Semantic Segmentation

Danna Gurari
University of Colorado Boulder
Fall 2021

CC BY
https://home.cs.colorado.edu/~DrG/Courses/RecentAdvancesInComputerVision/AboutCourse.html

Review

- Last lecture:
  - Overview of object detection algorithms
  - Baseline Model: R-CNN
  - Fast R-CNN
  - Faster R-CNN
  - YOLO

- Assignments (Canvas)
  - Reading assignment due earlier today
  - Reading assignments out that are due tomorrow and next week

- Questions?

Semantic Segmentation: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks

Semantic Segmentation: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks

Definition

- Locate all pixels that belong to a particular category; e.g.,

Note: instances of the same class are NOT separated

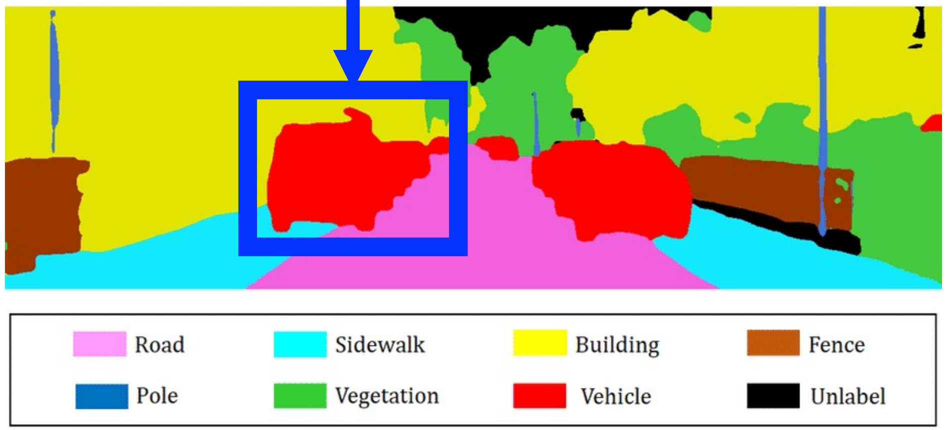


Image Source: Jeong, Yoon, and Park. Sensors 2018.

Object Segmentation vs Detection

- Why choose object “segmentation” over “detection”?

http://mmcheng.net/msra10k/

Semantic Segmentation: Today's Topics

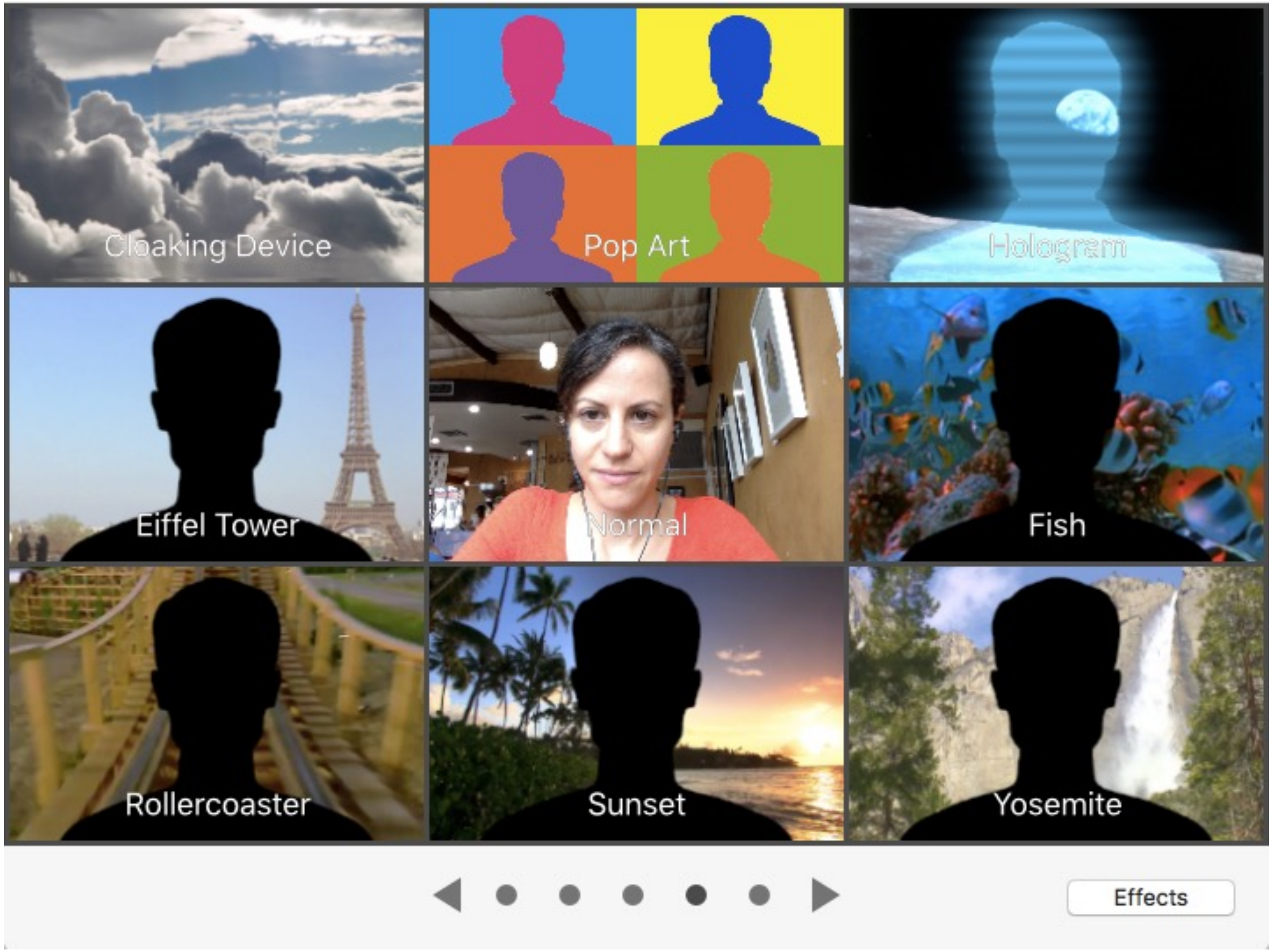
- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks

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

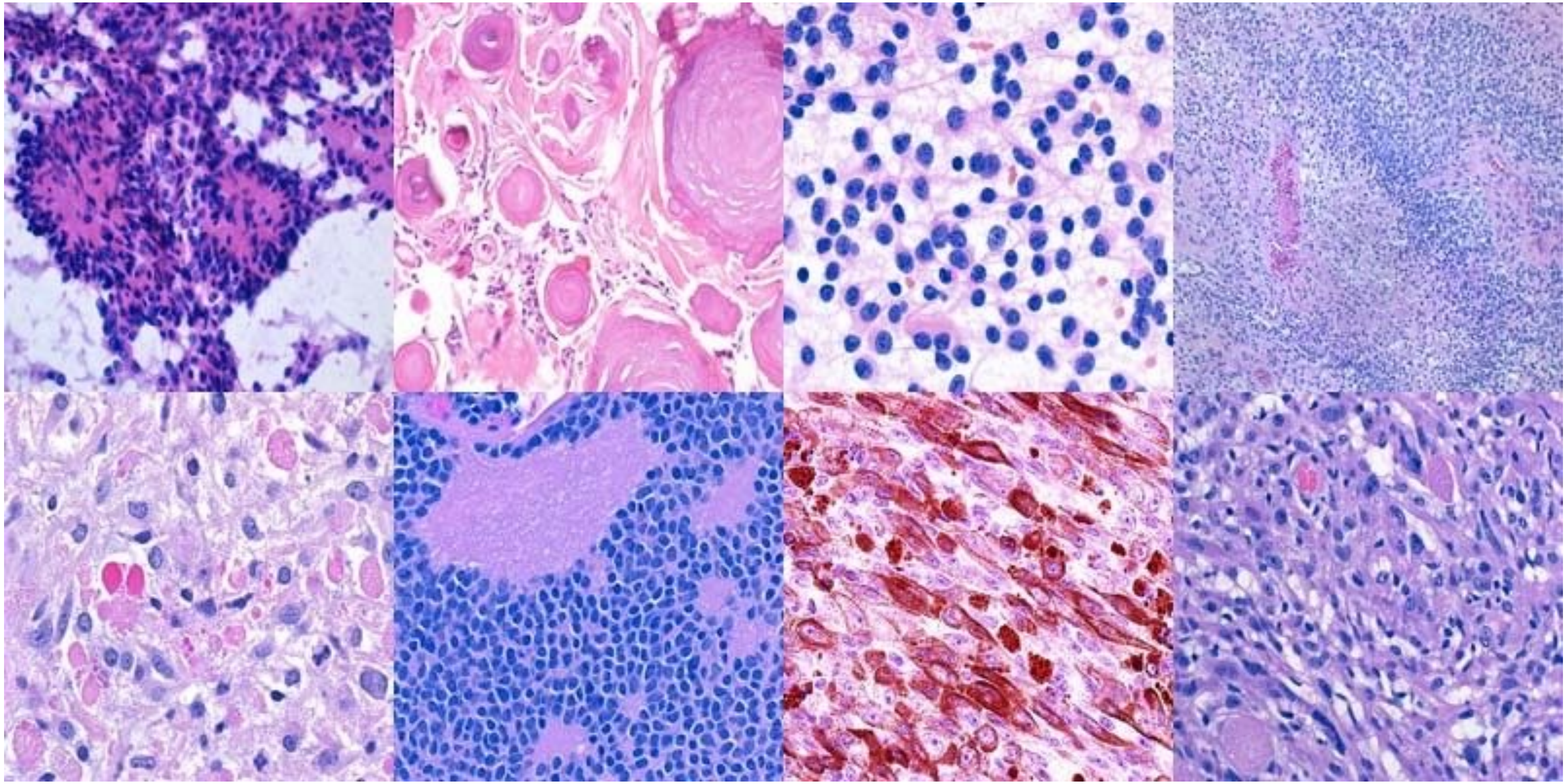
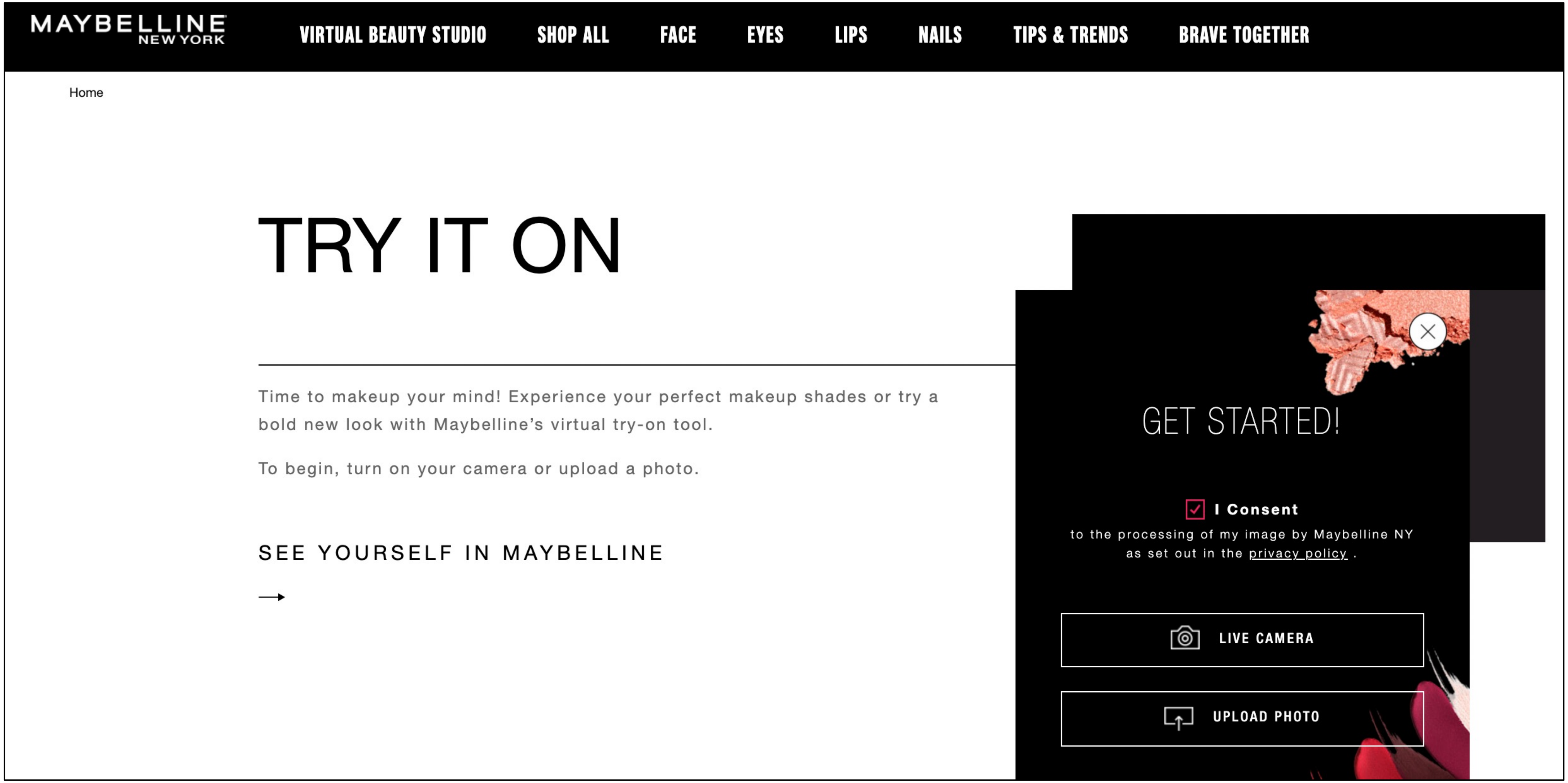


Figure Source: https://pathology.jhu.edu/brain-tumor/grading-classification

Face Makeover



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

Self-Driving Vehicles

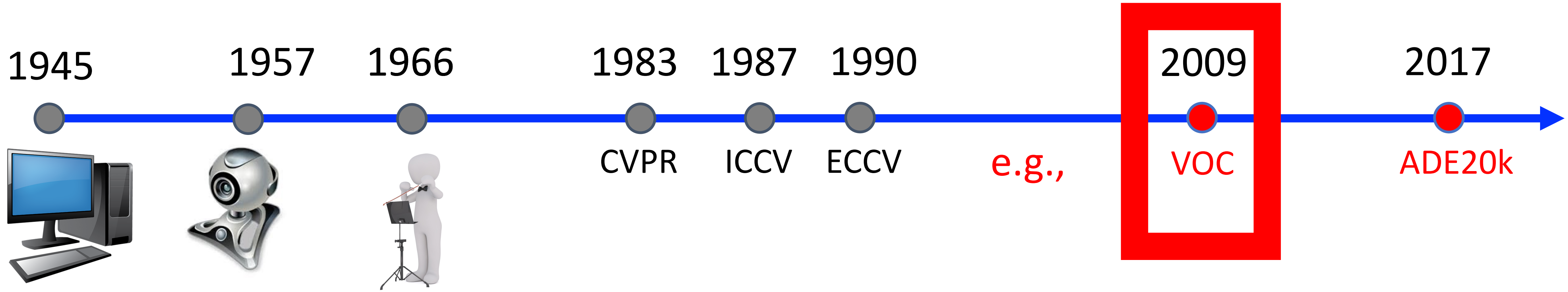
Figure Source: https://www.inc.com/kevin-j-ryan/self-driving-cars-powered-by-people-playing-games-mighty-ai.html

Can you think of any other potential applications?

Semantic Segmentation: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks

Datasets


1945
1957
1966
1983
1987
1990
2009
2017
CVPR
ICCV
ECCV
e.g.,
VOC
ADE20k

VOC

1. Image Collection

- A subset of images from the VOC detection dataset were used

2. Image Annotation

- Annotation party annually

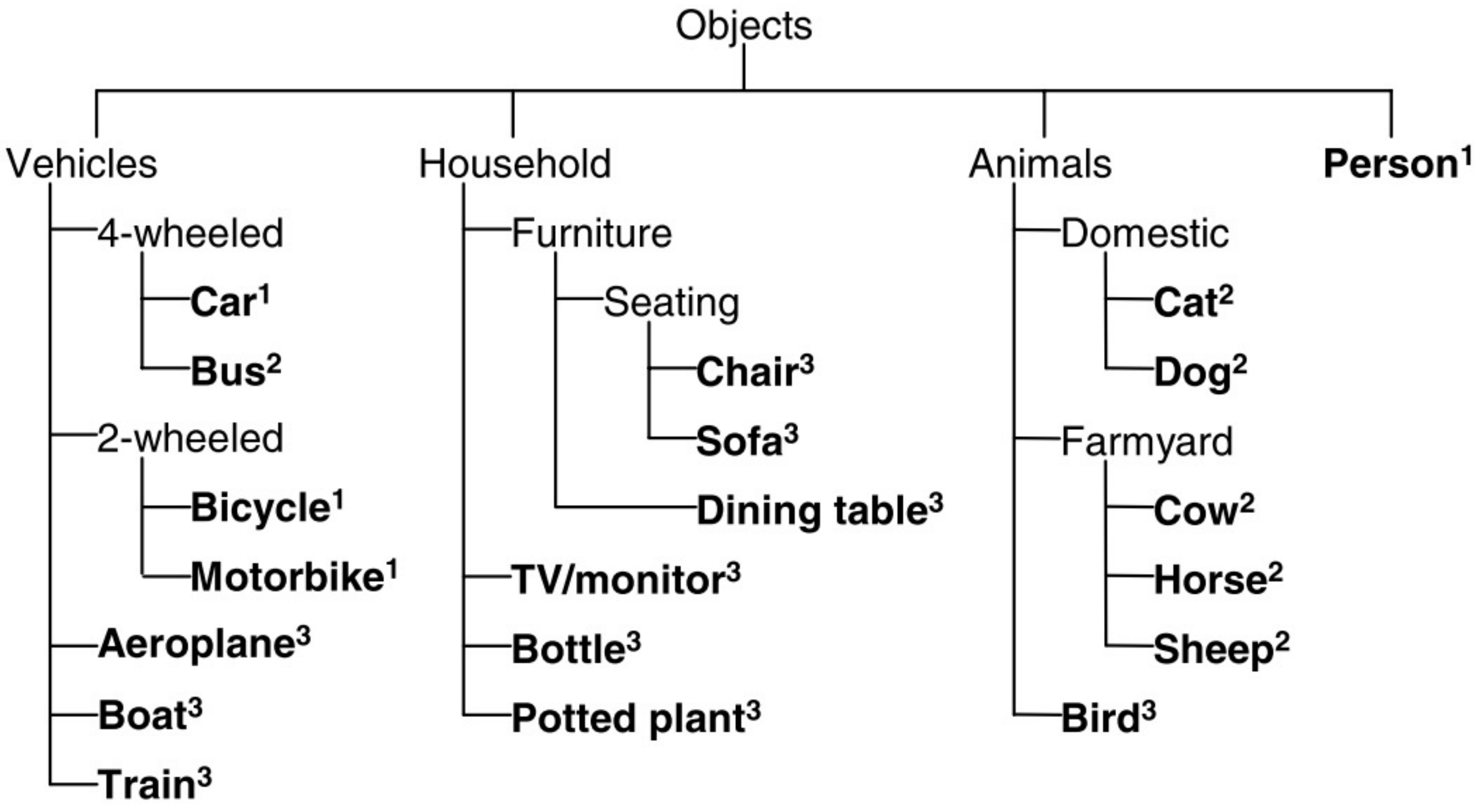
- Annotation guidelines & real-time assistance – refine detections into segmentations

- Post-hoc correction/feedback about the number and kind of errors made

- Annotations for each of the 20 object classes were merged into class-specific segmentation regions and 1 more class was added for background

