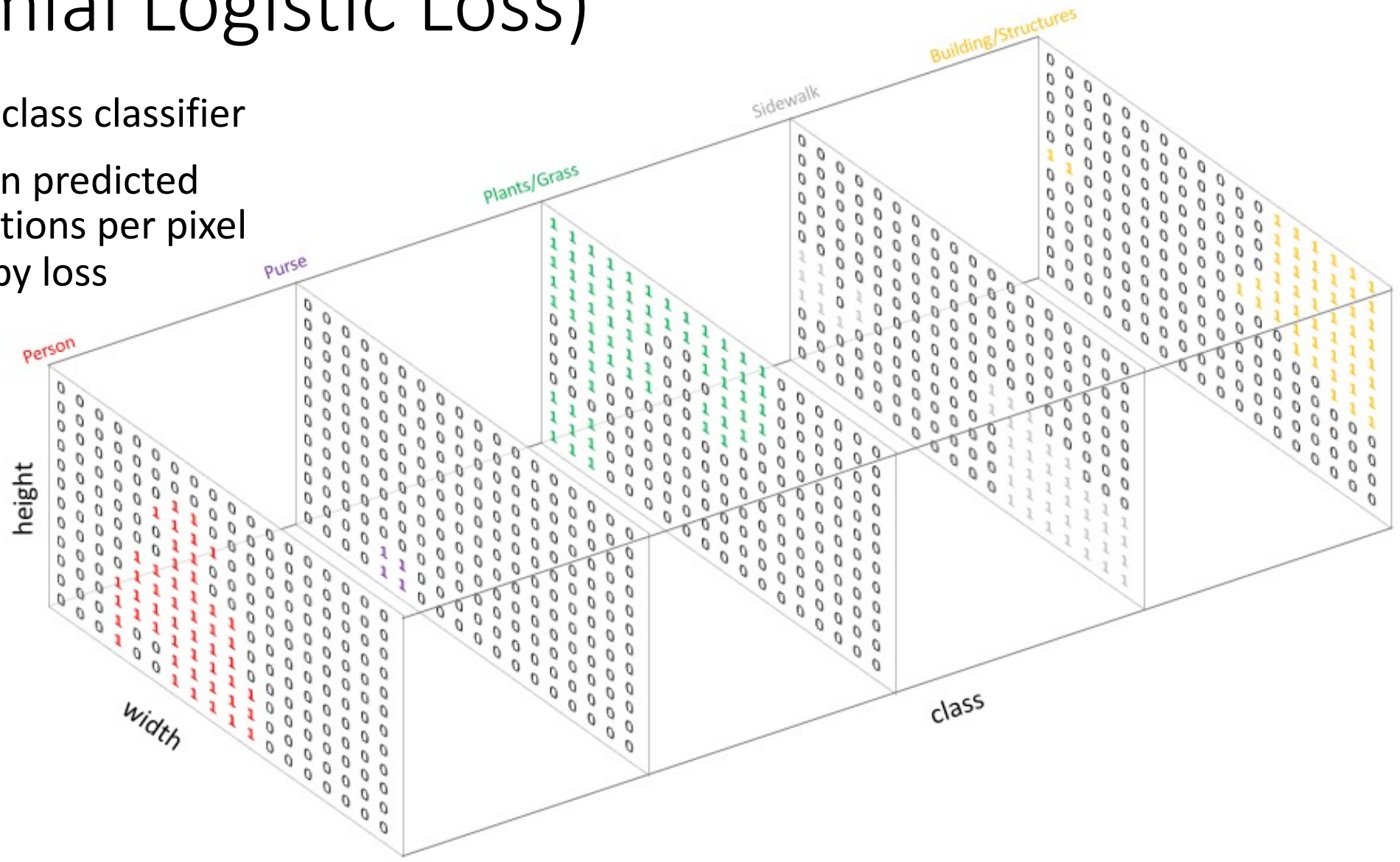
Sum across all pixels the distance between predicted and true distributions using cross entropy loss

Sum of gradients for all pixels (acts like a minibatch)

