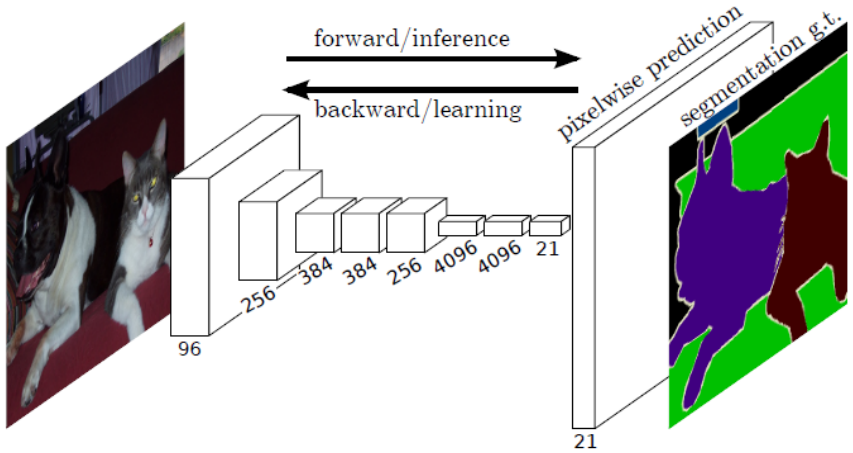


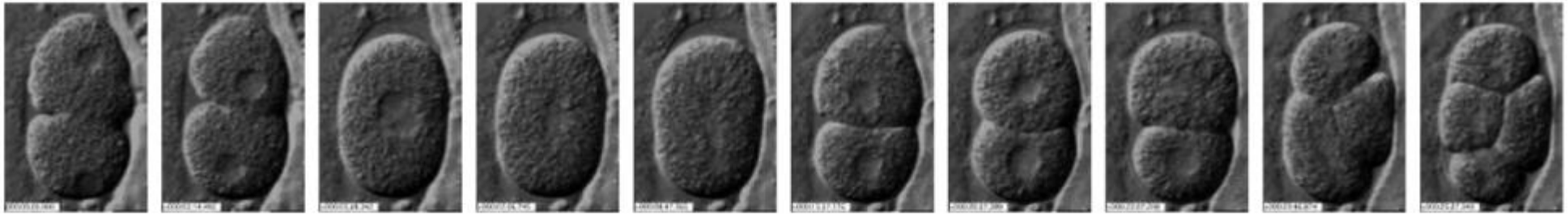

# Fully Convolutional Networks for Semantic Segmentation [1]

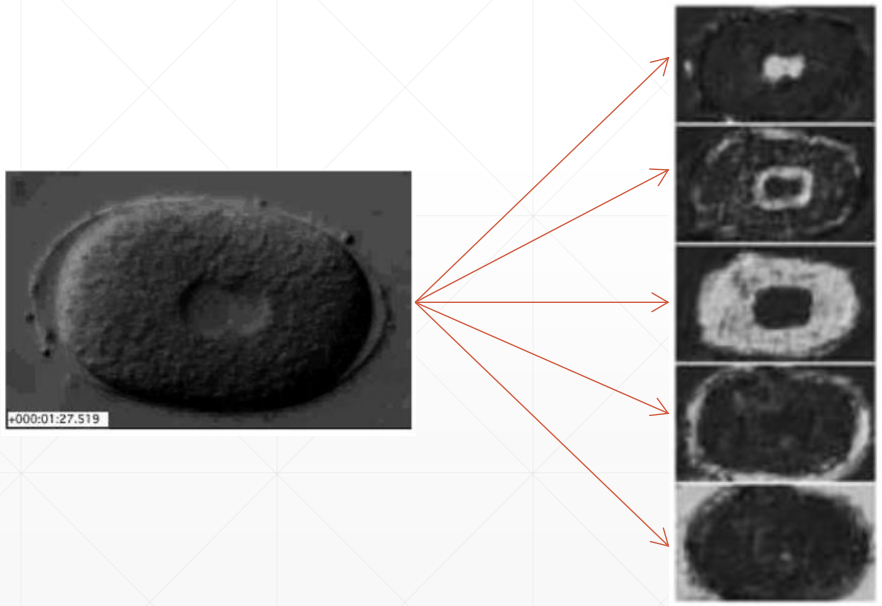
Jonathan Long, Evan Shelhamer and Trevor Darrell