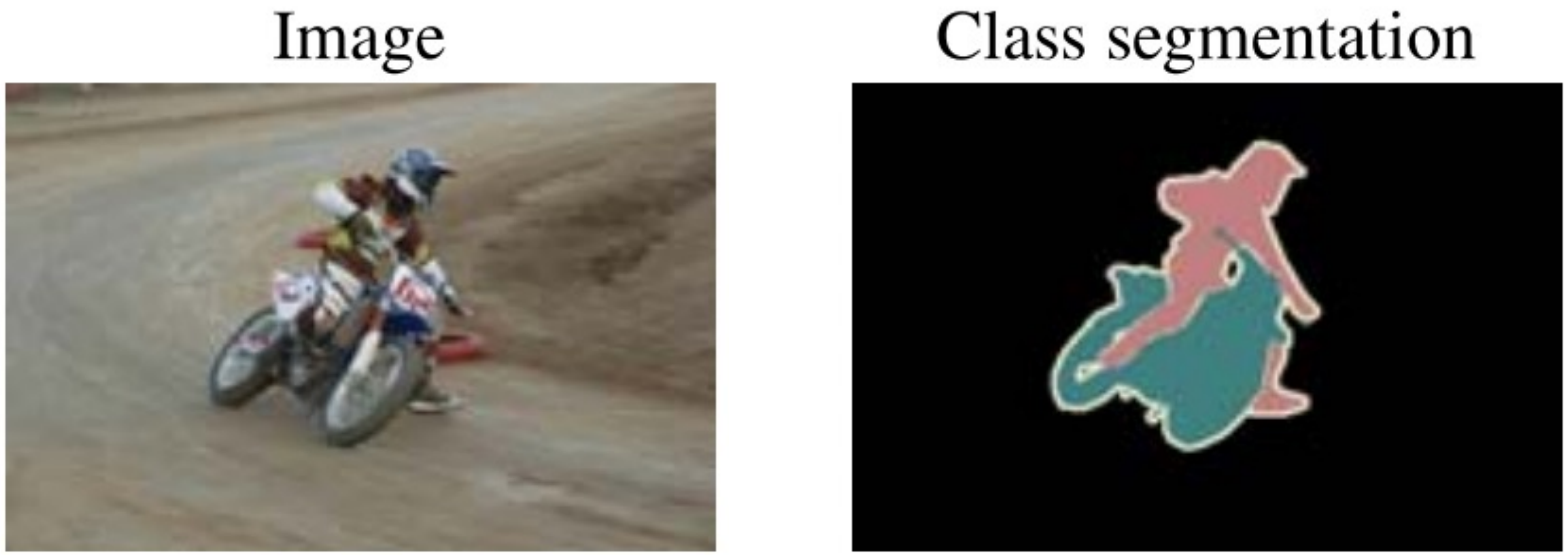
Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

VOC: Recall Categories Included (Leaf Nodes)



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

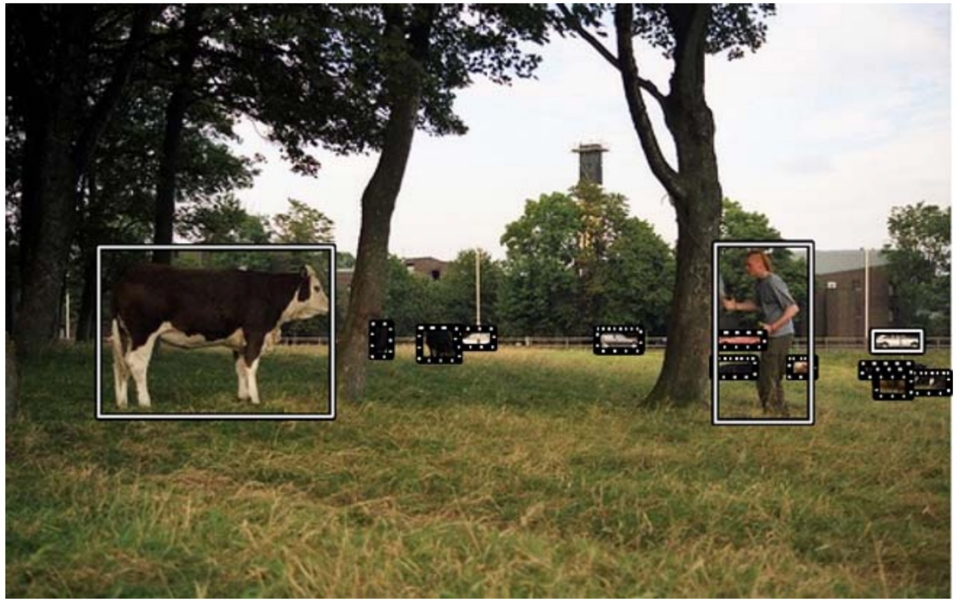
VOC: Boundary Accuracy Heuristic



"To give high accuracy but to keep the annotation time short enough to provide a large image set, a border area of 5 pixels width was allowed around each object where the pixels were labelled neither object nor background."

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

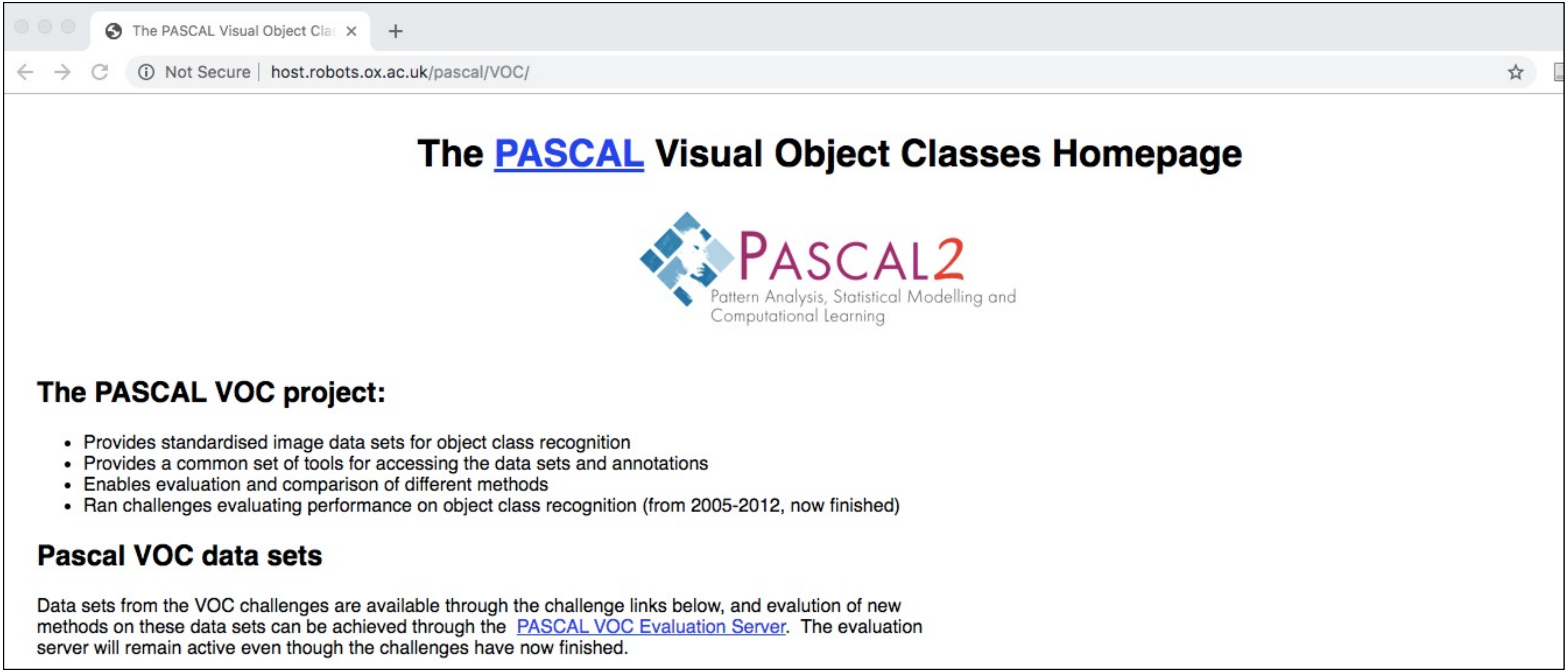
VOC: “Difficult” Objects Excluded



Objects that are challenging to recognize are discarded (i.e., dashed regions): flagged for reasons of “small size, illumination, image quality or the need to use significant contextual information… no penalty is incurred for detecting them. The aim of this annotation is to maintain a reasonable level of difficulty…”

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

VOC Annual Workshop


The PASCAL Visual Object Cla
Not Secure | host.robots.ox.ac.uk/pascal/VOC/

The PASCAL Visual Object Classes Homepage

The PASCAL VOC project:

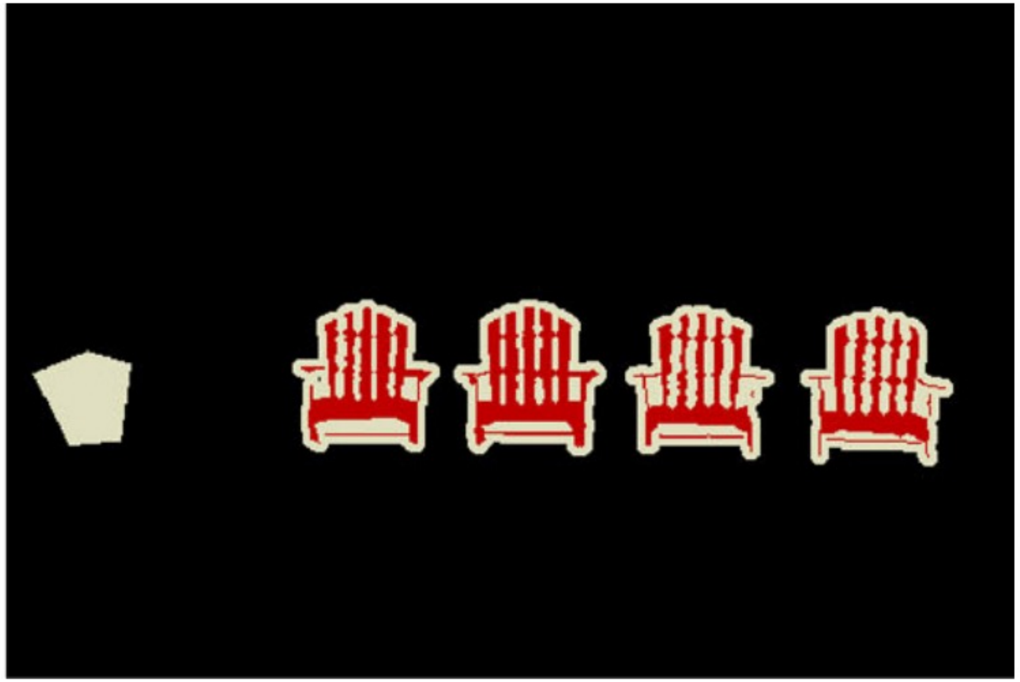
- Provides standardised image data sets for object class recognition
- Provides a common set of tools for accessing the data sets and annotations
- Enables evaluation and comparison of different methods
- Ran challenges evaluating performance on object class recognition (from 2005-2012, now finished)

Pascal VOC data sets

Data sets from the VOC challenges are available through the challenge links below, and evalution of new methods on these data sets can be achieved through the PASCAL VOC Evaluation Server. The evaluation server will remain active even though the challenges have now finished.

http://host.robots.ox.ac.uk/pascal/VOC/

What is a Limitation of Datasets Built Around Specific Categories (e.g., Objects)?

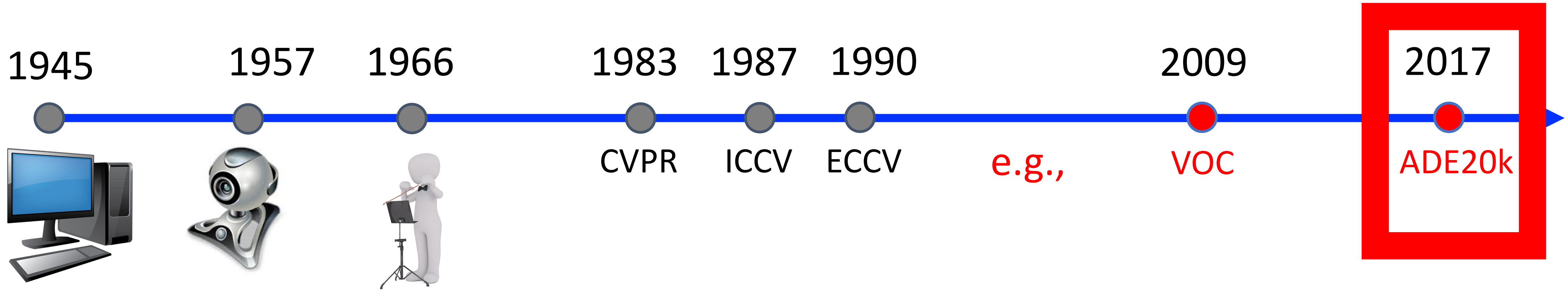


No knowledge that anything else is in the scene, such as a house, trees or flowers!

A further consequence is that the majority of pixels are labeled as `background'.

Mark Everingham et al. The PASCAL Visual Object Classes Challenge: A Retrospective. IJCV 2015.

Datasets


1945
1957
1966
1983
1987
1990
2009
2017
CVPR
ICCV
ECCV
e.g.,
VOC
ADE20k

ADE20K

1. Image Collection

- 25,210 images collected from existing datasets (SUN, Places, and LabelMe)

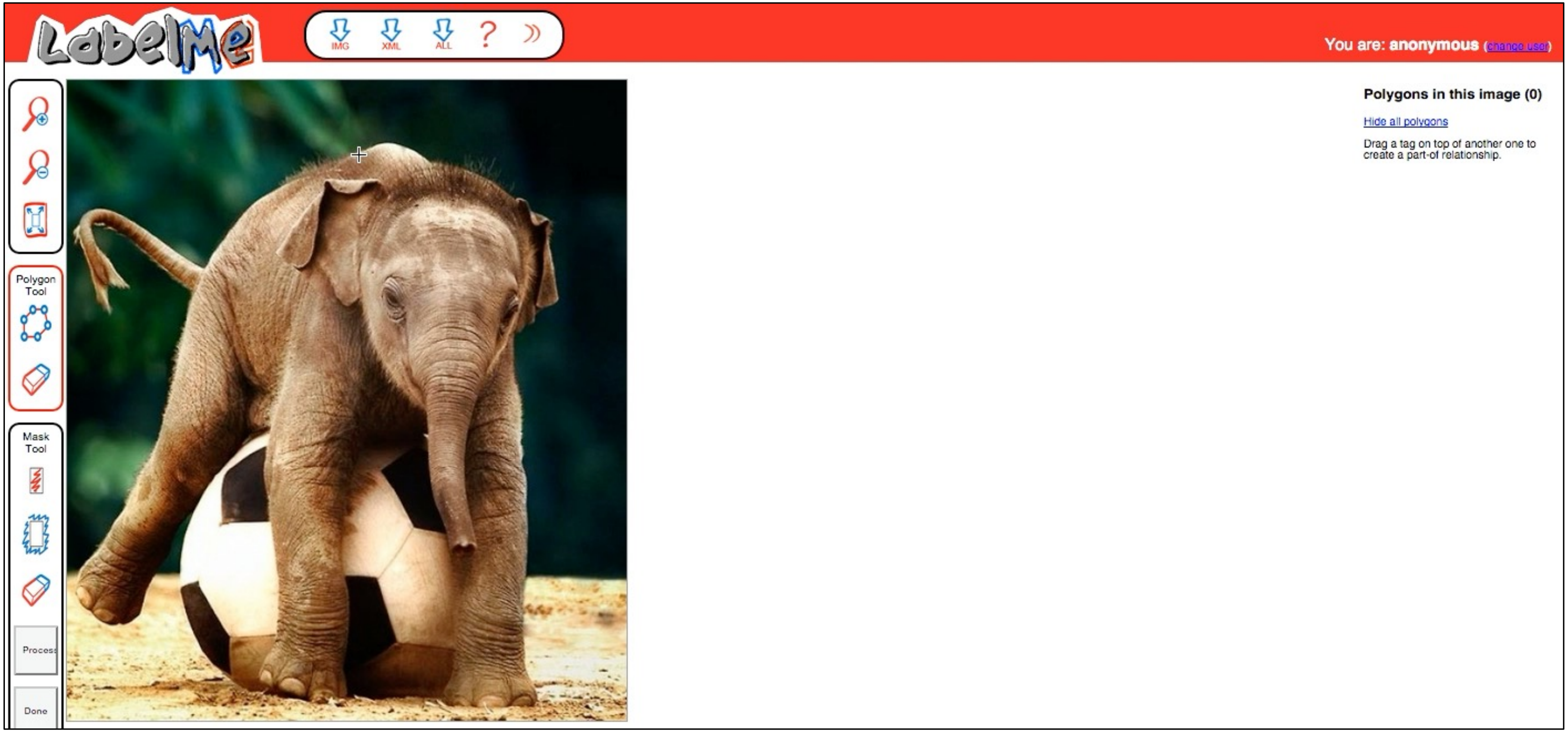
- Selected to capture all scene categories defined in SUN

→

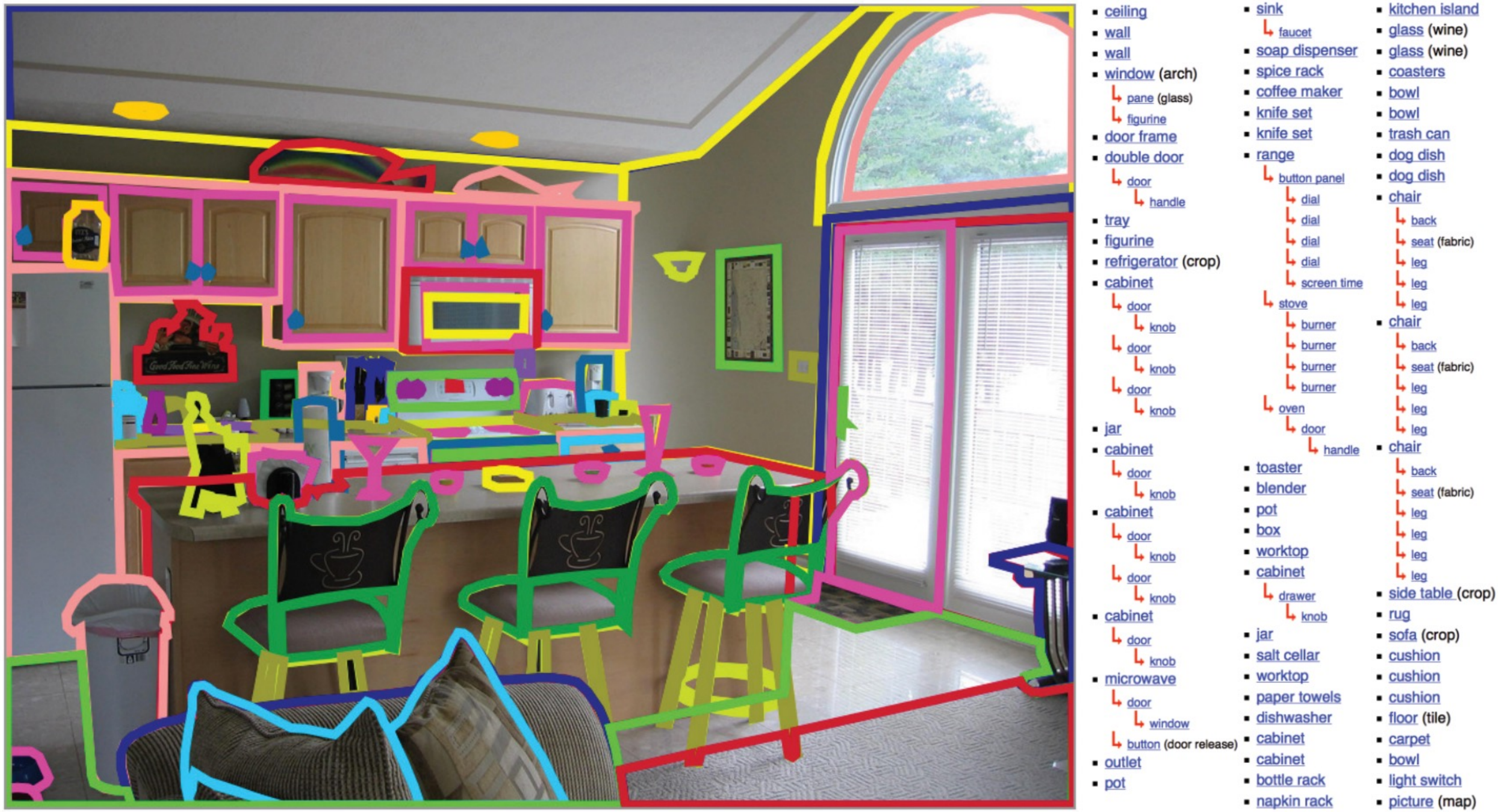
2. Region Localization and Category Assignment

- A single person annotated all images into three types and kept adding new categories as they were observed: (1) objects, (2) object parts, and (3) attributes (e.g., occluded)

ADE20K: User Annotation Tool



ADE20K: User Annotation Tool



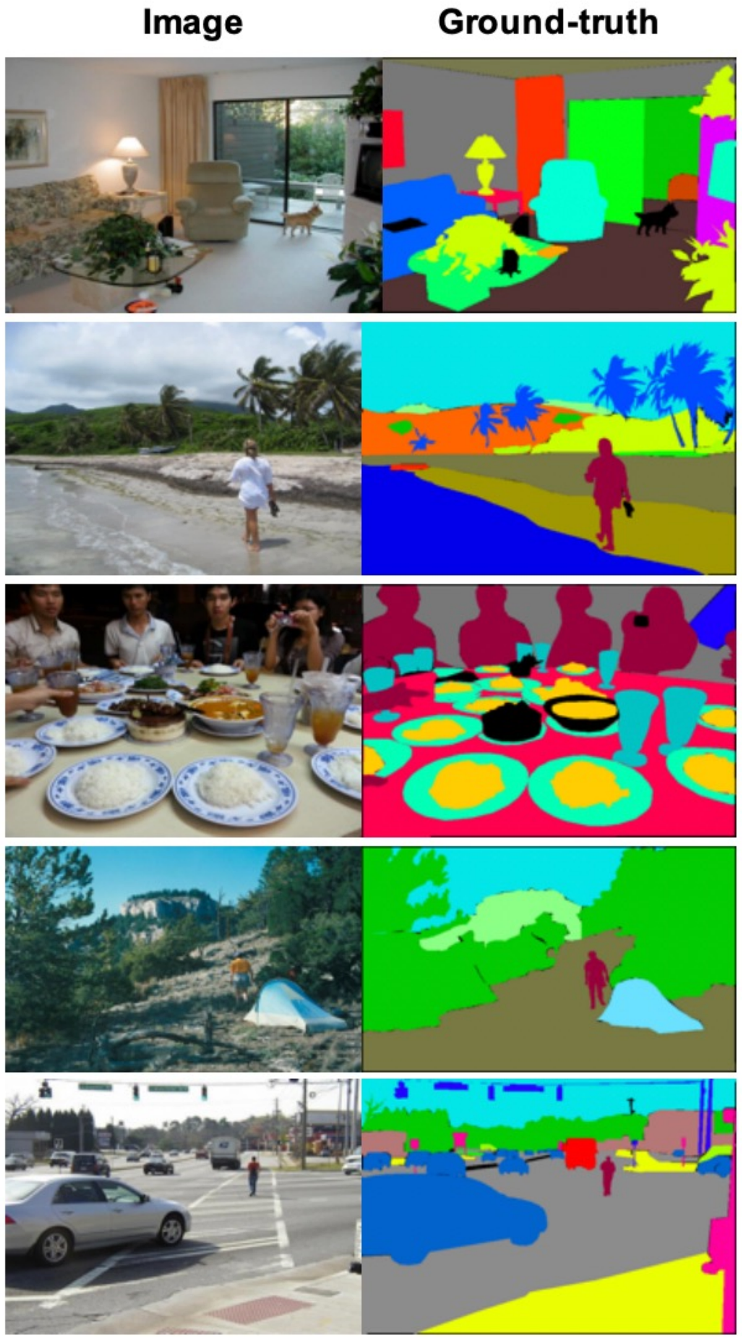
Bolei Zhou et al. Scene Parsing through ADE20K Dataset. CVPR 2017.