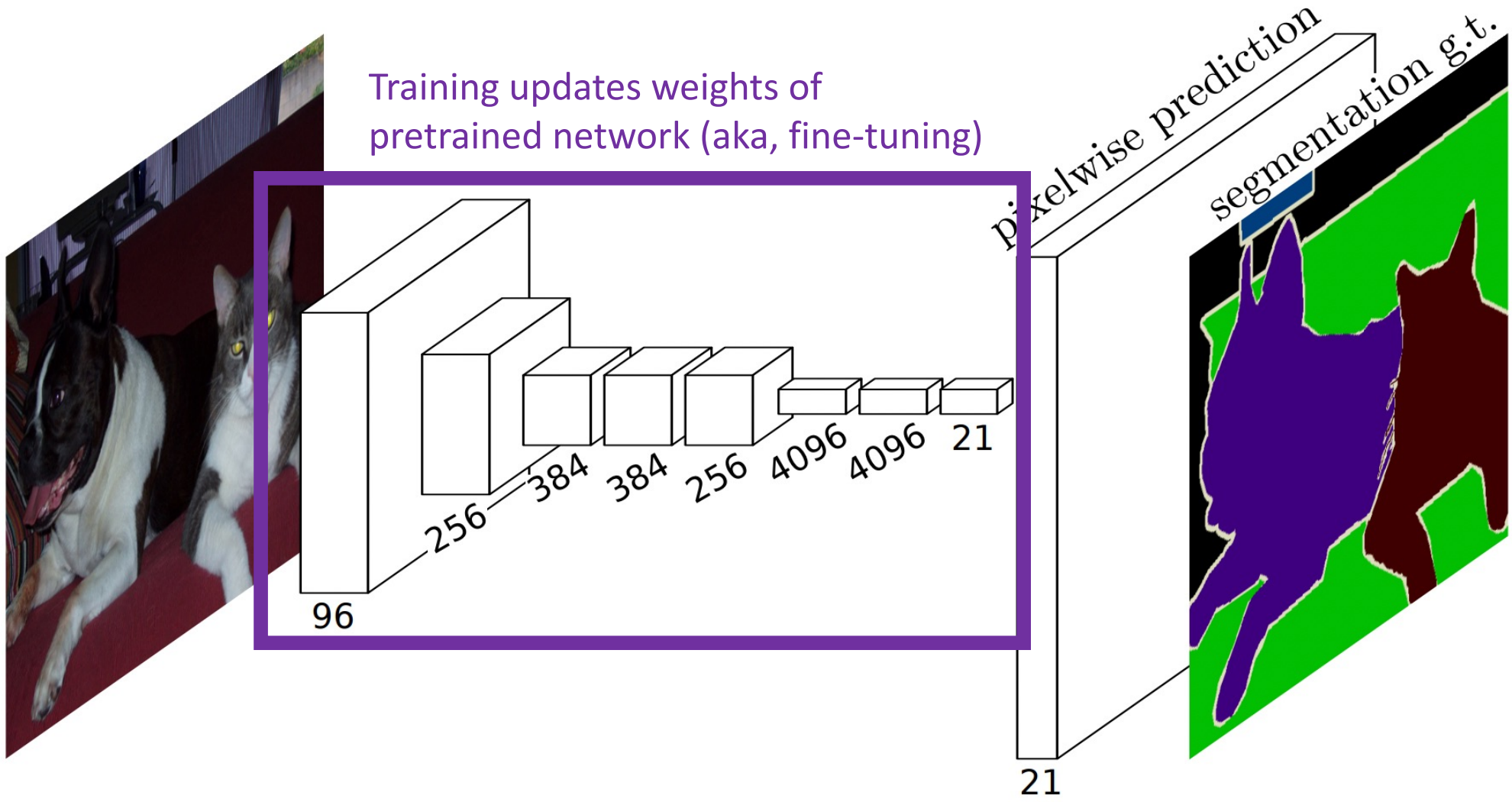
Figure from: Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, Jeffrey Mark Siskind; Automatic Differentiation in Machine Learning: a Survey; 2018

Training: Cross Entropy Loss (Multinomial Logistic Loss)

- e.g., assume a 5-class classifier
- Distance between predicted and true distributions per pixel with cross entropy loss



Architecture: Algorithm Training



Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Results

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [16]	52.6	51.6	$\sim$ 50 s
FCN-8s	**62.7**	**62.2**	$\sim$ **175 ms**

Compared to existing methods, produces better results at a faster speed!

Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Improved Architecture: U-Net

Passes information lost in the encoder to the decoder from **each downsampling layer** in the encoder to its corresponding upsampling layer in the decoder, while also keeping the computation low.

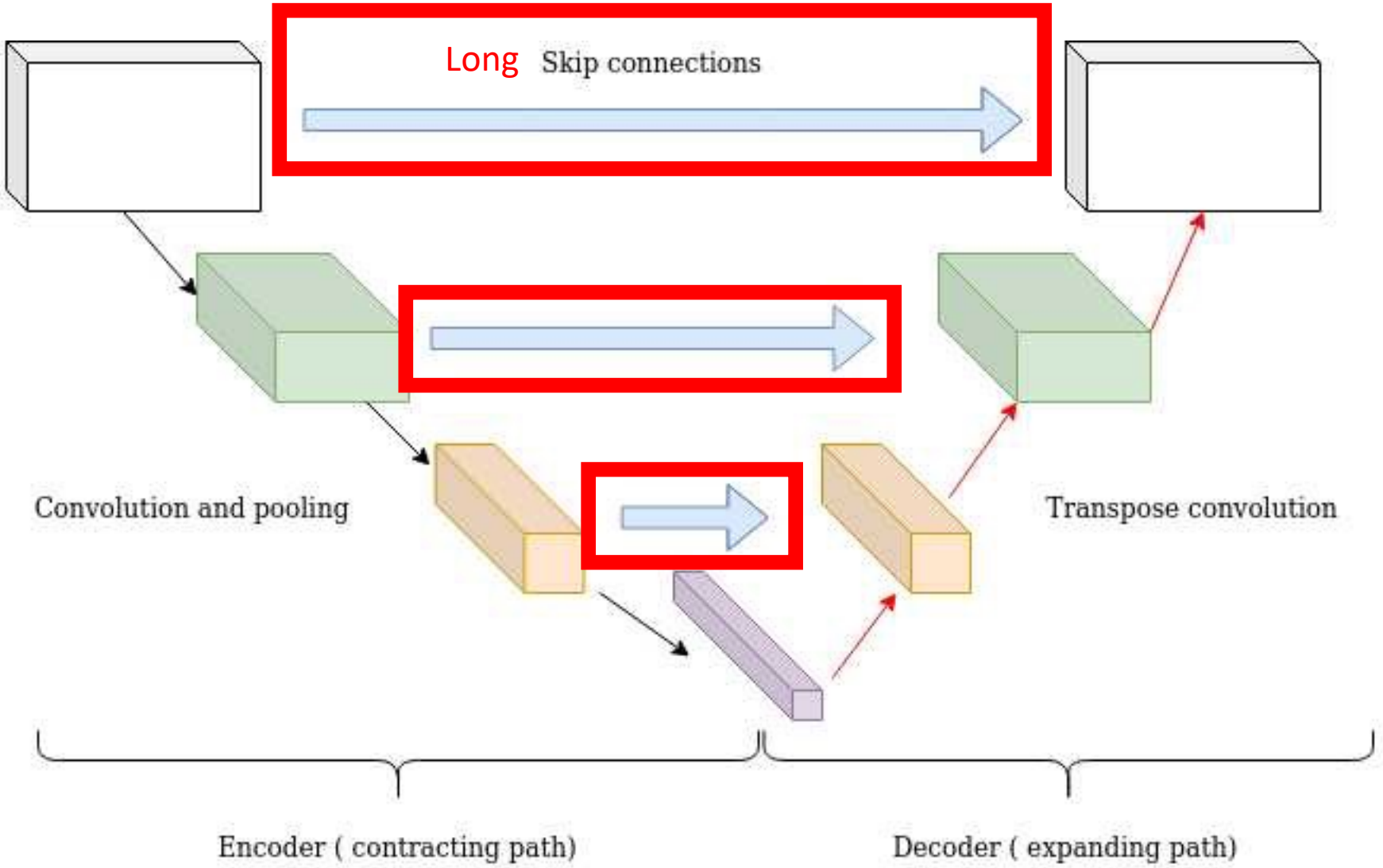
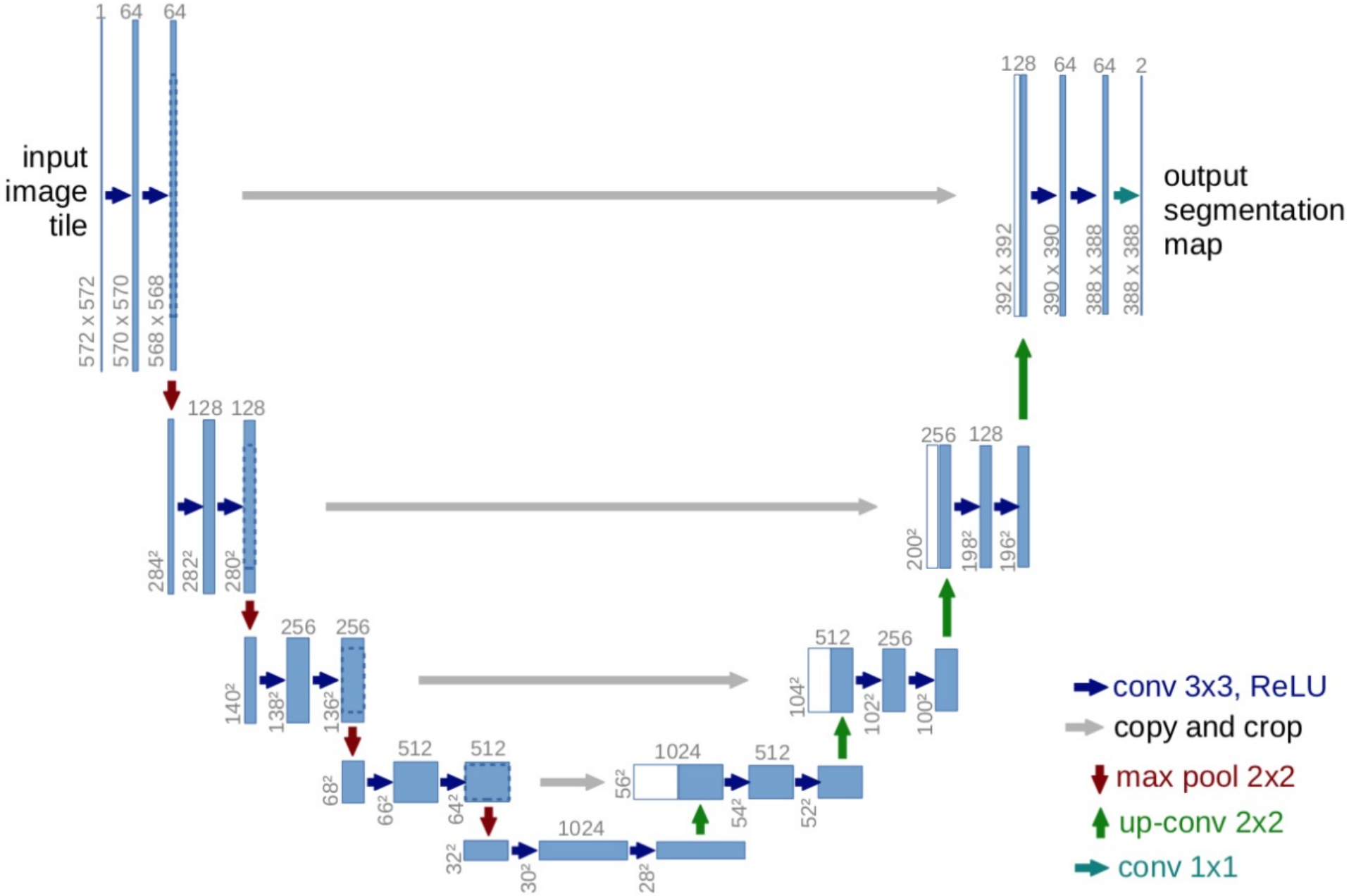


Image Source: https://theaisummer.com/skip-connections/

U-Net



Ronneberger, Fischer, and Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015.

Today's Topics

- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

Today’s Topics

- Problems
- Applications
- PASCAL VOC semantic segmentation challenge
- Fully convolutional networks
- Programming tutorial

The End