ECE 289G: Paper Presentation

Philipp Gysel

# Analyse Genome of *C Elegans*






1) Nucleus

2) Nucleus membrane

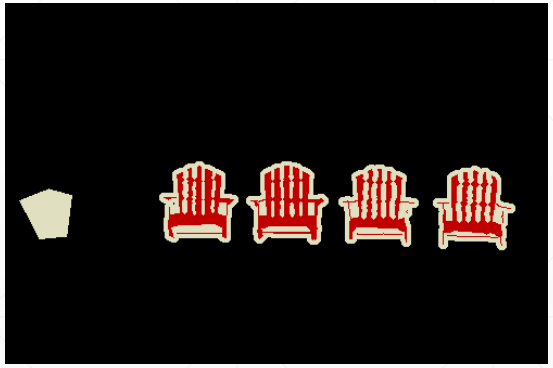
3) Cytoplasm

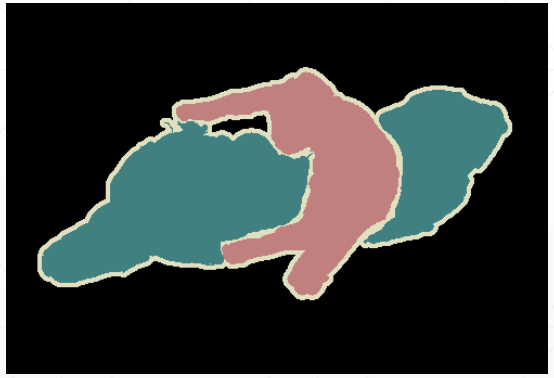
4) Cell wall

5) External medium

[2] and http://www.godandscience.org

# PASCAL VOC 2011





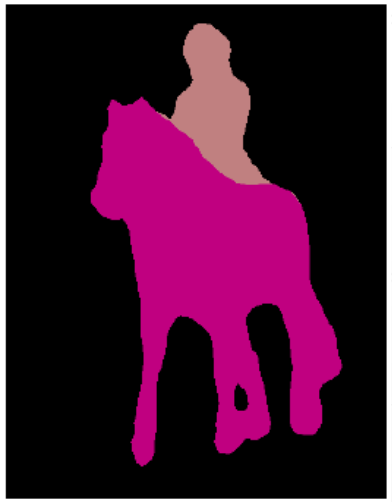
- Person
- Motorbike
- Chair

[3]

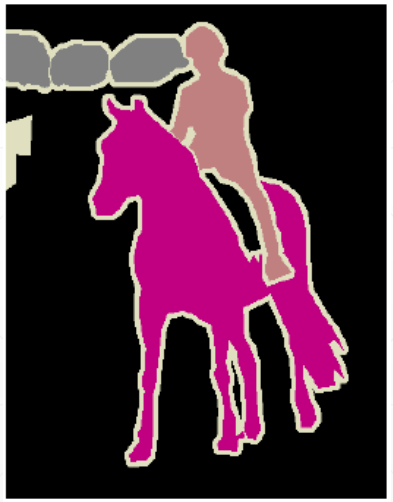
# Accuracy Metric

- Mean intersection over union (IU): $(1/n_{\mathrm{cl}}) \sum_i n_{ii} / \left( t_i + \sum_j n_{ji} - n_{ii} \right)$
- Pixel Accuracy: $(1/n_{\mathrm{cl}}) \sum_i n_{ii} / t_i$