ADE20K

- Includes:

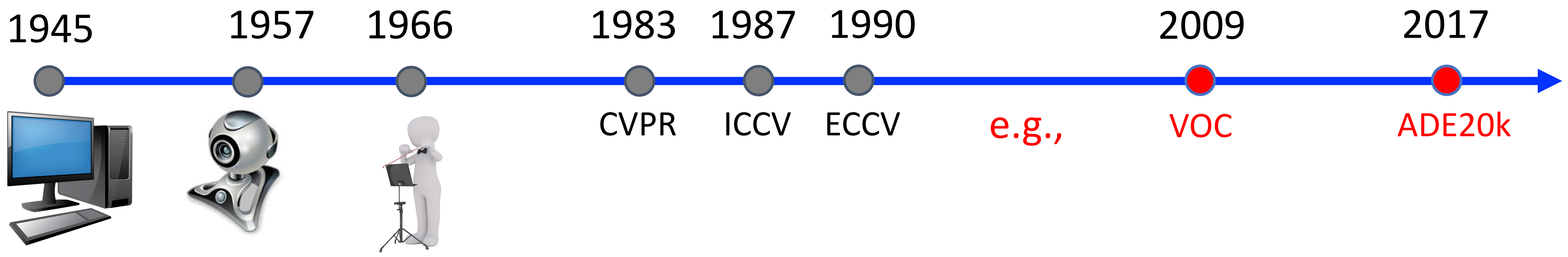
- “things”: objects that can easily be labeled; e.g., person, chair

- “stuff”: objects with no clear boundaries; e.g., sky, grass



Bolei Zhou et al. Scene Parsing through ADE20K Dataset. CVPR 2017.

Datasets



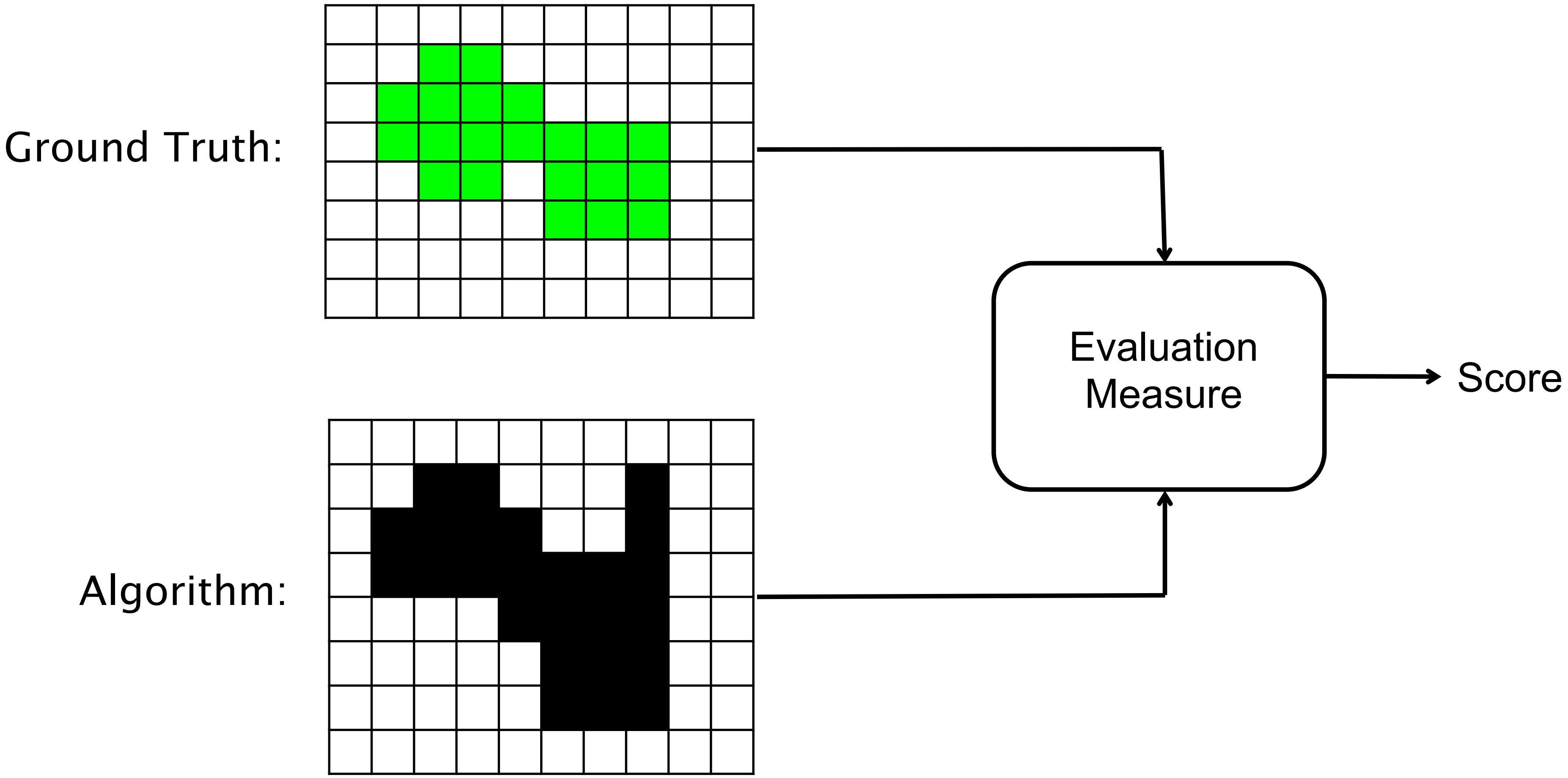
	VOC	ADE20k
# Categories:	21	3169
# Images:	1112 train/val	25,210

Trend: build bigger datasets

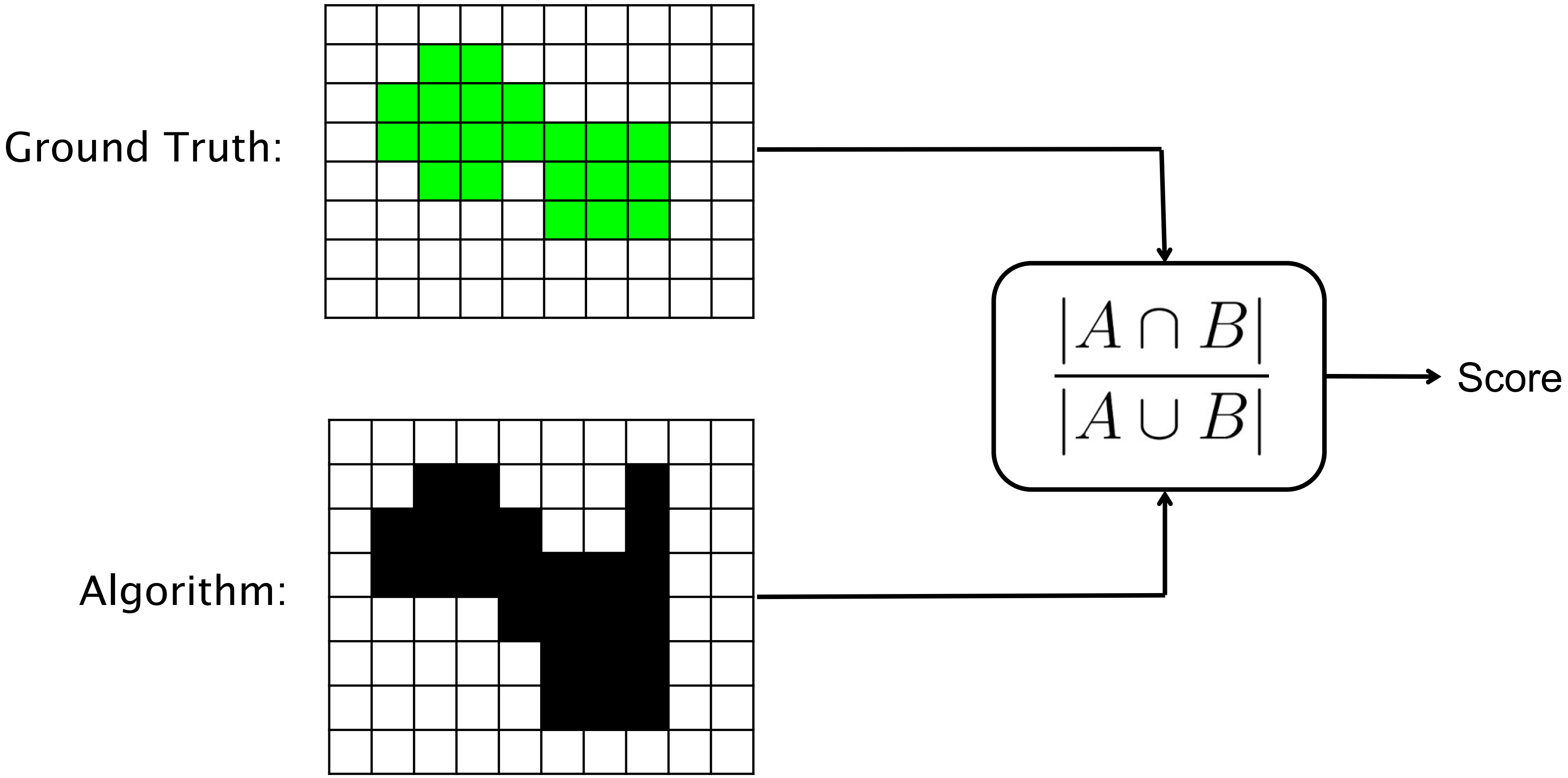
Semantic Segmentation: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks

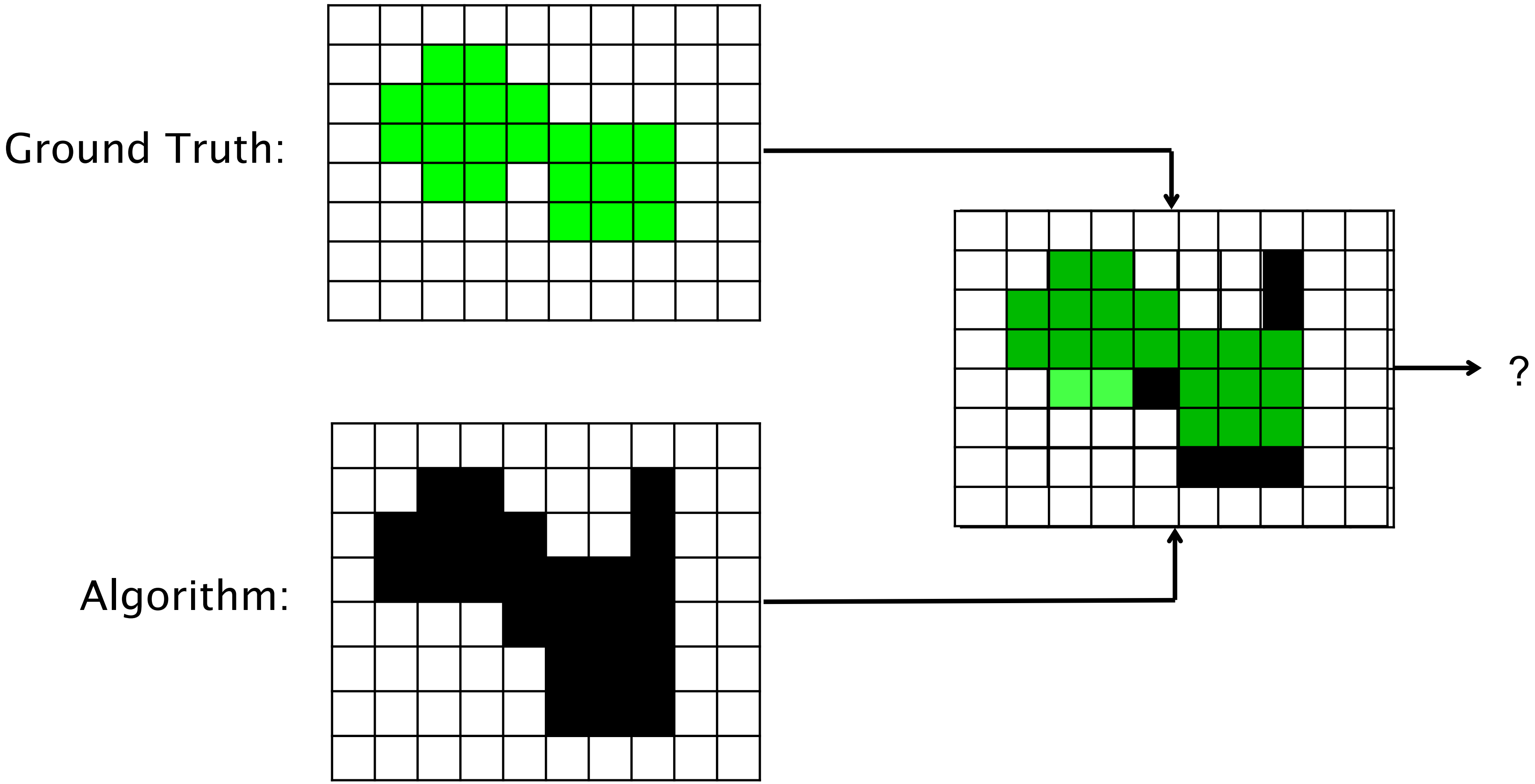
Evaluation Metric



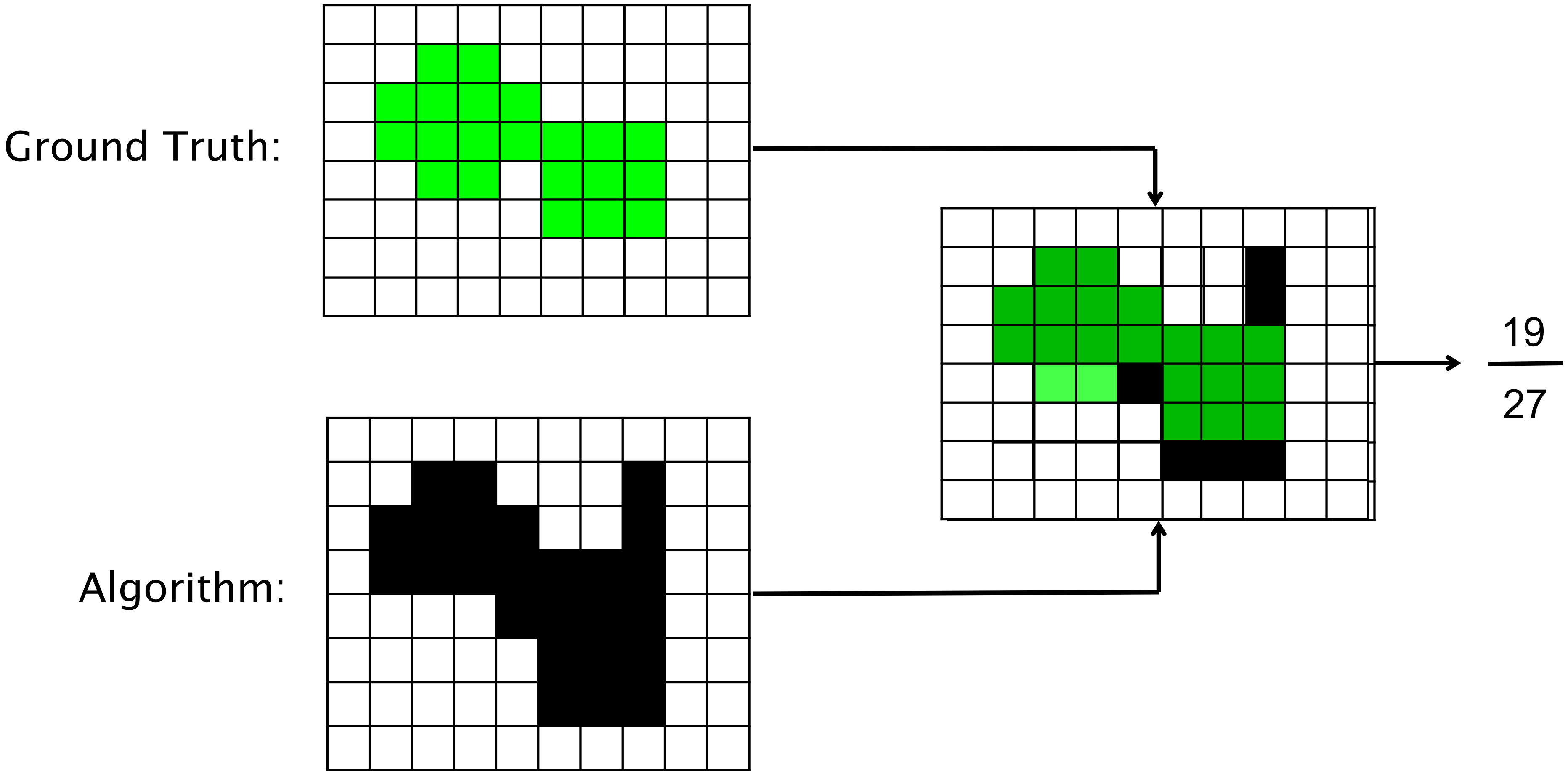
Recall: IoU Metric



Recall: IoU Metric



Recall: IoU Metric



Semantic Segmentation

- **Mean IoU**: IoU between predicted and ground-truth pixels, averaged over all categories
- **Weighted IoU**: IoU weighted by the total pixel ratio of each category
- **Pixel accuracy**: proportion of correctly classified pixels
- **Mean accuracy**: proportion of correctly classified pixels, averaged over all categories

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, & Antonio Torralba. Scene Parsing through ADE20K Dataset. ICCV 2017.

Semantic Segmentation: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks

Why Fully Convolutional Network?