Prediction



Ground truth

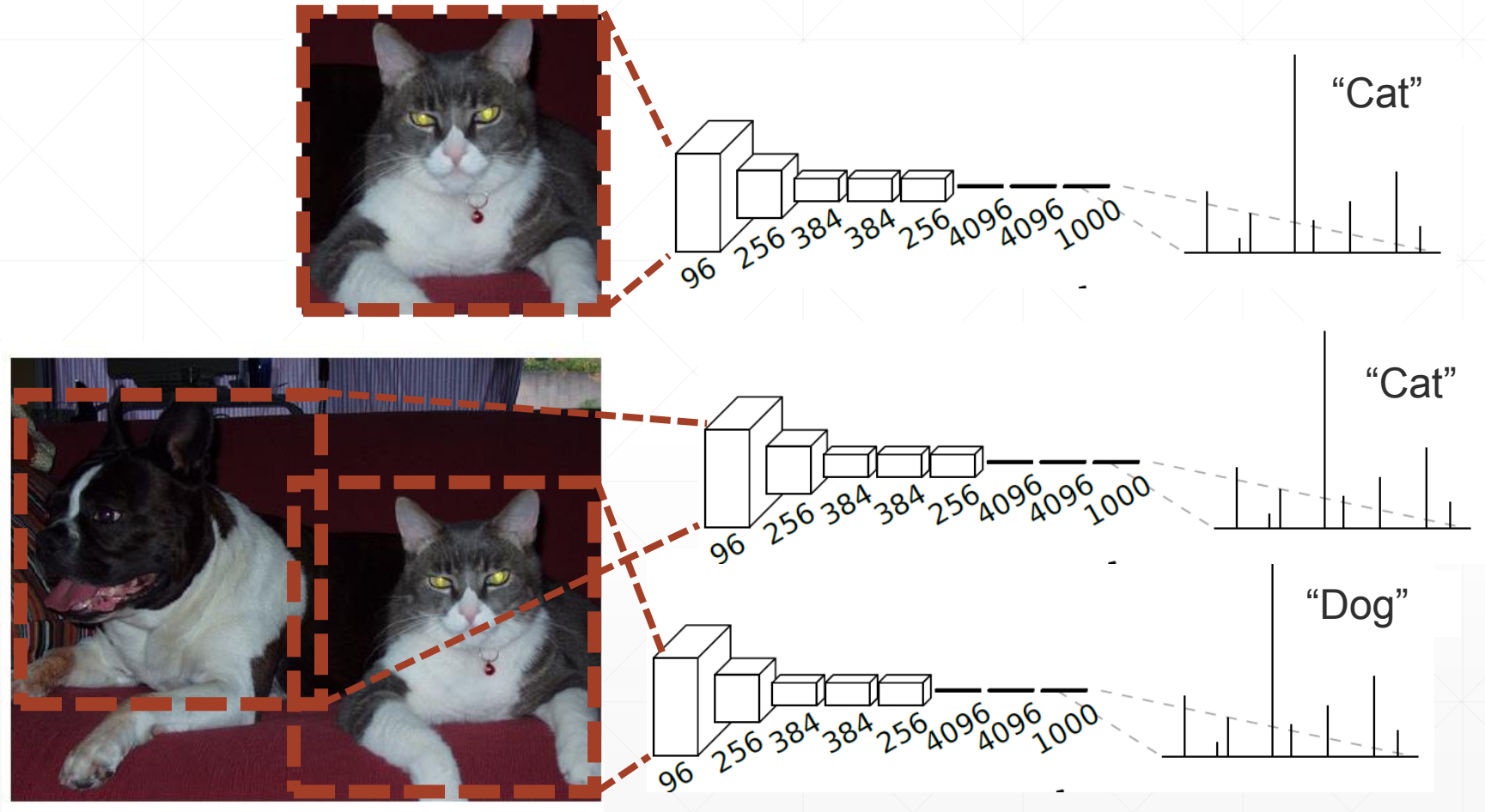


[1]