Named after the proposed technique that excludes fully connected layers:

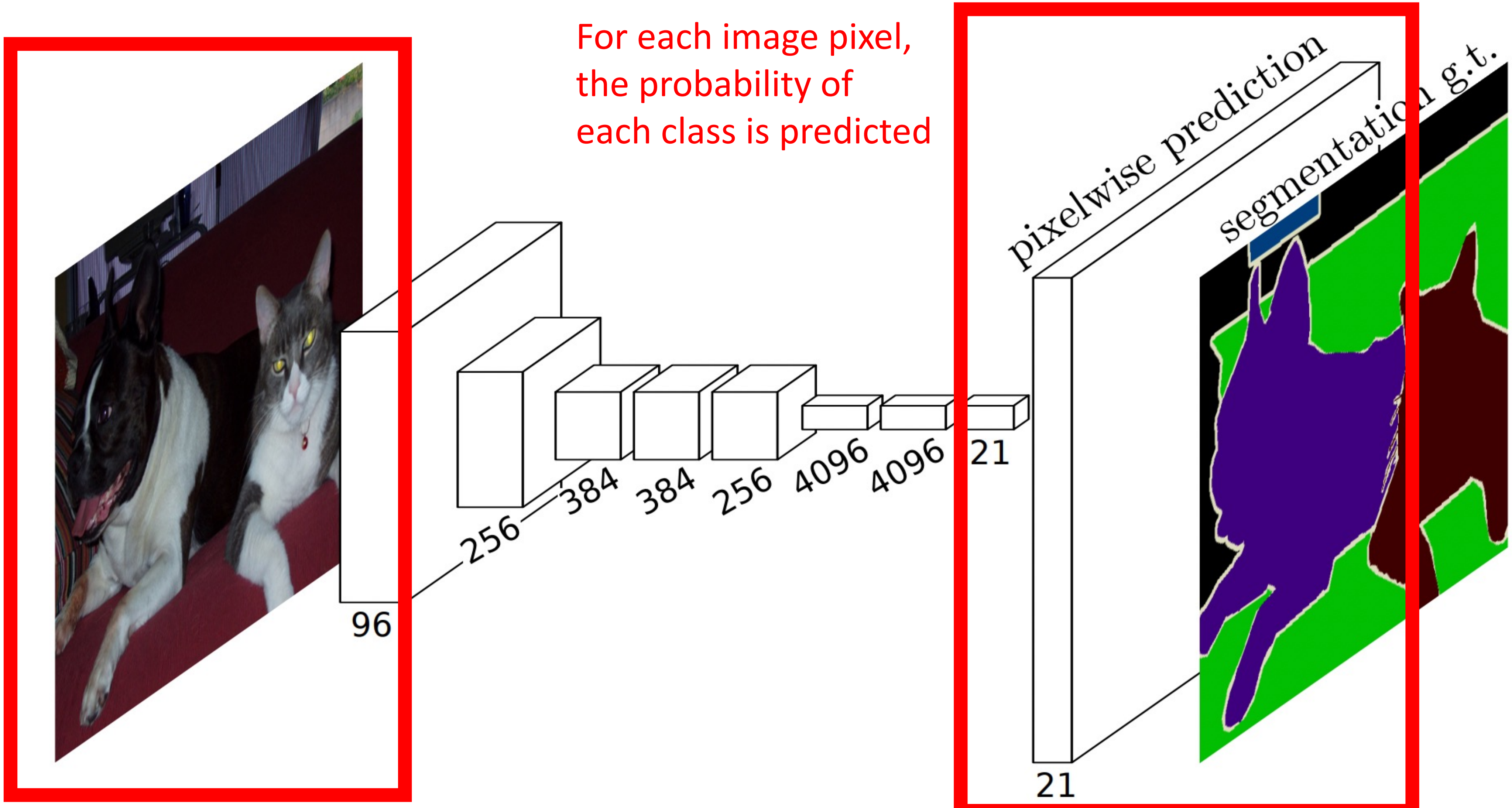
Jonathon Long, Evan Shelhamer, and Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation.” CVPR 2015.

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



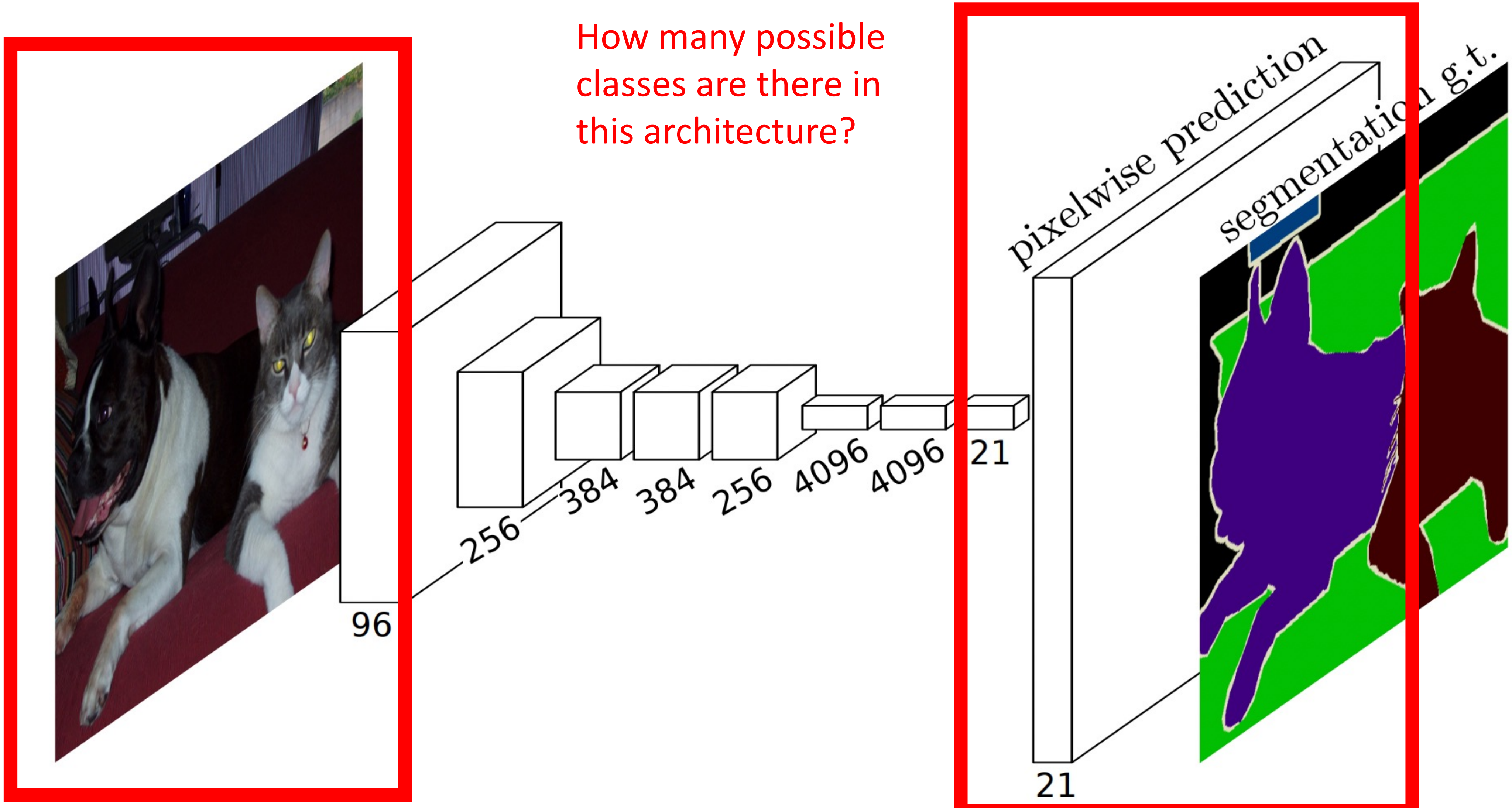
Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture

Input: RGB image of ANY size

Output: Image of same size as input

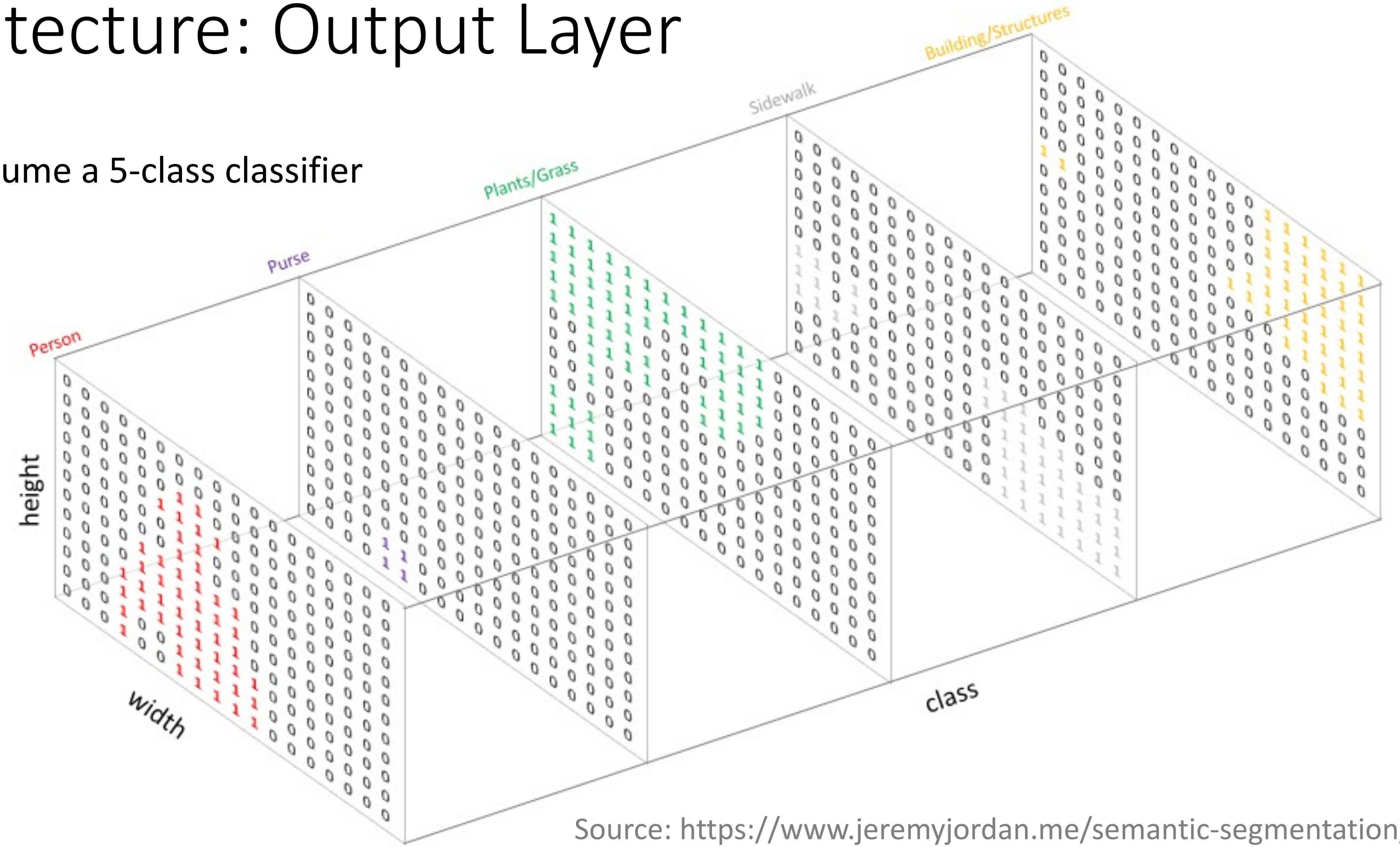
How many possible classes are there in this architecture?



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

Architecture: Output Layer

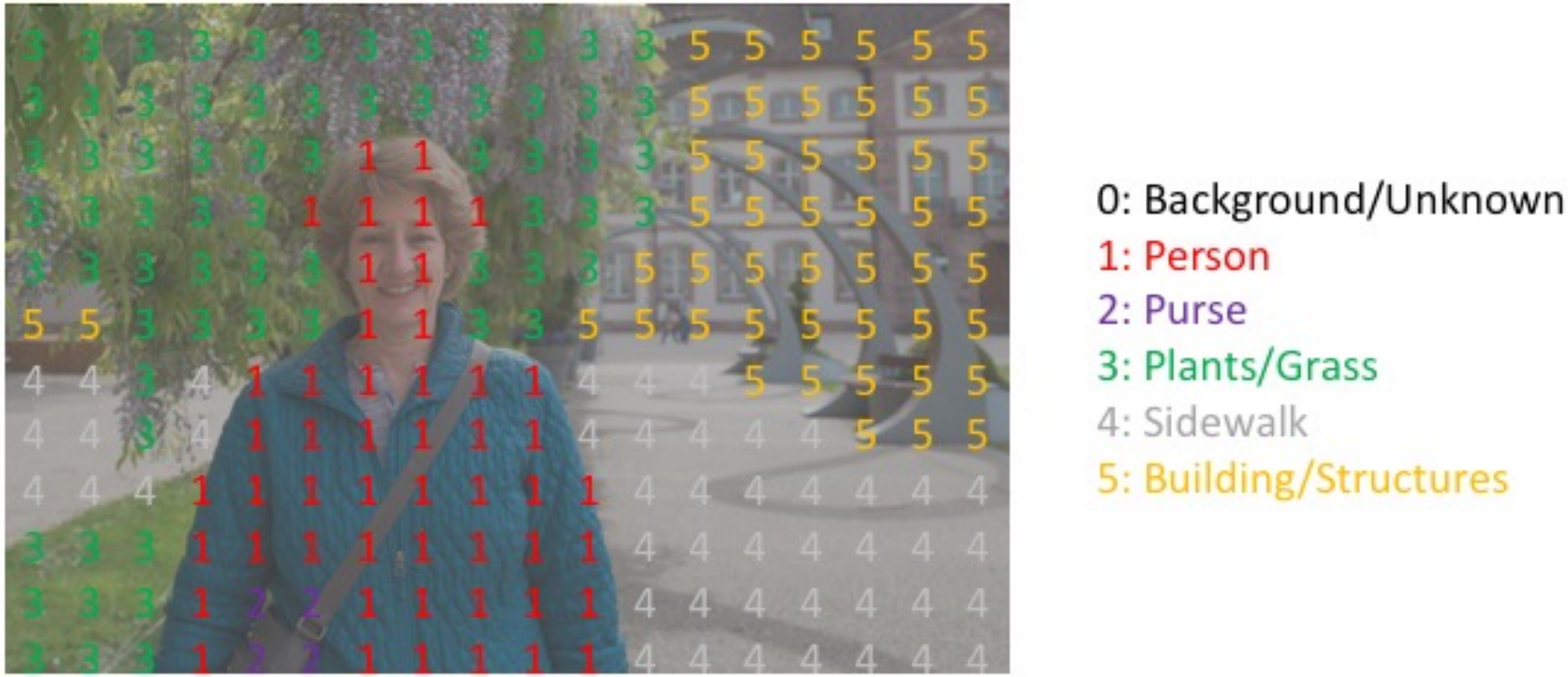
- e.g., assume a 5-class classifier



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

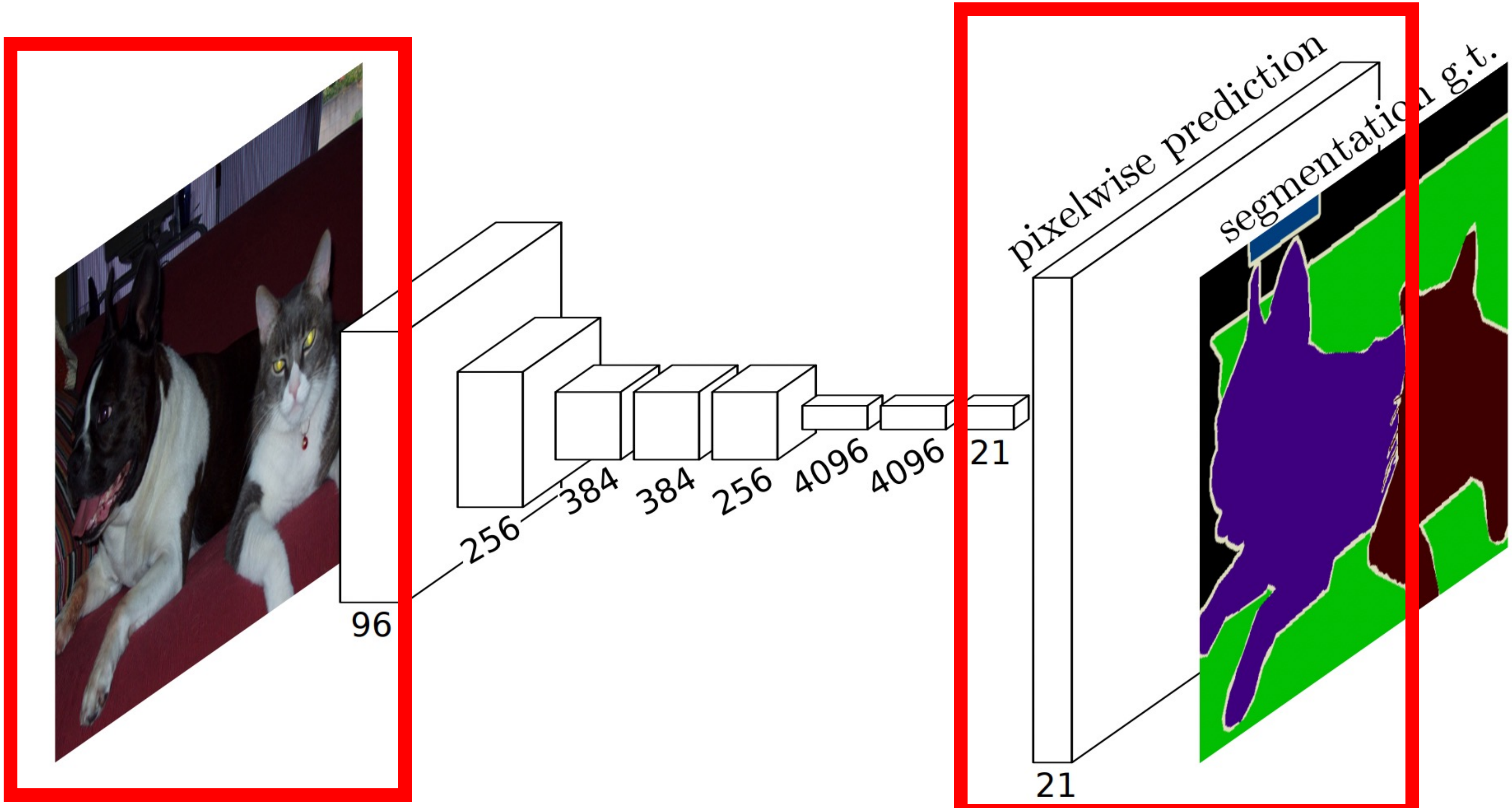


Source: https://www.jeremyjordan.me/semantic-segmentation/

Architecture

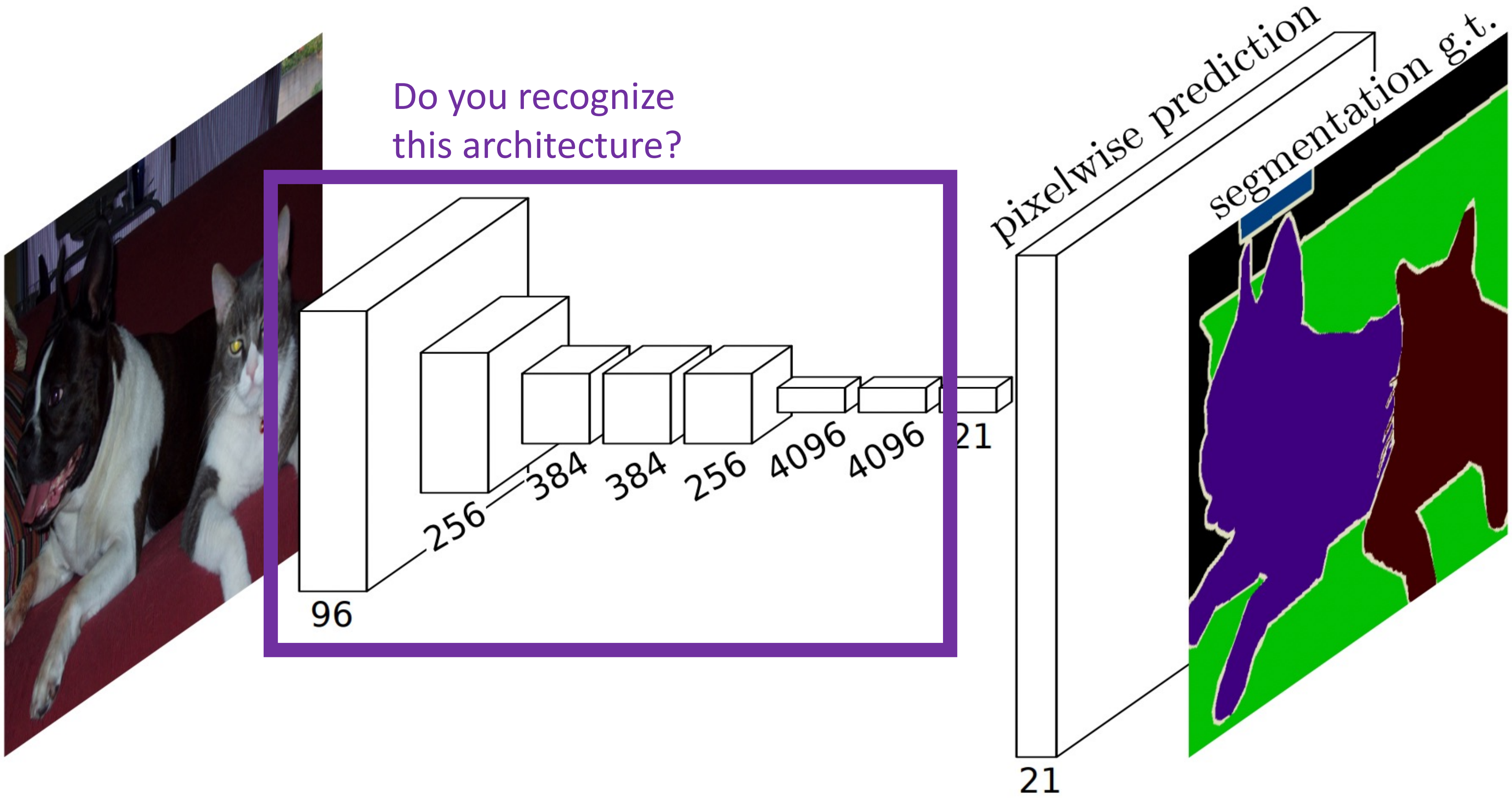
Input: RGB image of ANY size

Output: Image of same size as input



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

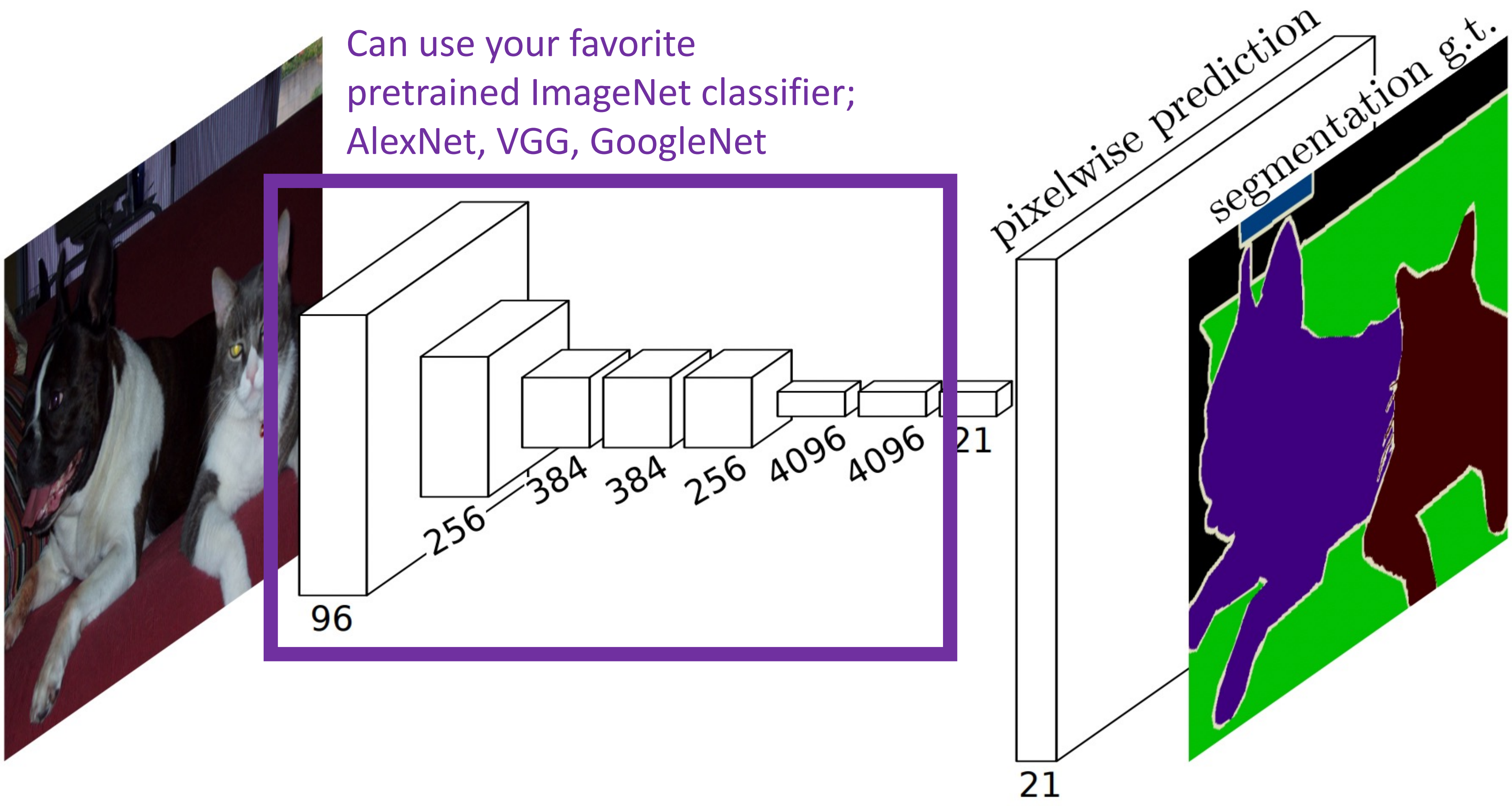
Architecture



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

Architecture

Can use your favorite
pretrained ImageNet classifier;
AlexNet, VGG, GoogleNet

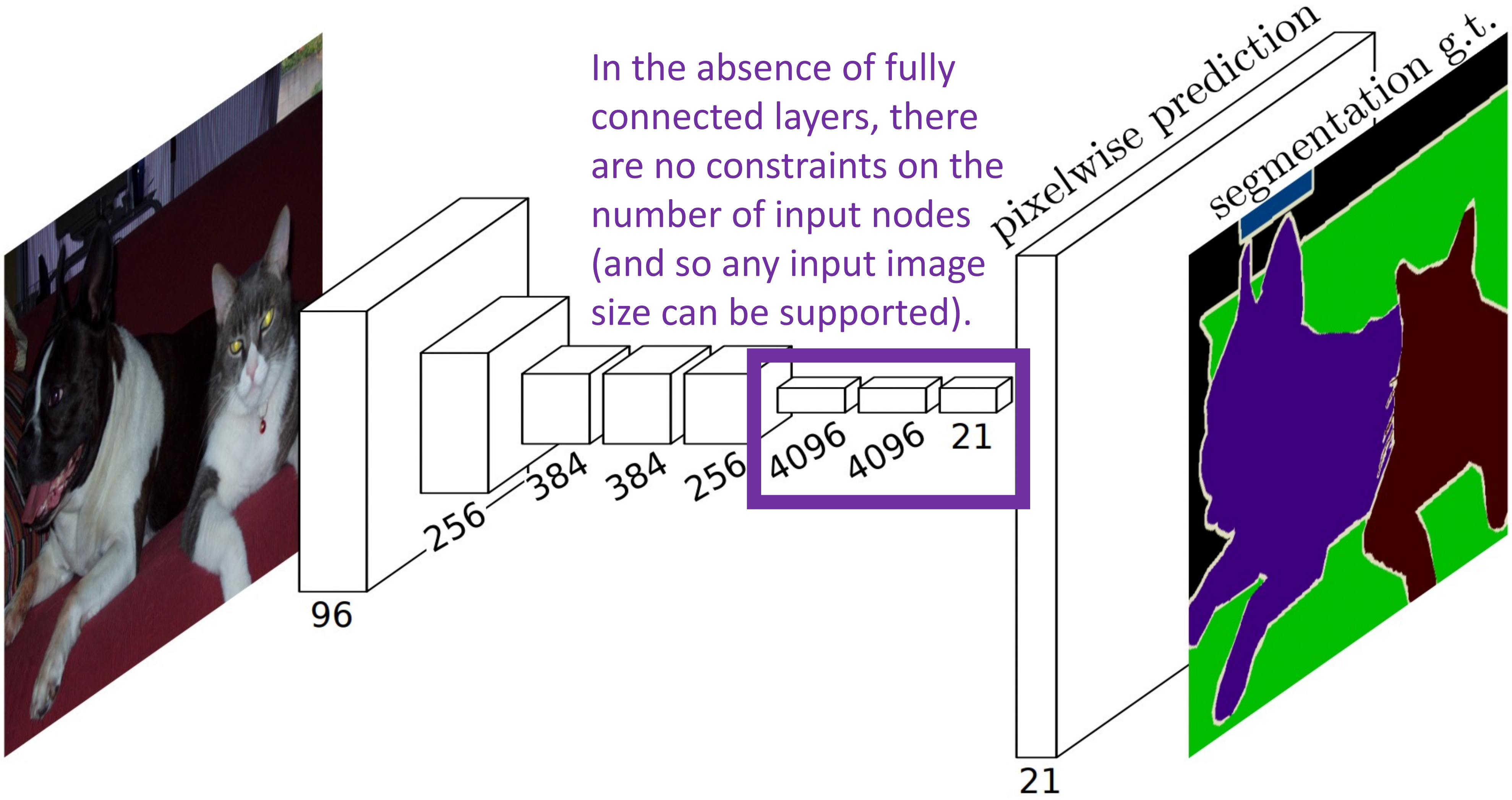


Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture

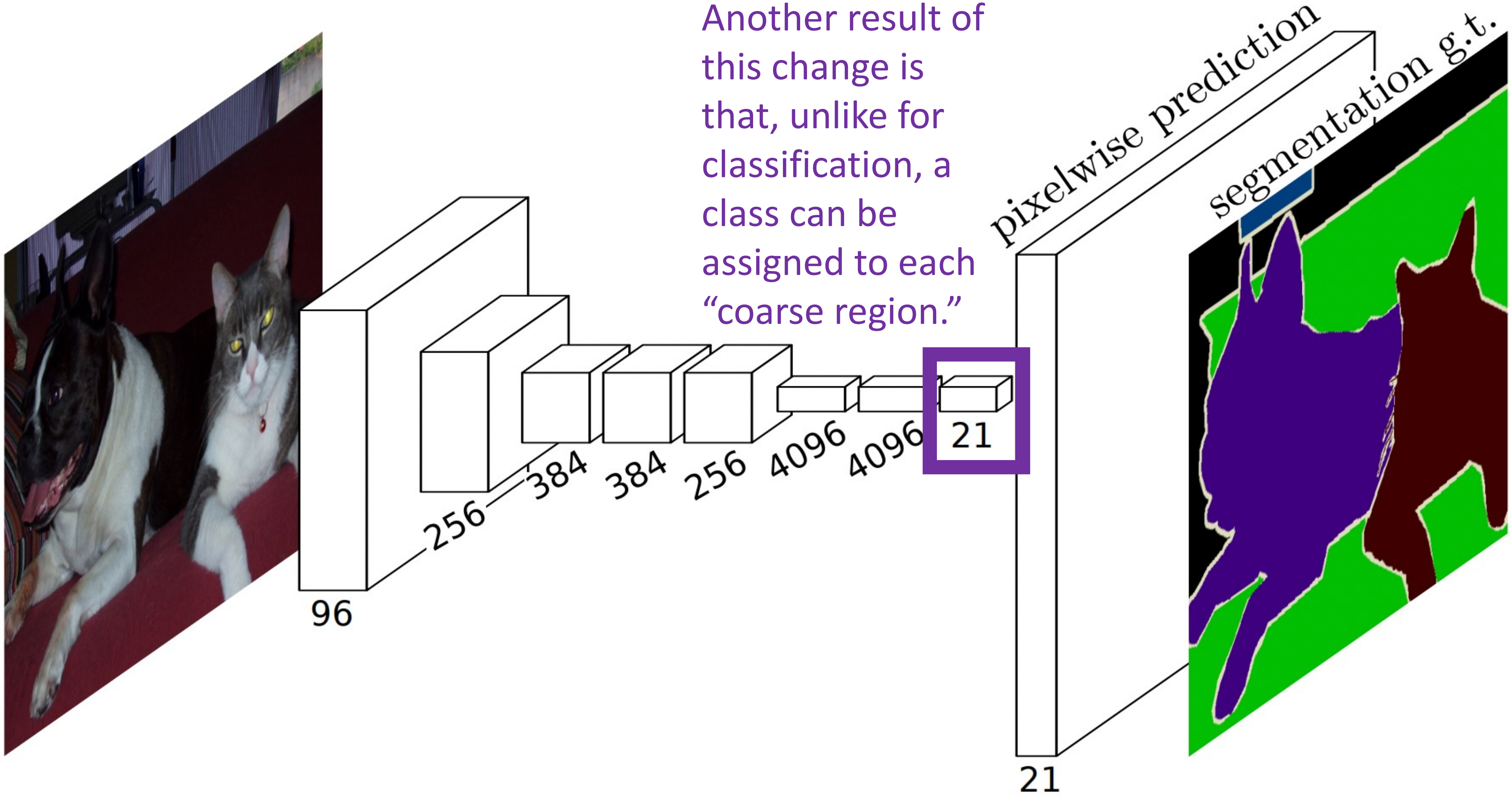
To make the architecture fully convolutional, fully connected layers are converted to convolutional layers.

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

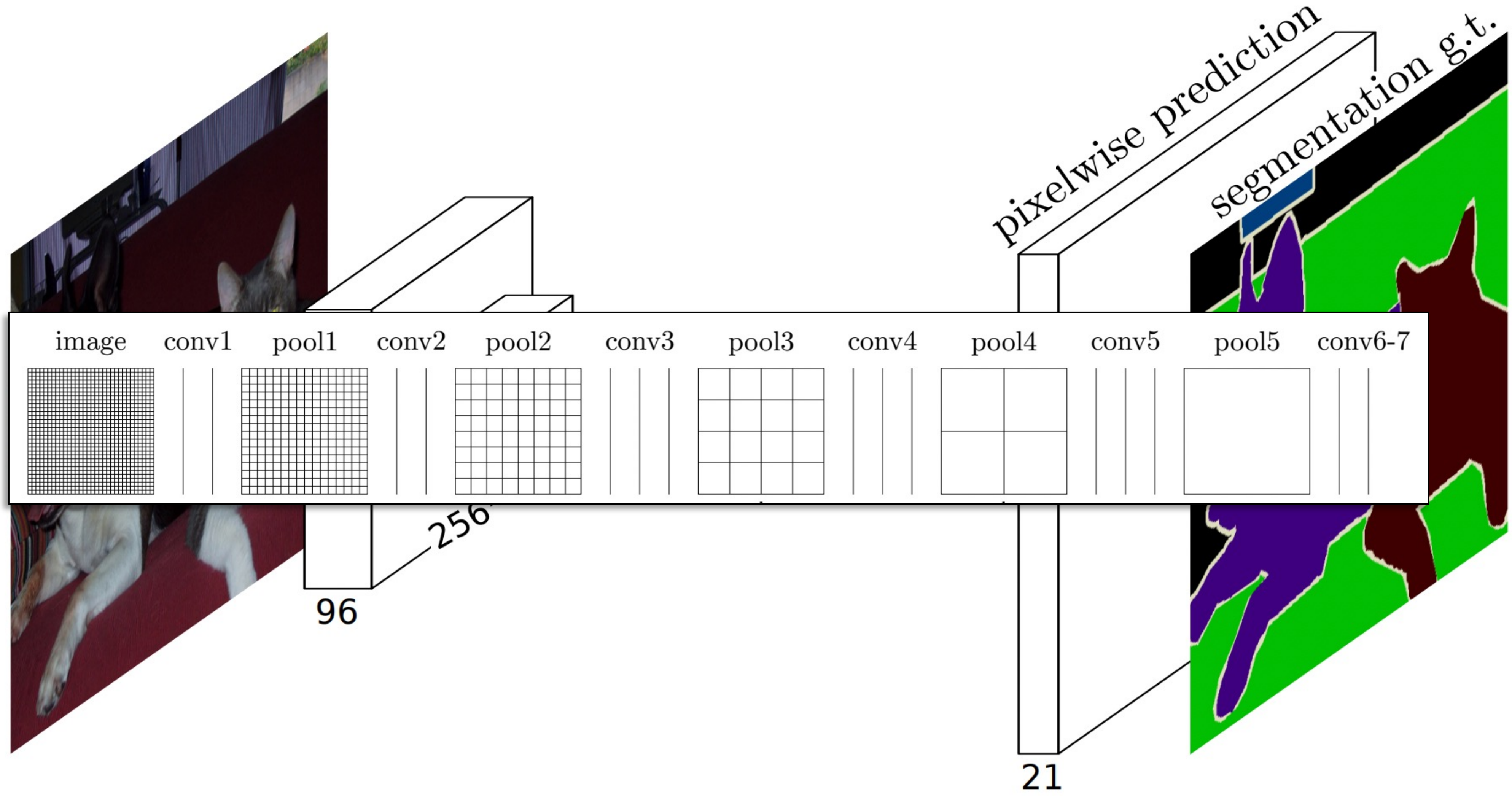


Another result of this change is that, unlike for classification, a class can be assigned to each "coarse region."

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

Architecture: Coarse Region Classification (Recall Intuition)

Using VGG16 instead:

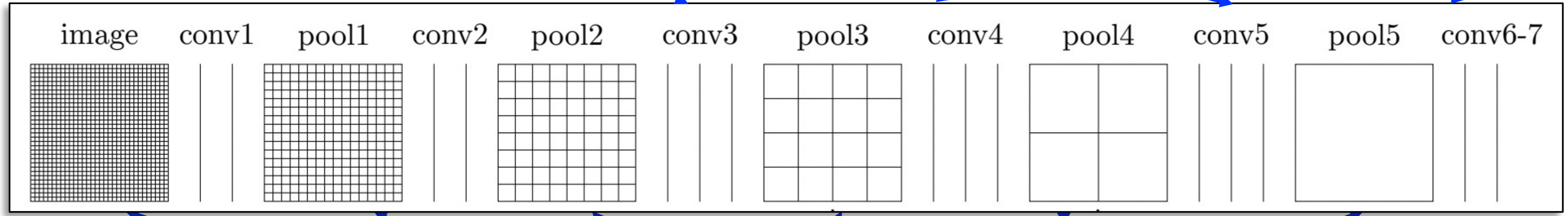


Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Coarse Region Classification (Recall Intuition)

Each line represents a convolutional layer

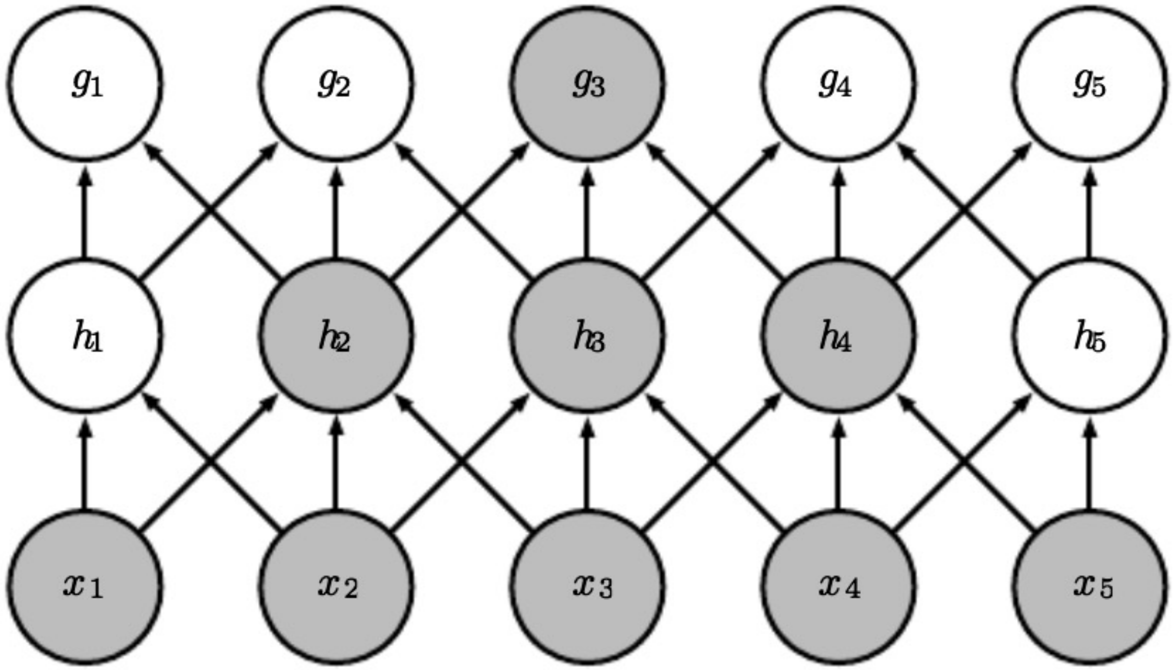
Using VGG16 instead:



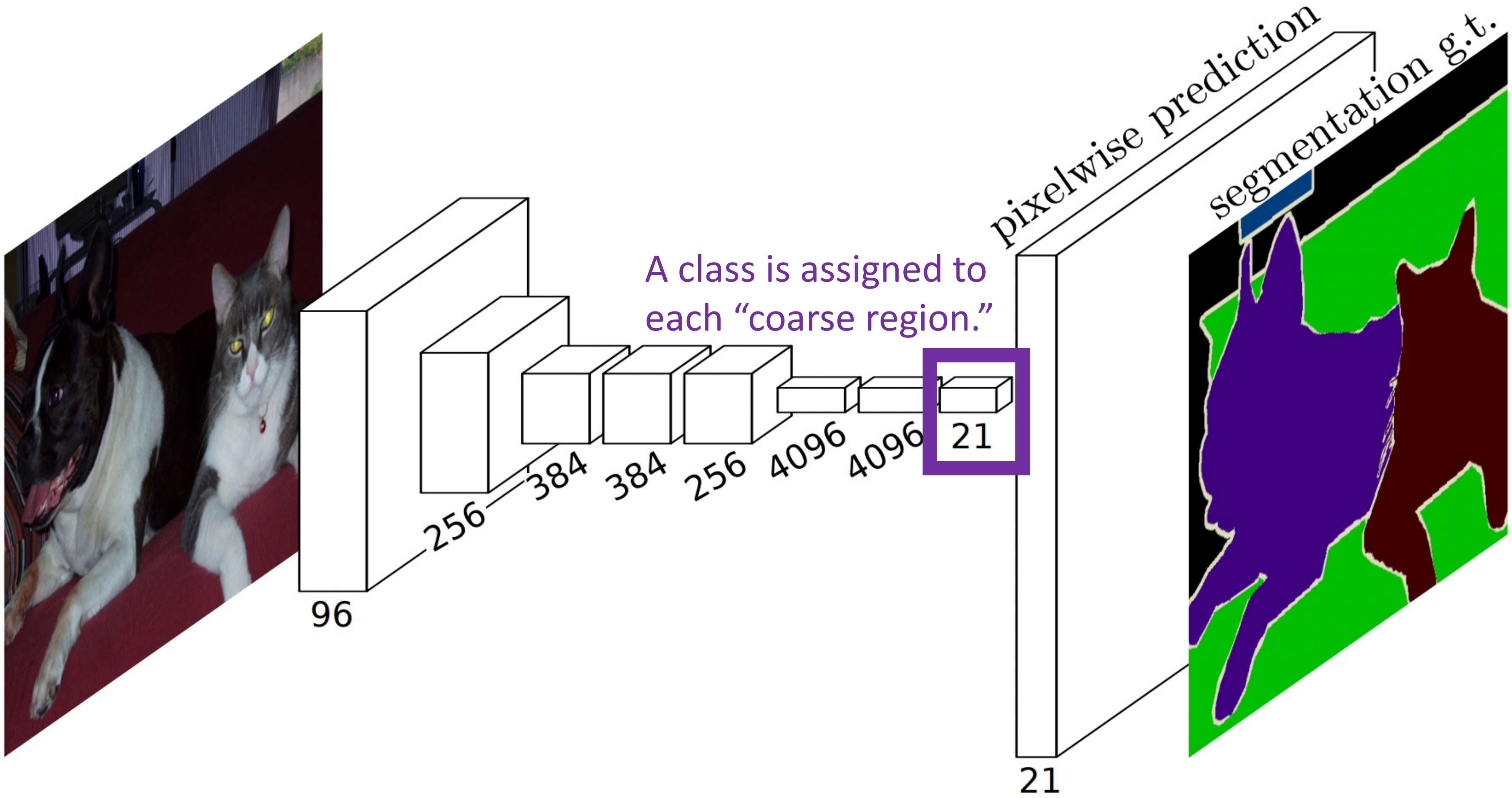
Grids reflect relative spatial coarseness at each layer

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.**



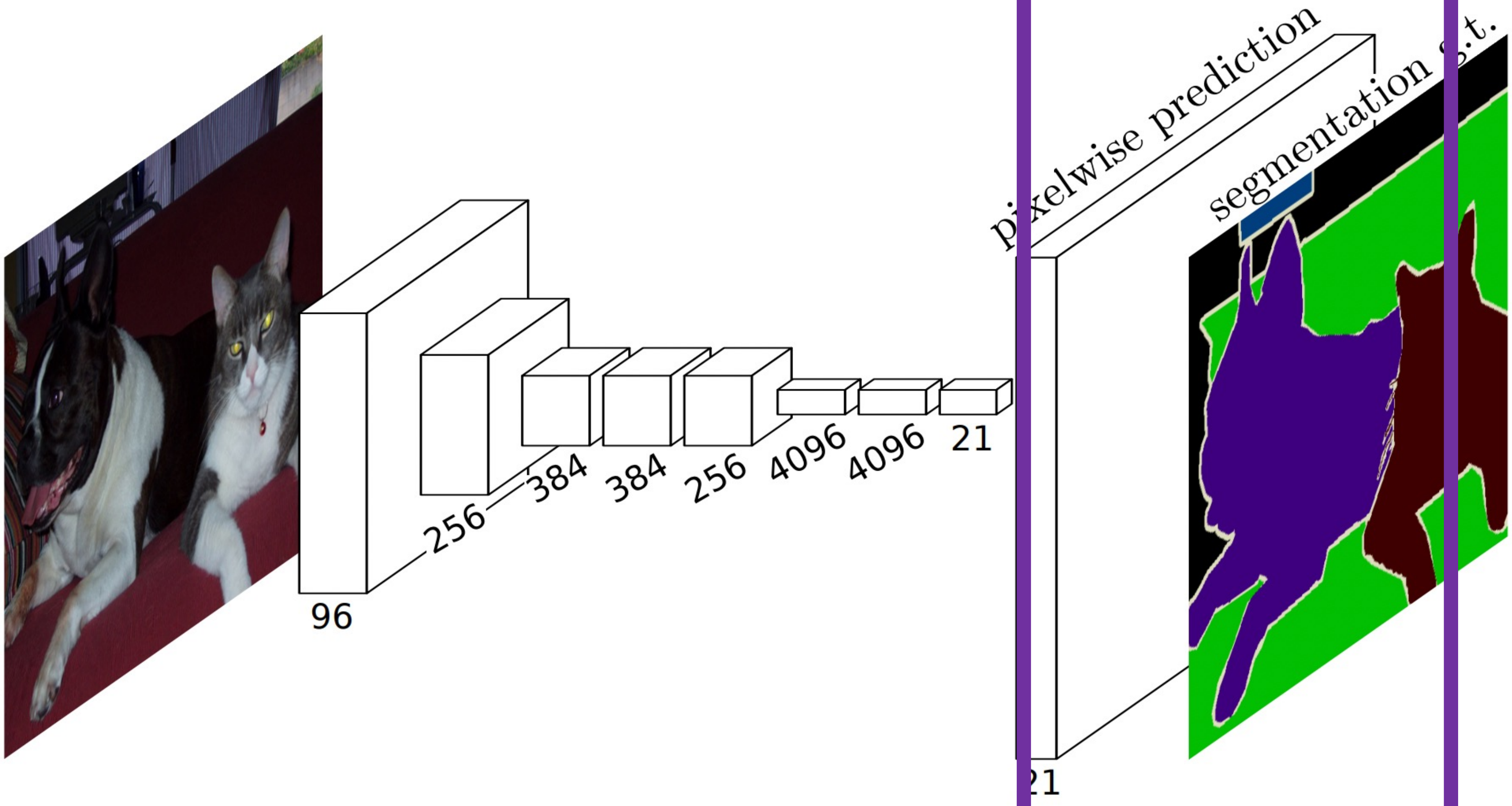
Architecture: Coarse Region Classification (Recall Intuition)



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

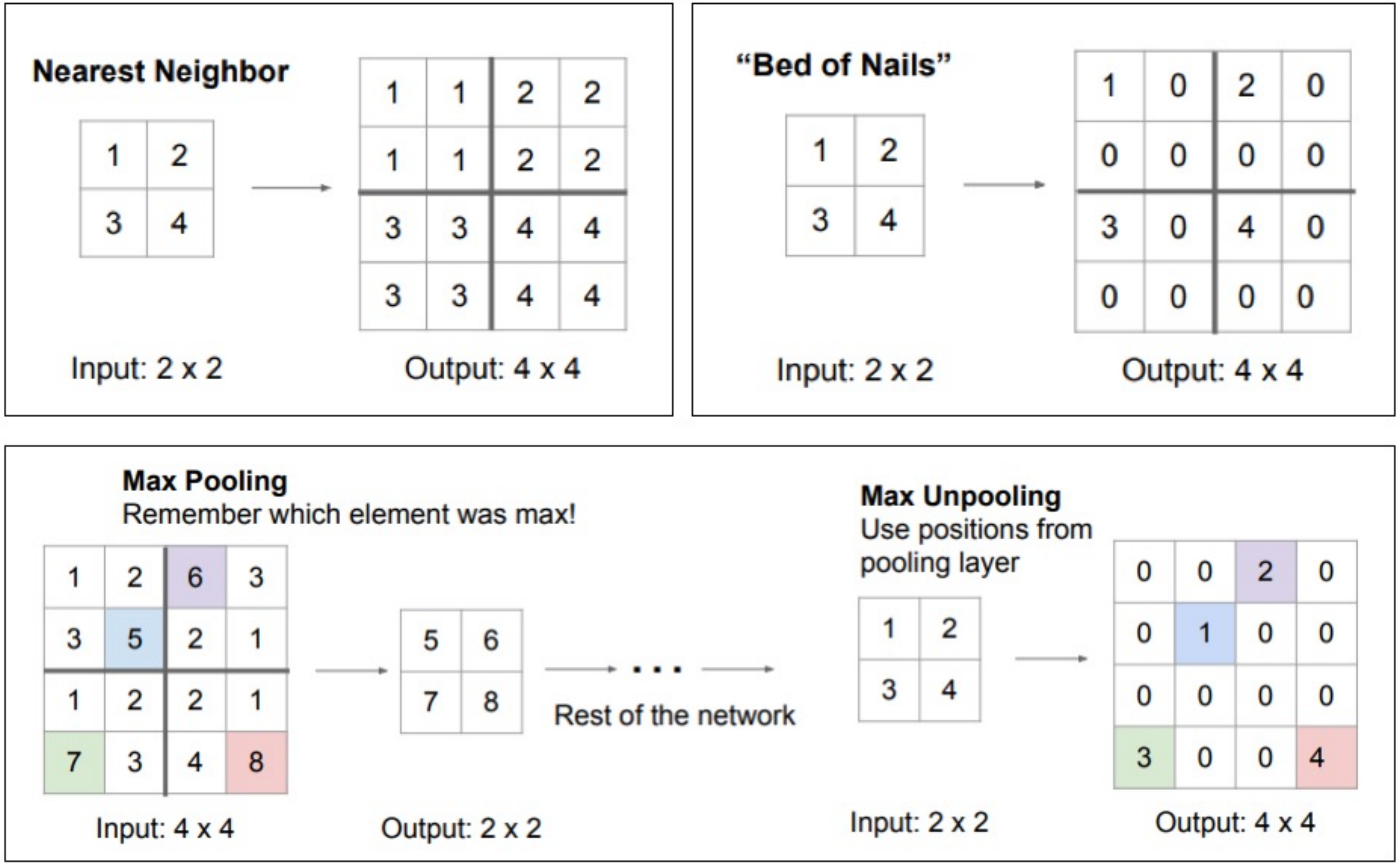
Architecture

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)

- Prior approaches used a convolutional layer to clean-up/refine the hard-coded upsampling approaches
- Idea: learn filters to refine the subsampled image *while* upsampling
- Implementation: looks like convolution in that the number of filters and kernel size of each filter must be specified; stride differs though by appearing like a fractional input, e.g. with a stride of f=1/2 insert rows and columns of 0.0 to achieve the desired stride.
- Also called "fractional convolutional layer" and, incorrectly, "deconvolution layer"

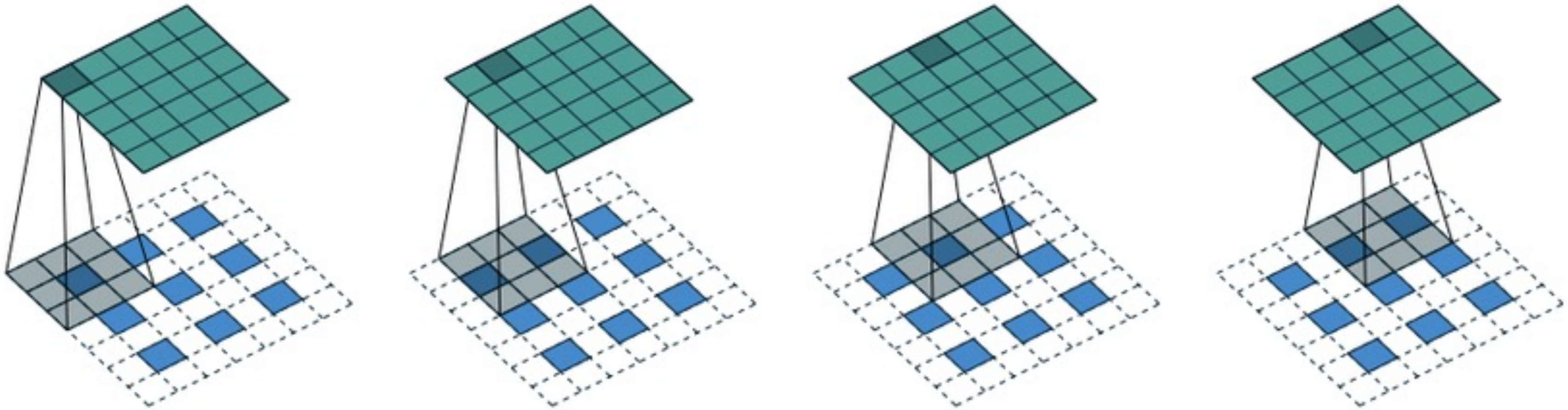
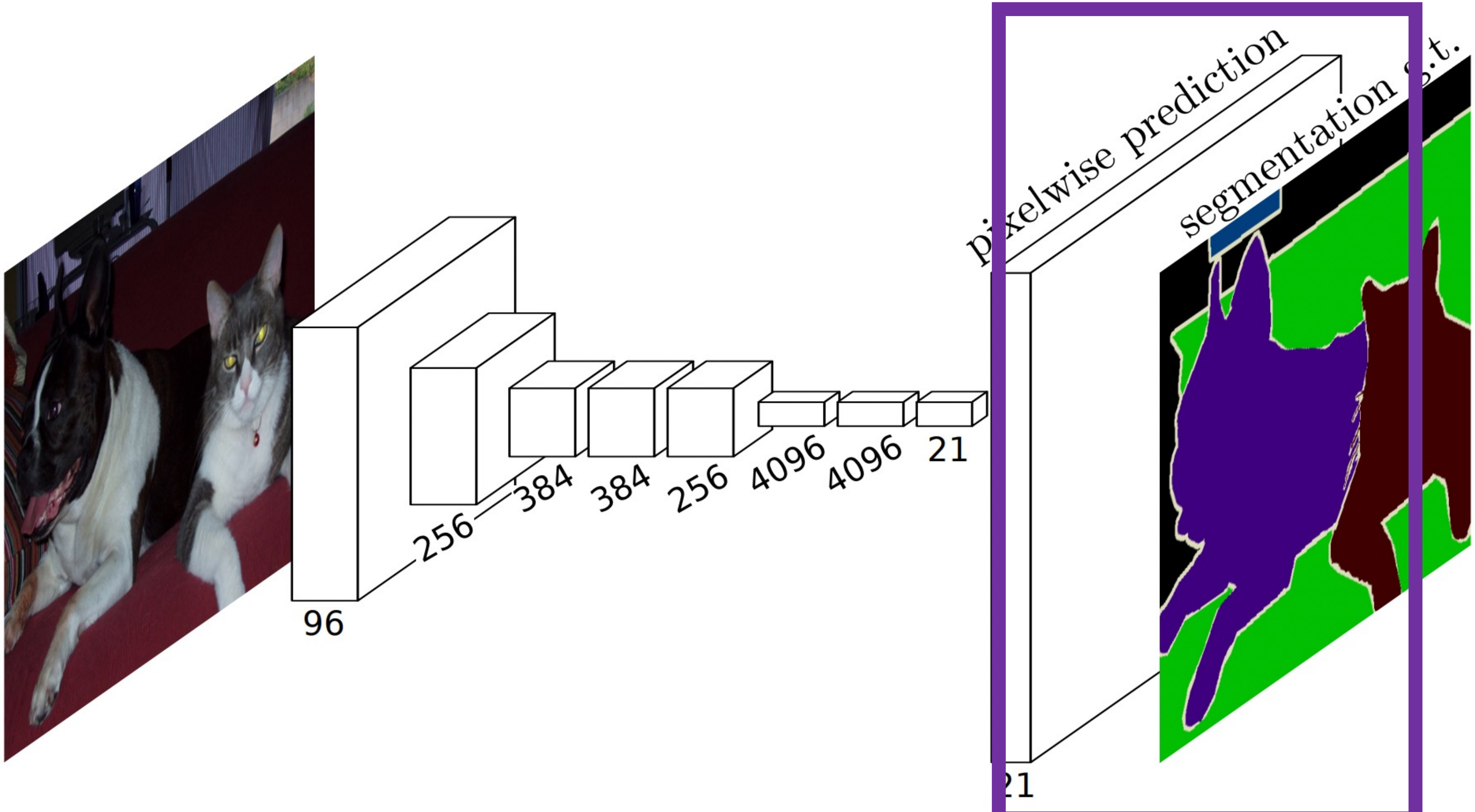


Image Source: https://www.researchgate.net/publication/324783775_Text_to_Image_Synthesis_Using_Generative_Adversarial_Networks

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?

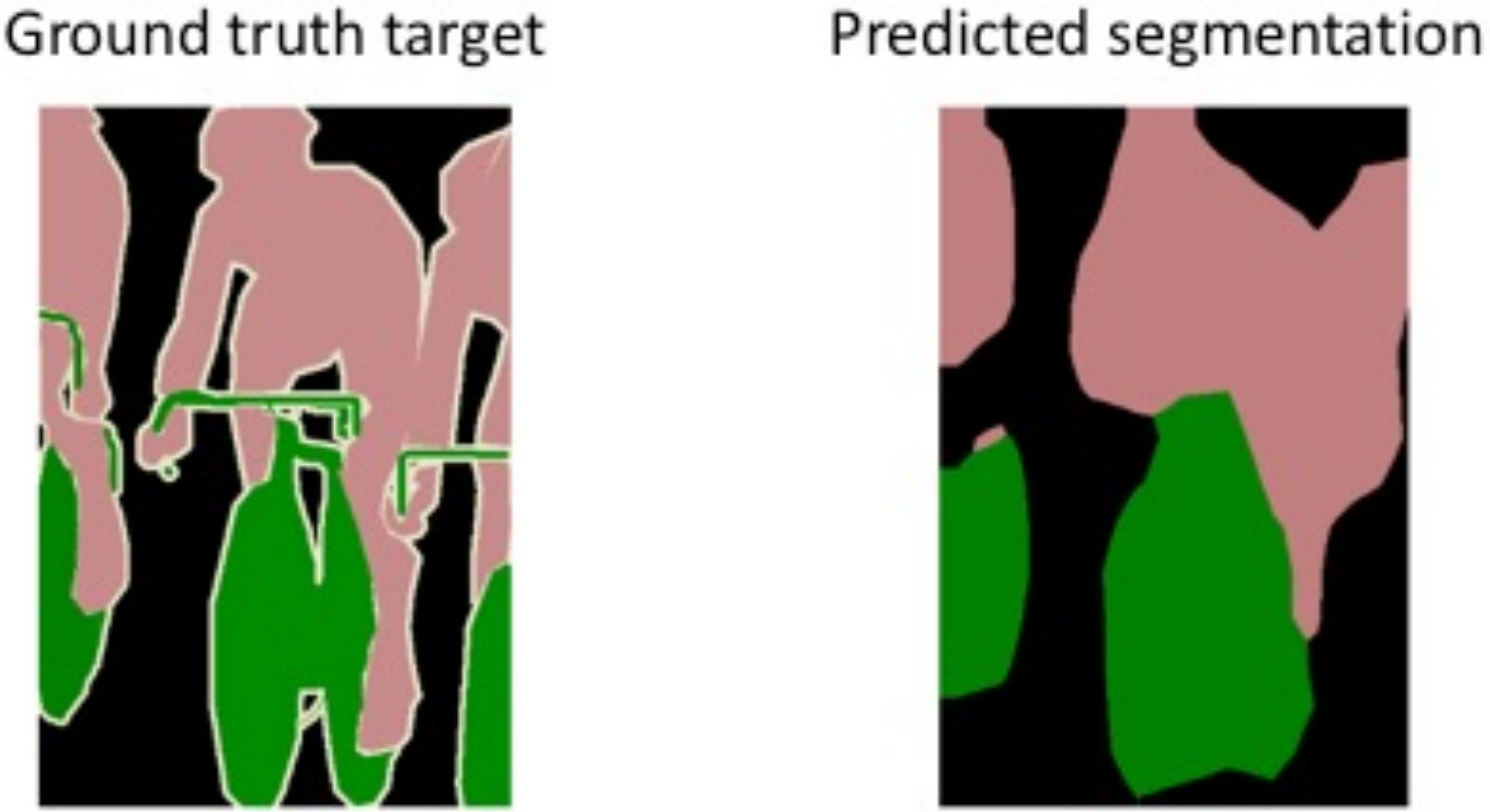
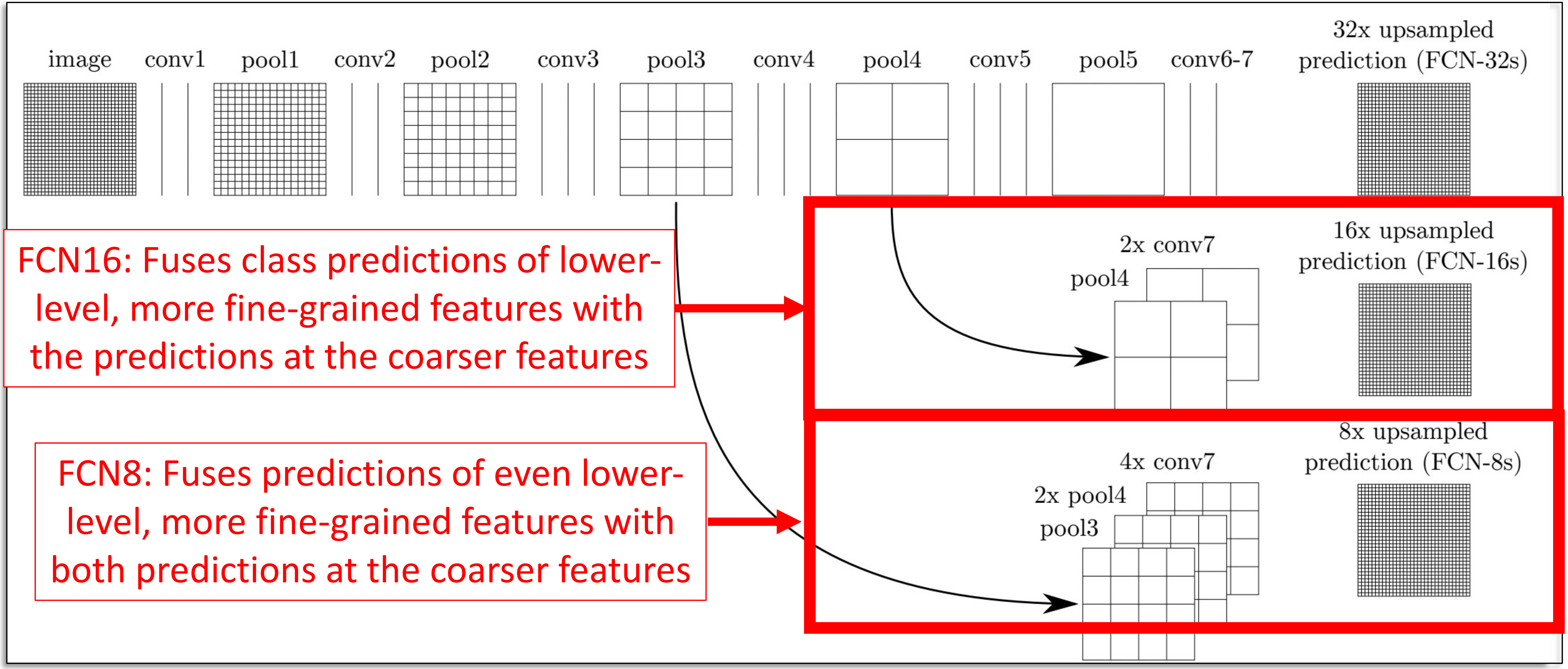