# FCN: Fully Connected CNN

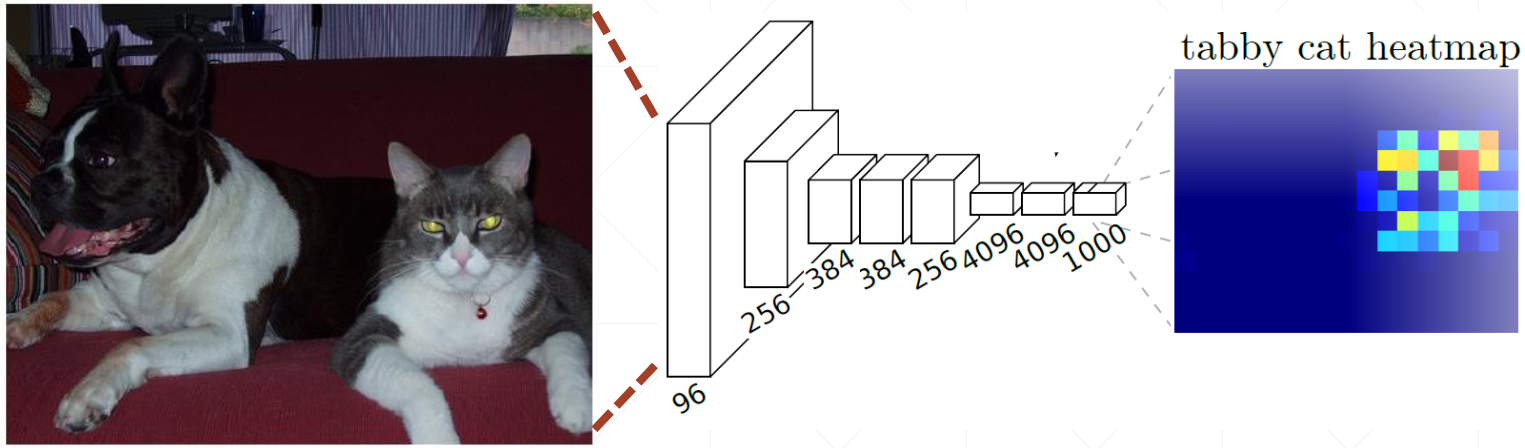
CNN:

FCN:

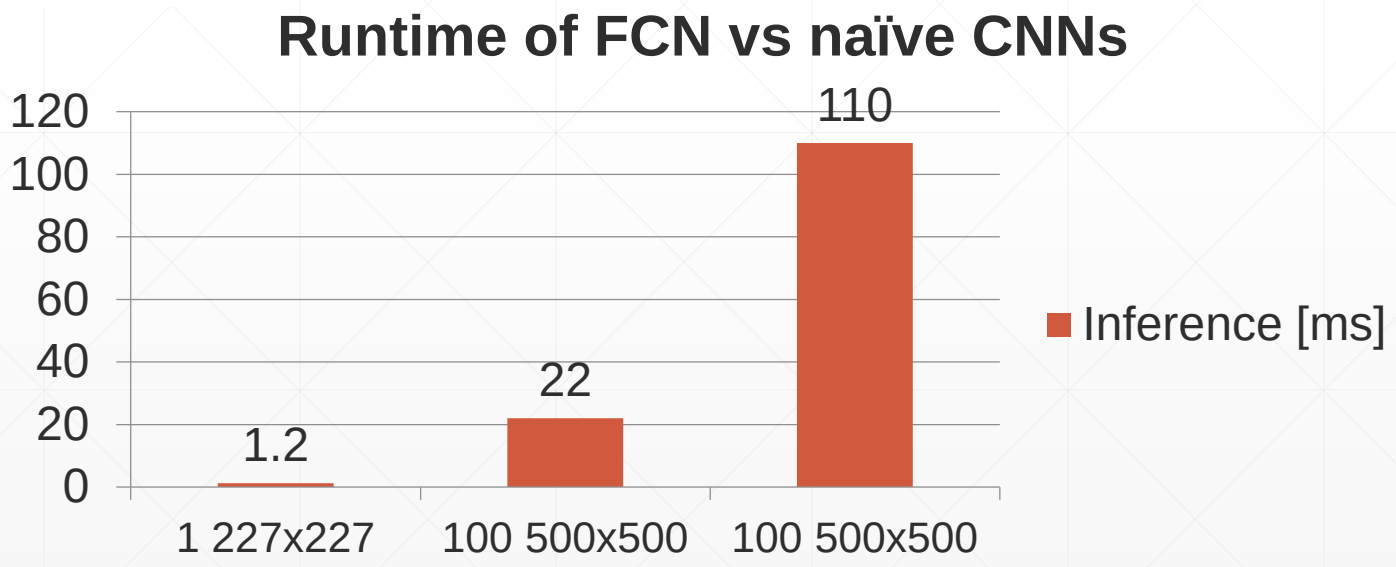



[1]

# FCN Speedup

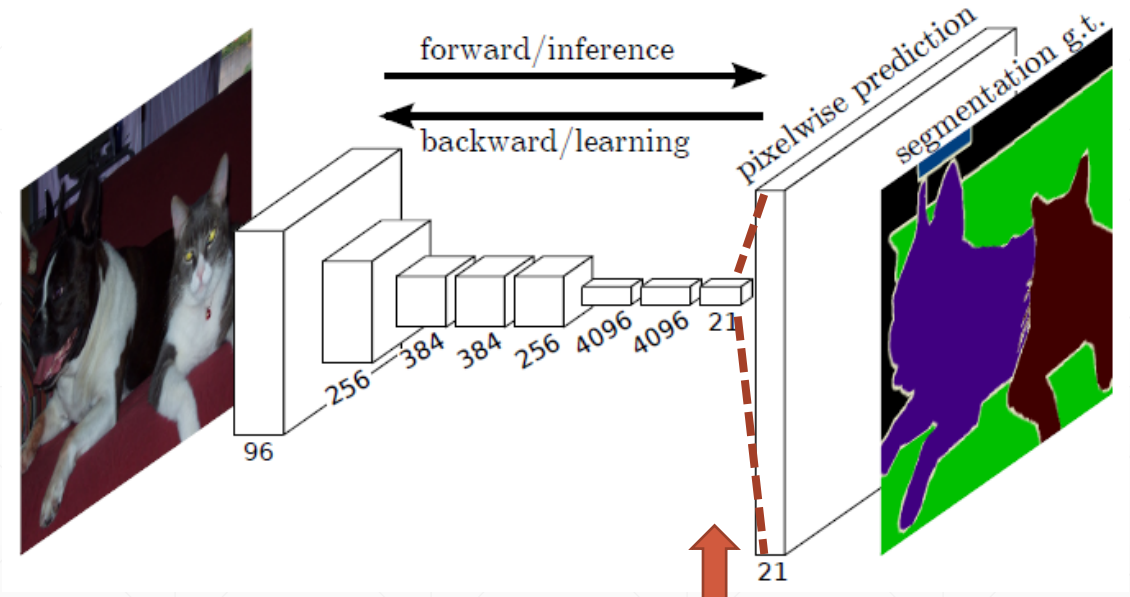


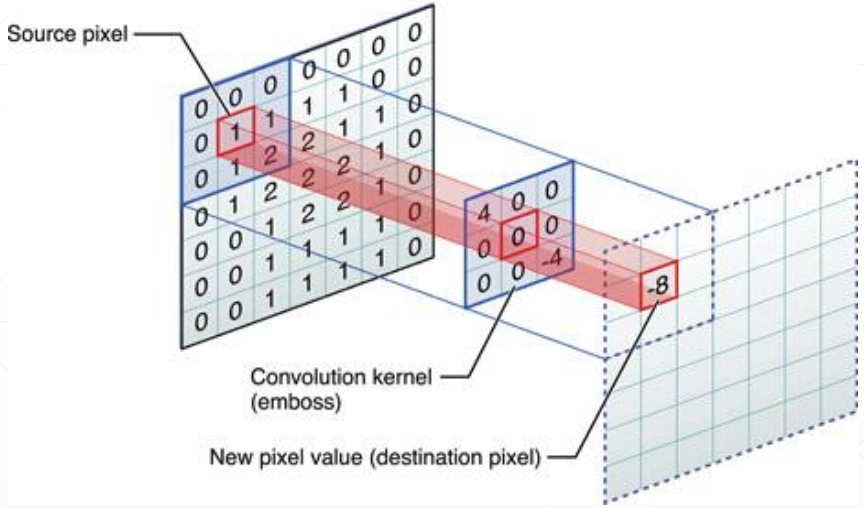

- Keep kernel sizes and strides
- Replace dense layer with convolution




[1]