Figure source: https://www.jeremyjordan.me/semantic-segmentation/

Architecture: Update to Use Skip Connections



Architecture: Results

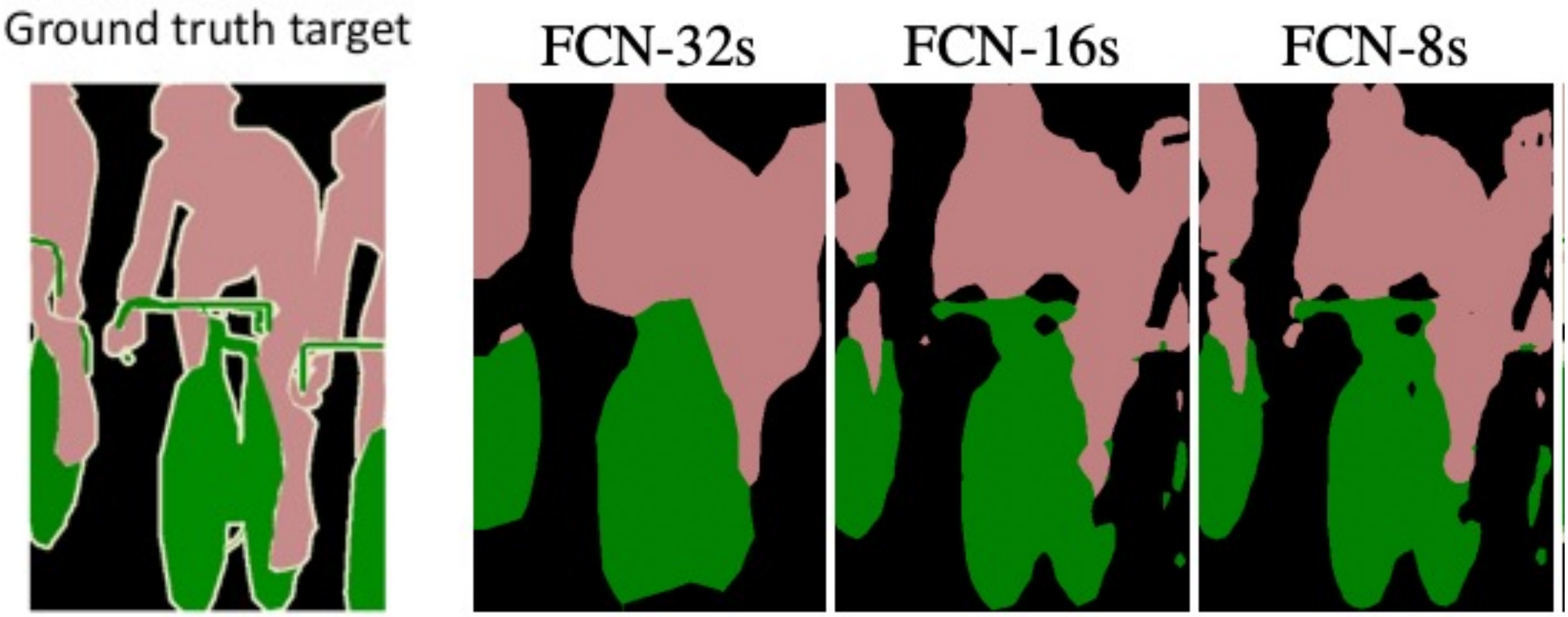
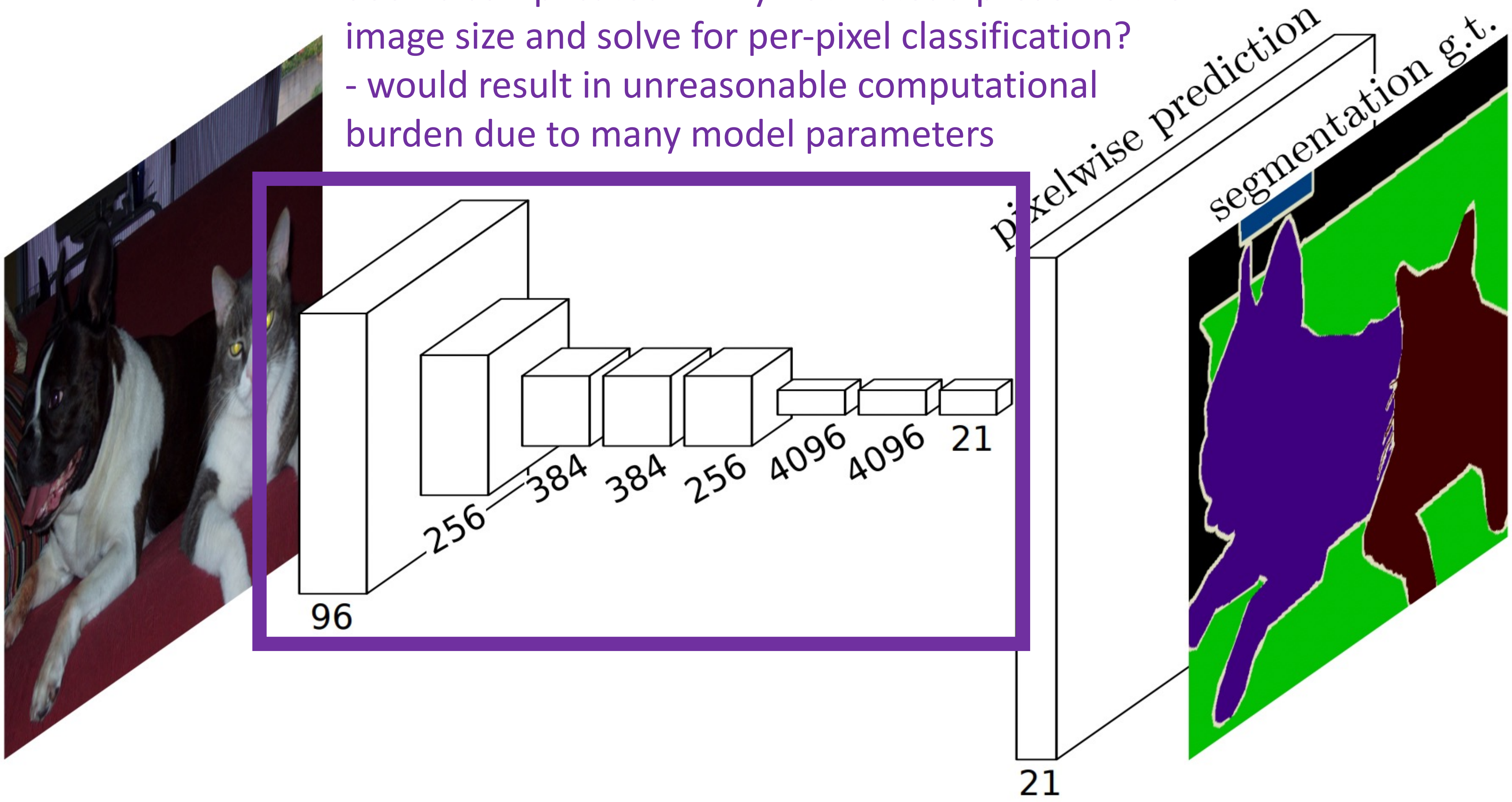


Figure source: https://www.jeremyjordan.me/semantic-segmentation/

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

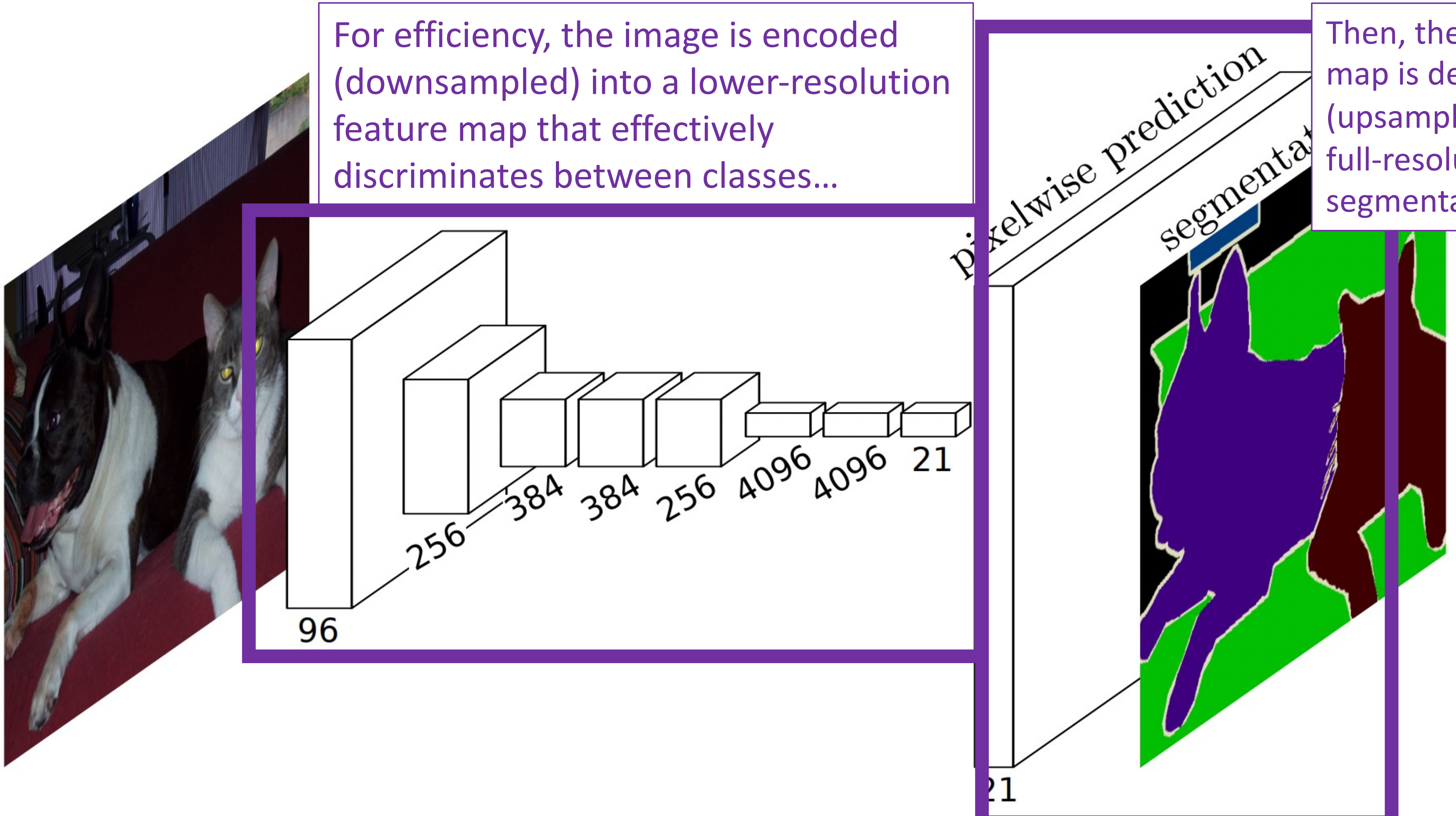


Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Encoder Decoder Architecture

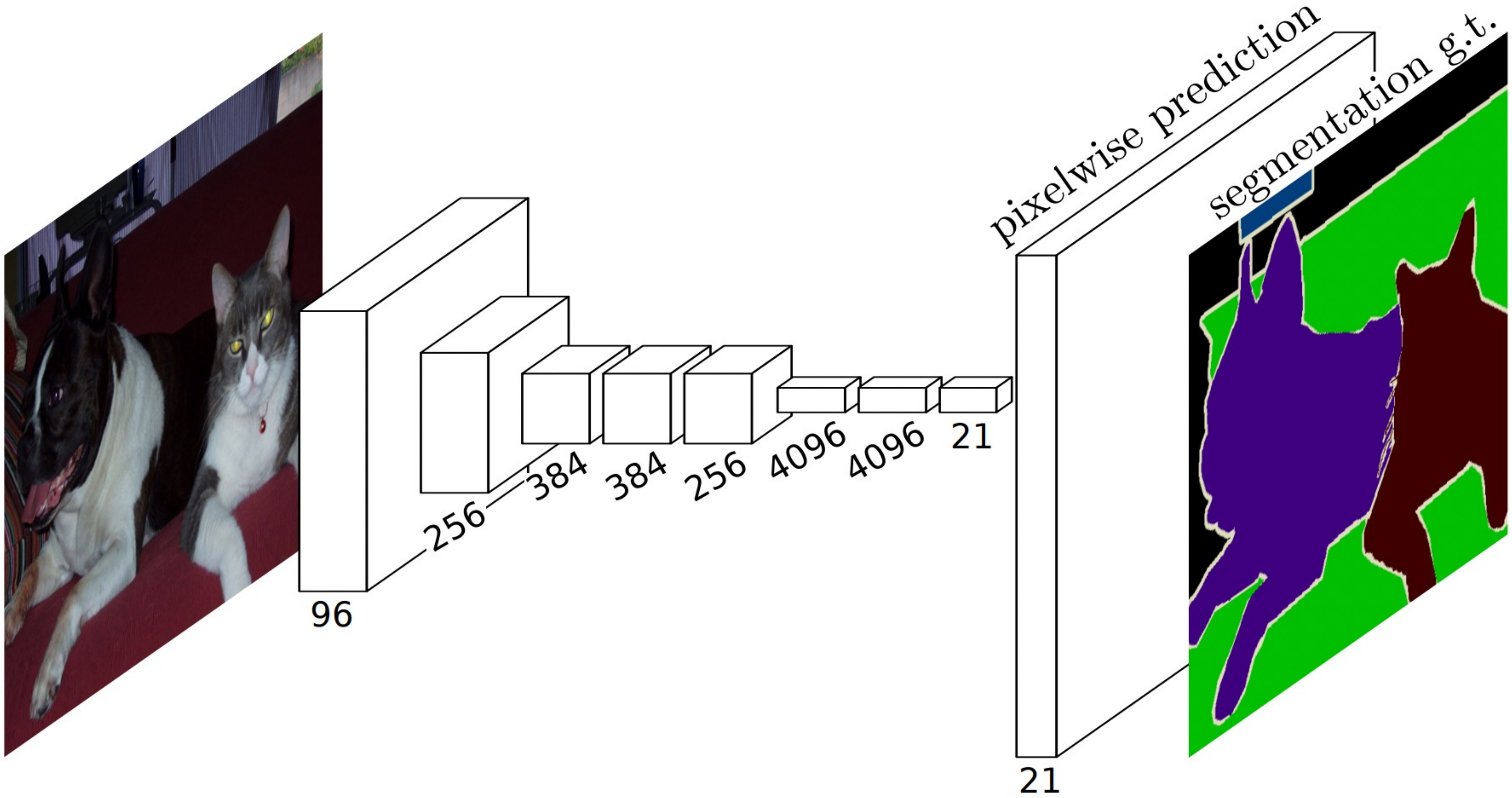
For efficiency, the image is encoded (downsampled) into a lower-resolution feature map that effectively discriminates between classes…

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



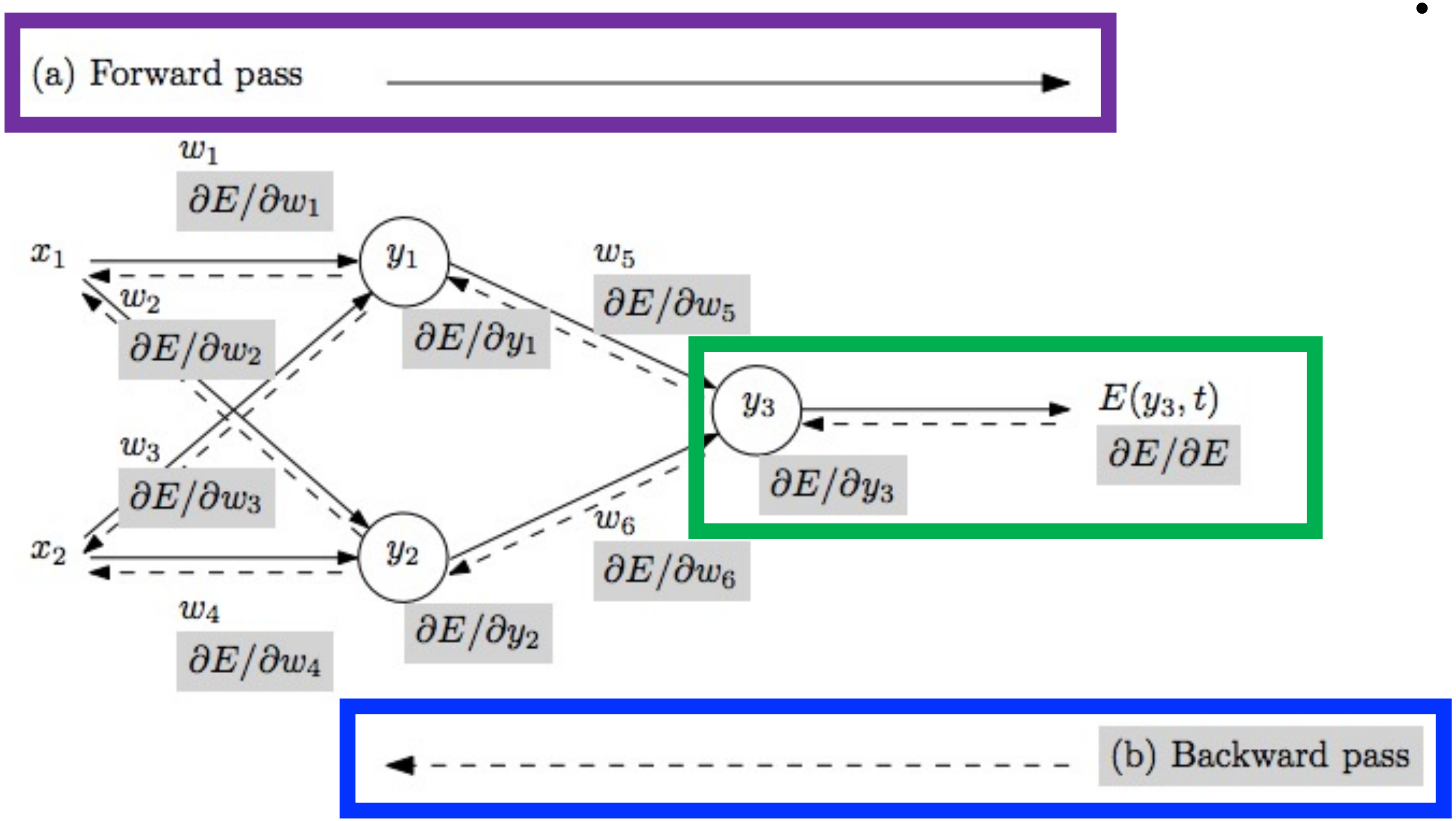
Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Architecture: Algorithm Training