# Upsampling: Backwards strided convolution



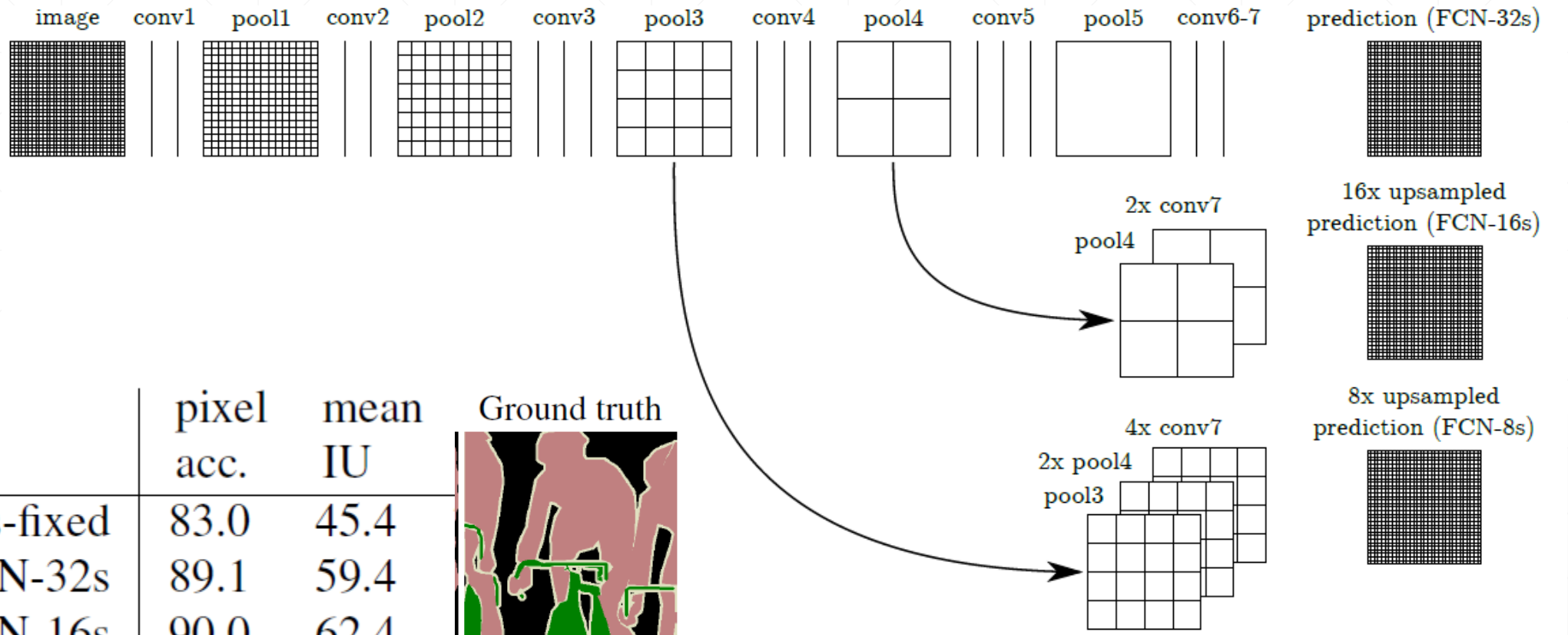




In-network upsampling

pixelwise loss

[1], https://developer.apple.com

# Deep jet



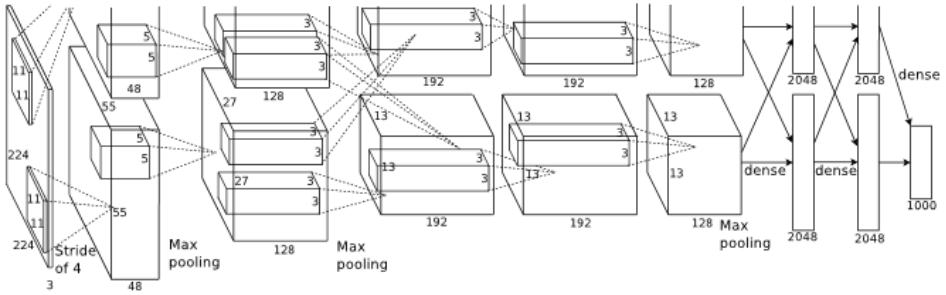

| | pixel acc. | mean IU |
|---|---|---|
| FCN-32s-fixed | 83.0 | 45.4 |
| FCN-32s | 89.1 | 59.4 |
| FCN-16s | 90.0 | 62.4 |
| FCN-8s | **90.3** | **62.7** |

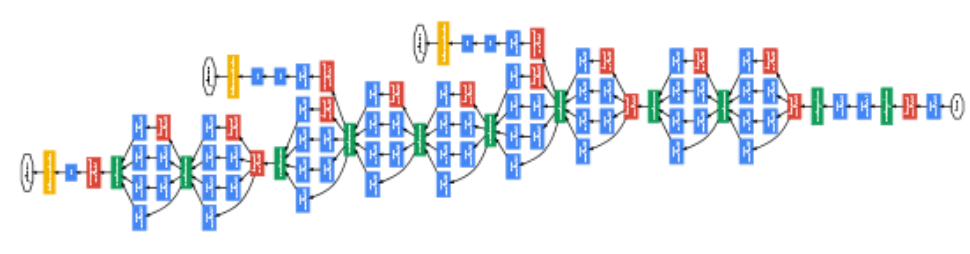
Ground truth

[1]

# Experiment setup

- Pre-trained networks: **AlexNet**, **VGG** and **GoogLeNet**
- Convert dense to convolutional layer
- Discard final classifier
- Add deconvolution layer for up-sampling
- Fine-tuning end-to-end on
  - PASCAL VOC
  - NYUDv2
  - SIFT



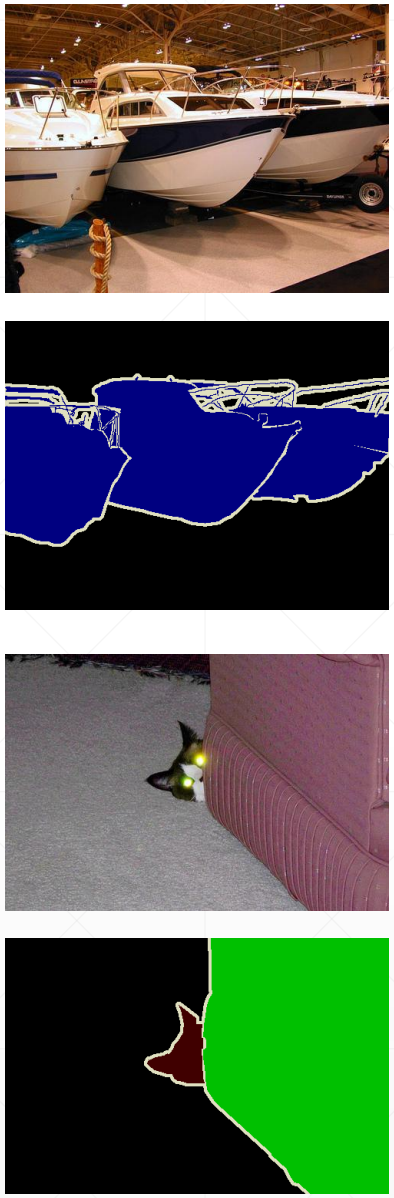



[4], [5]

# Experiment #1: PASCAL VOC 2011

- 20 classes (e.g. airplane, boat, bicycle, person, cat)
- Relative margin of 20% to previous state-of-art
- Inference time is reduced 114x and 286x respectively

| | mean IU VOC2011 test | mean IU VOC2012 test | inference time |
|---|---|---|---|
| R-CNN [12] | 47.9 | - | - |
| SDS [17] | 52.6 | 51.6 | ~ 50 s |
| FCN-8s | **62.7** | **62.2** | **~ 175 ms** |