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

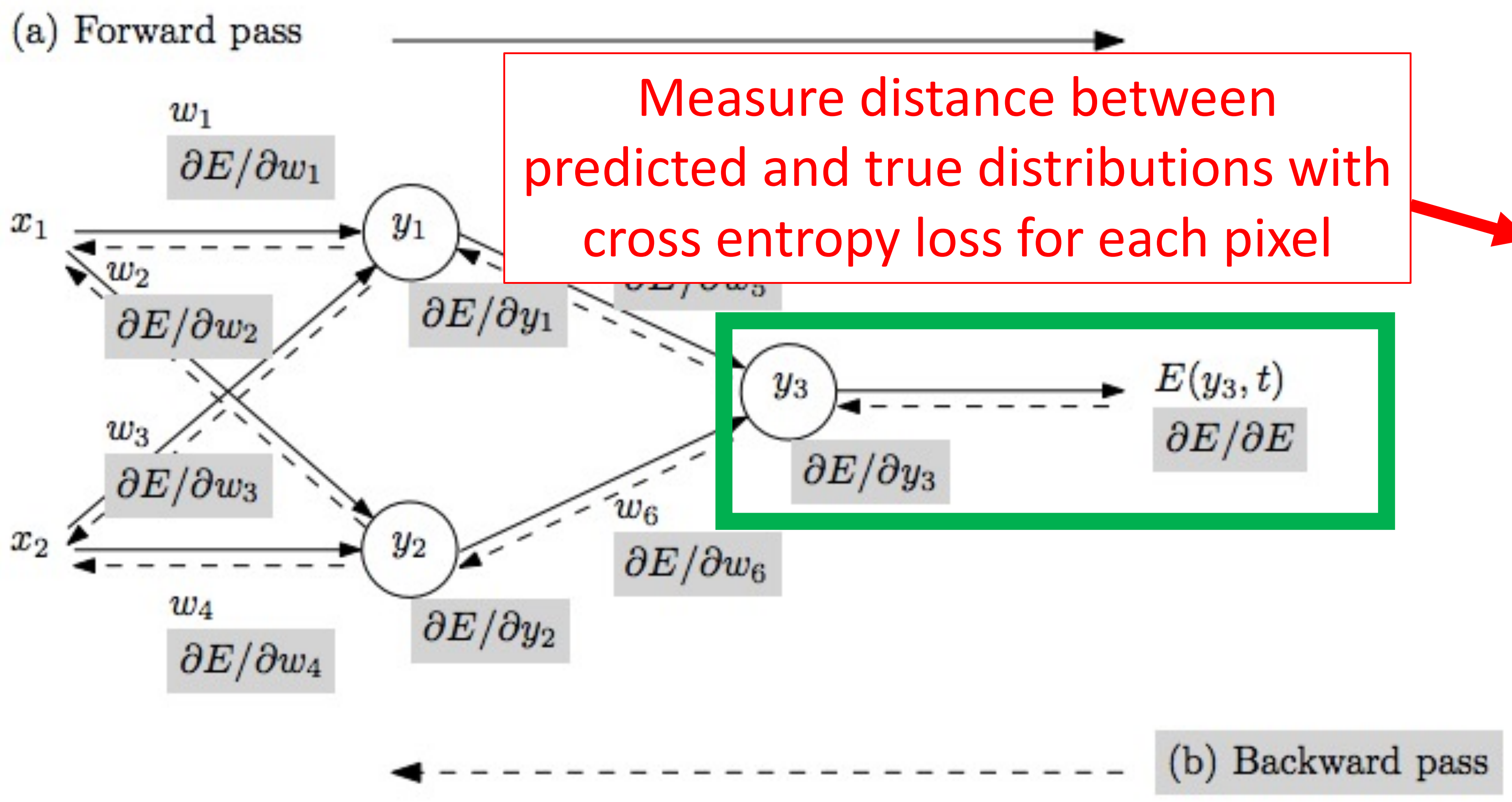
Algorithm Training: Recall How NNs Learn



- 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

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

Algorithm Training: CNN



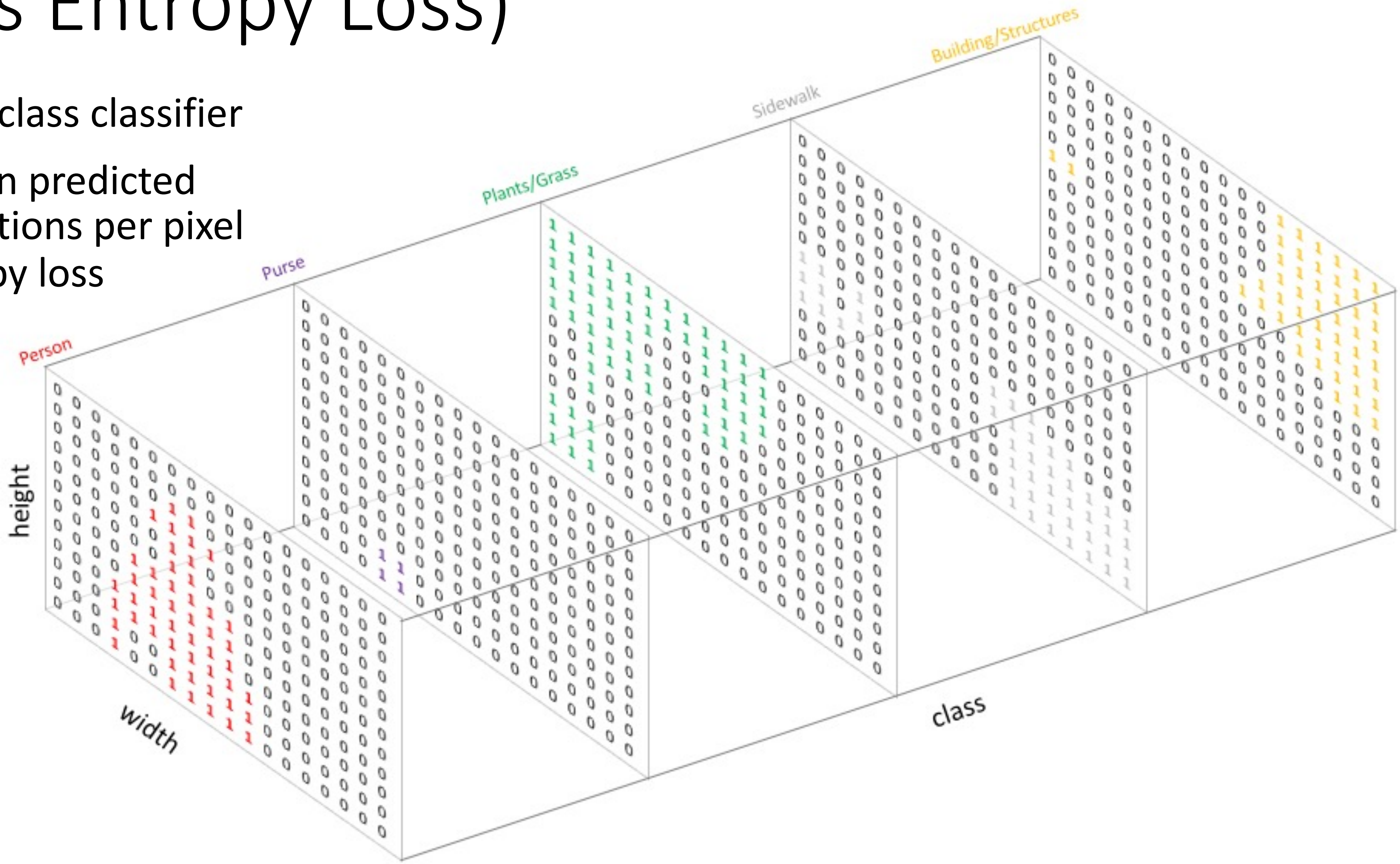
Measure distance between predicted and true distributions with cross entropy loss for each pixel

- 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

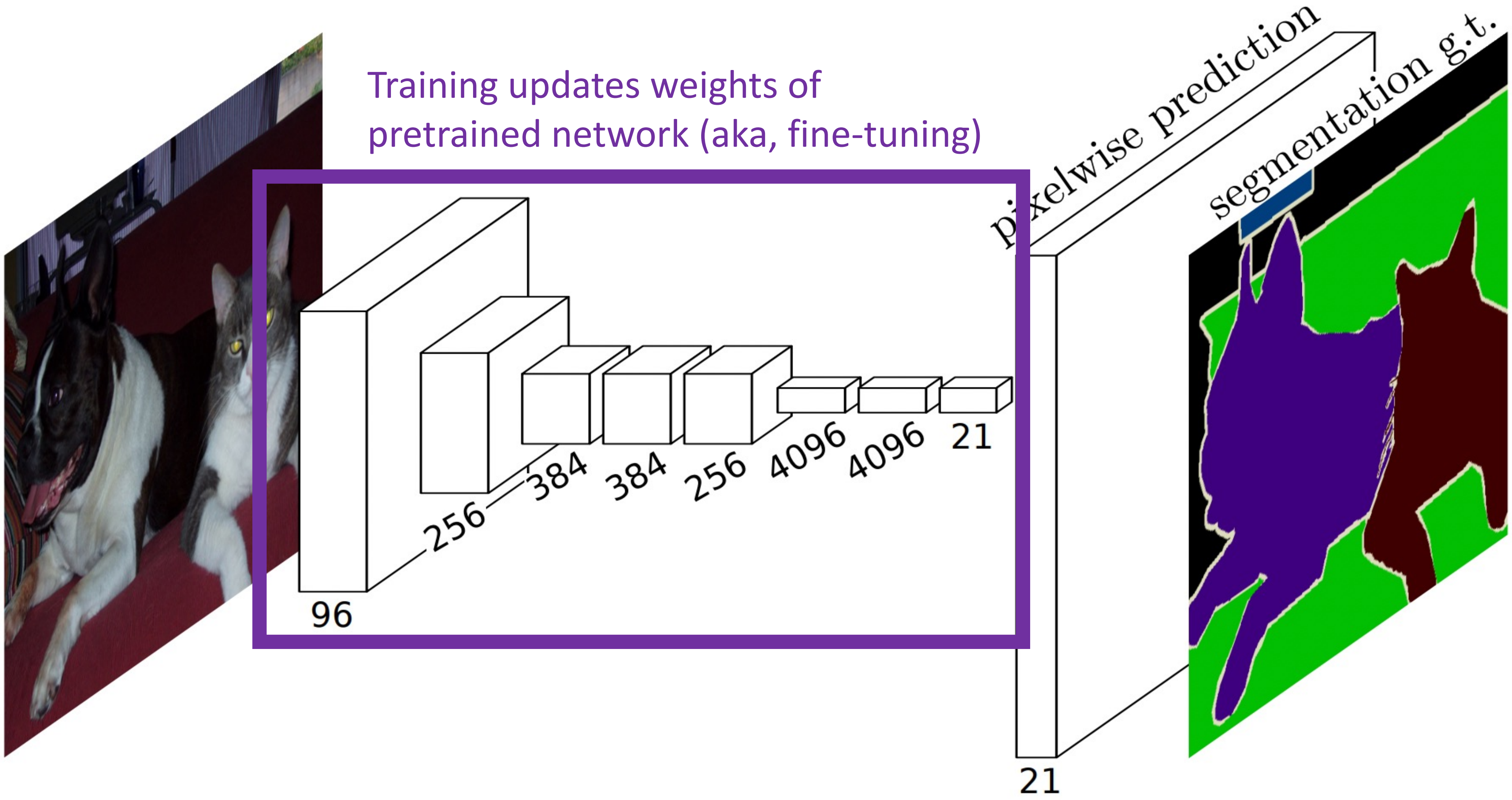
Figure from: Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, Jeffrey Mark Siskind; Automatic Differentiation in Machine Learning: a Survey; 2018

Algorithm Training: Multinomial Logistic Loss (i.e., Cross Entropy 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.

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.

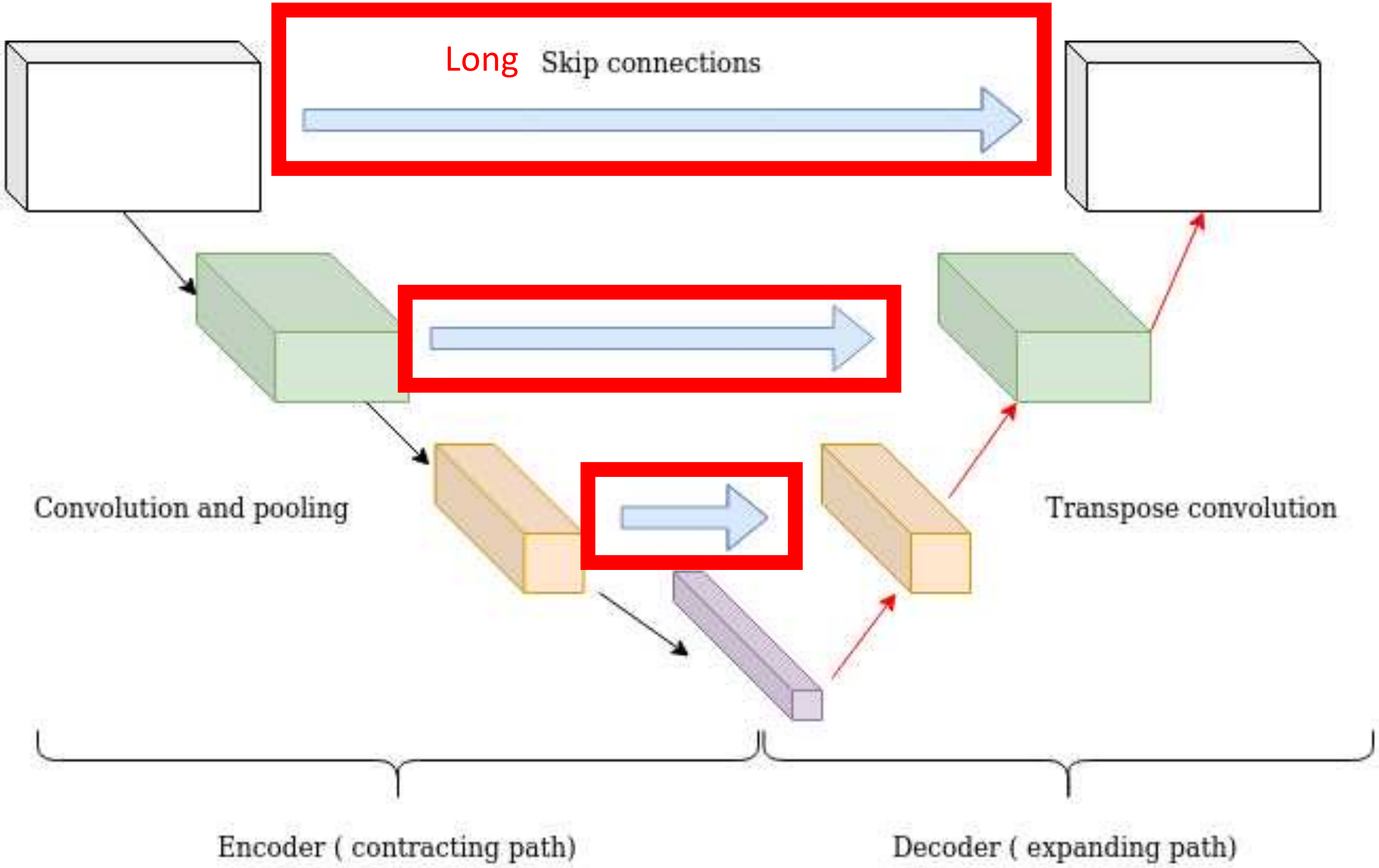
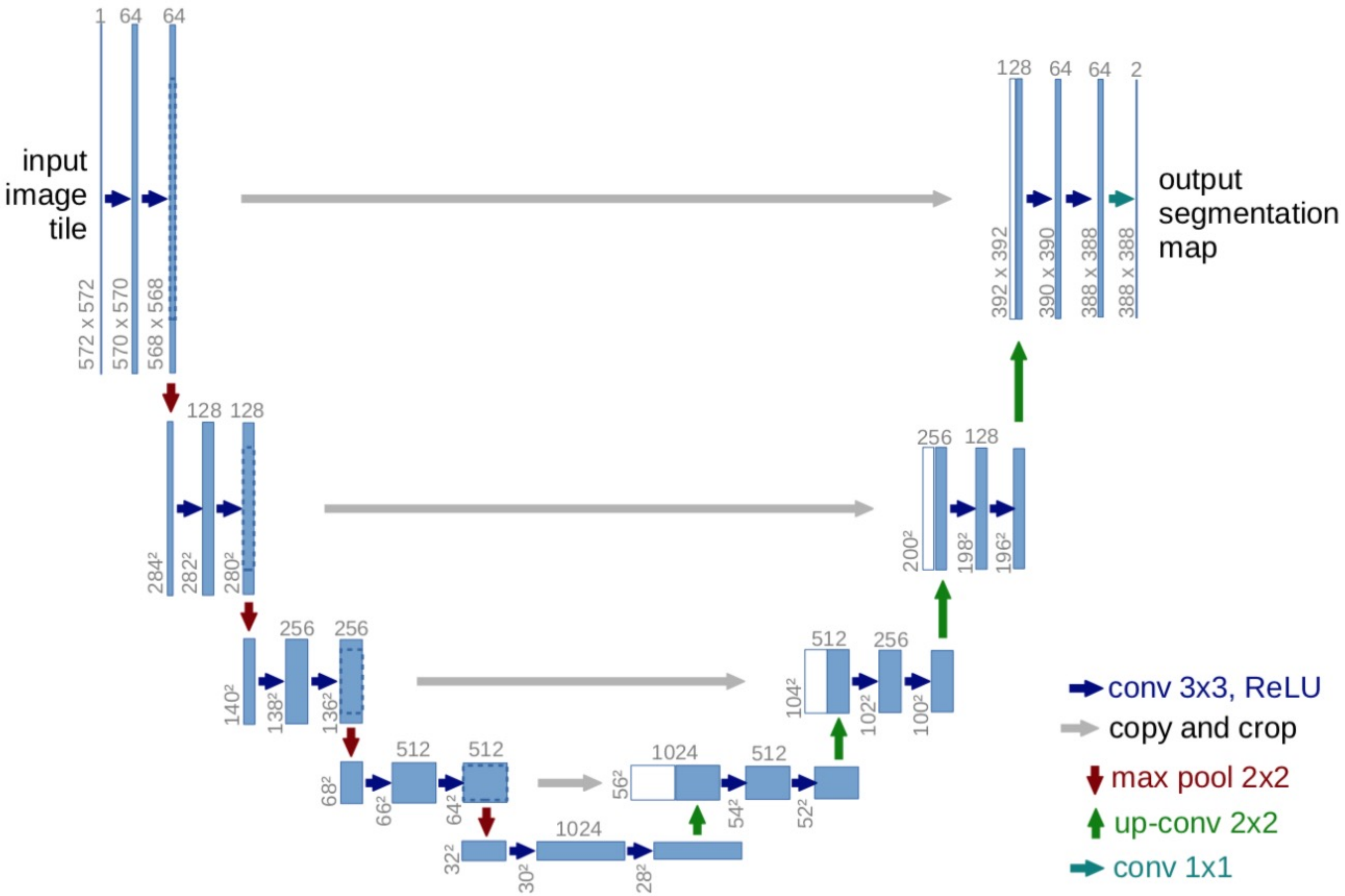


Image Source: https://theaisummer.com/skip-connections/

U-Net



1
64
64
input
image
tile
572 x 572
570 x 570
568 x 568
128
128
284²
282²
280²
256
256
140²
138²
136²
512
512
68²
66²
64²
1024
32²
30²
28²
1024
512
56²
54²
52²
512
256
104²
102²
100²
256
128
200²
198²
196²
128
64
64
2
392 x 392
390 x 390
388 x 388
388 x 388
output
segmentation
map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

Semantic Segmentation: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Computer vision models: fully convolutional networks


The End