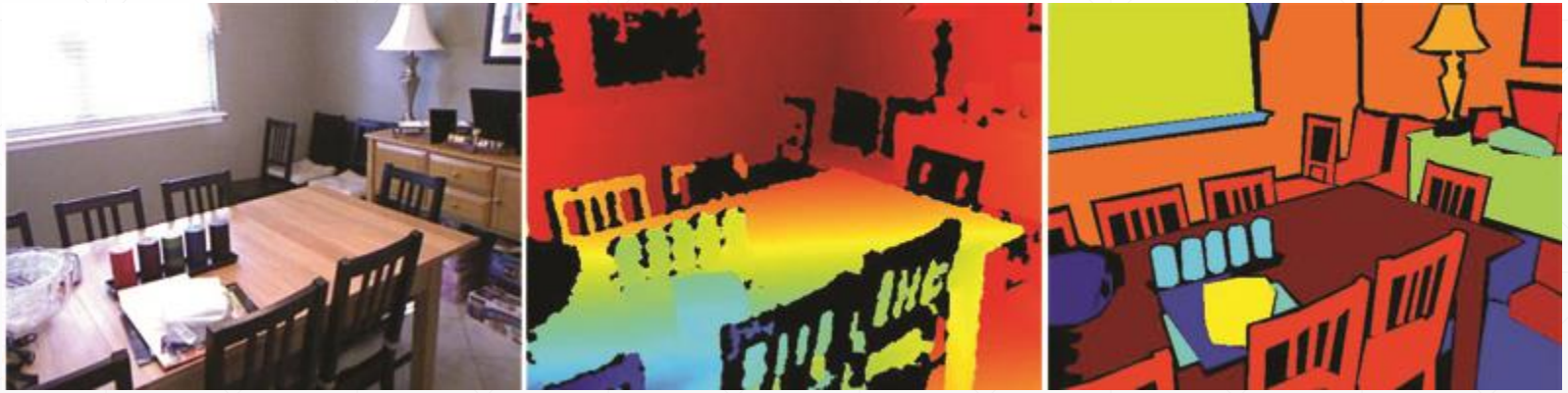


[1], [3]

# Experiment #2: NYUDv2

- 40 classes from indoor scenes
- Densely labeled RGB and depth images

| | pixel acc. | mean IU |
|---|---|---|
| Gupta *et al.* [15] | 60.3 | 28.6 |
| FCN-32s RGB | 60.0 | 29.2 |
| FCN-32s RGBD | 61.5 | 30.5 |
| FCN-32s HHA | 57.1 | 24.2 |
| FCN-32s RGB-HHA | 64.3 | 32.8 |
| FCN-16s RGB-HHA | **65.4** | **34.0** |



[1], http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

# Experiment #3: SIFT Flow

- 33 semantic categories (e.g. bridge, mountain)
- 3 geometric categories (horizontal, vertical)
- Two-headed FCN

| | pixel acc. | geom. acc. |
|---|---|---|
| Liu *et al.* [25] | 76.7 | - |
| Tighe *et al.* [36] | - | 90.8 |
| Tighe *et al.* [37] 1 | 75.6 | - |
| Tighe *et al.* [37] 2 | 78.6 | - |
| Farabet *et al.* [9] 1 | 72.3 | - |
| Farabet *et al.* [9] 2 | 78.5 | - |
| Pinheiro *et al.* [31] | 77.7 | - |
| FCN-16s | **85.2** | **94.3** |

[1]

# Questions?

# References

[1] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *arXiv preprint arXiv:1411.4038* (2014).

[2] Ning, Feng, et al. "Toward automatic phenotyping of developing embryos from videos." *Image Processing, IEEE Transactions on* 14.9 (2005): 1360-1371.

[3] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." *International journal of computer vision* 88.2 (2010): 303-338.

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[5] Szegedy, Christian, et al. "Going deeper with convolutions." *arXiv preprint arXiv:1409.4842* (2014).