Convolutional Networks

Lecture 03

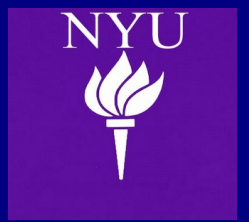
Yann Le Cun

Facebook AI Research,

Center for Data Science, NYU

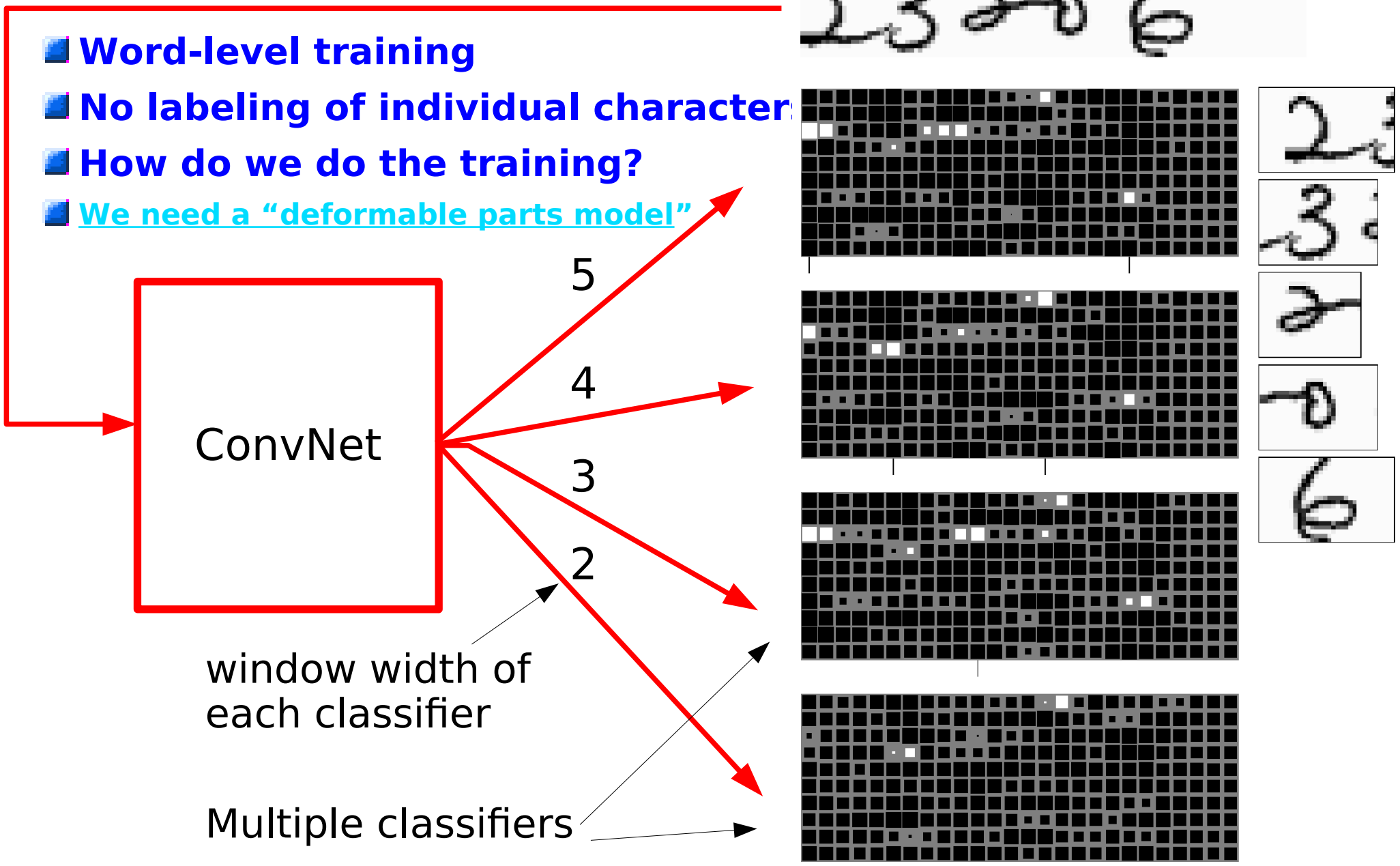
Courant Institute of Mathematical Sciences, NYU

http://yann.lecun.com



Word-level training with weak supervision [Matan et al 1992]

- Word-level training
- No labeling of individual character
- How do we do the training?
- We need a “deformable parts model”



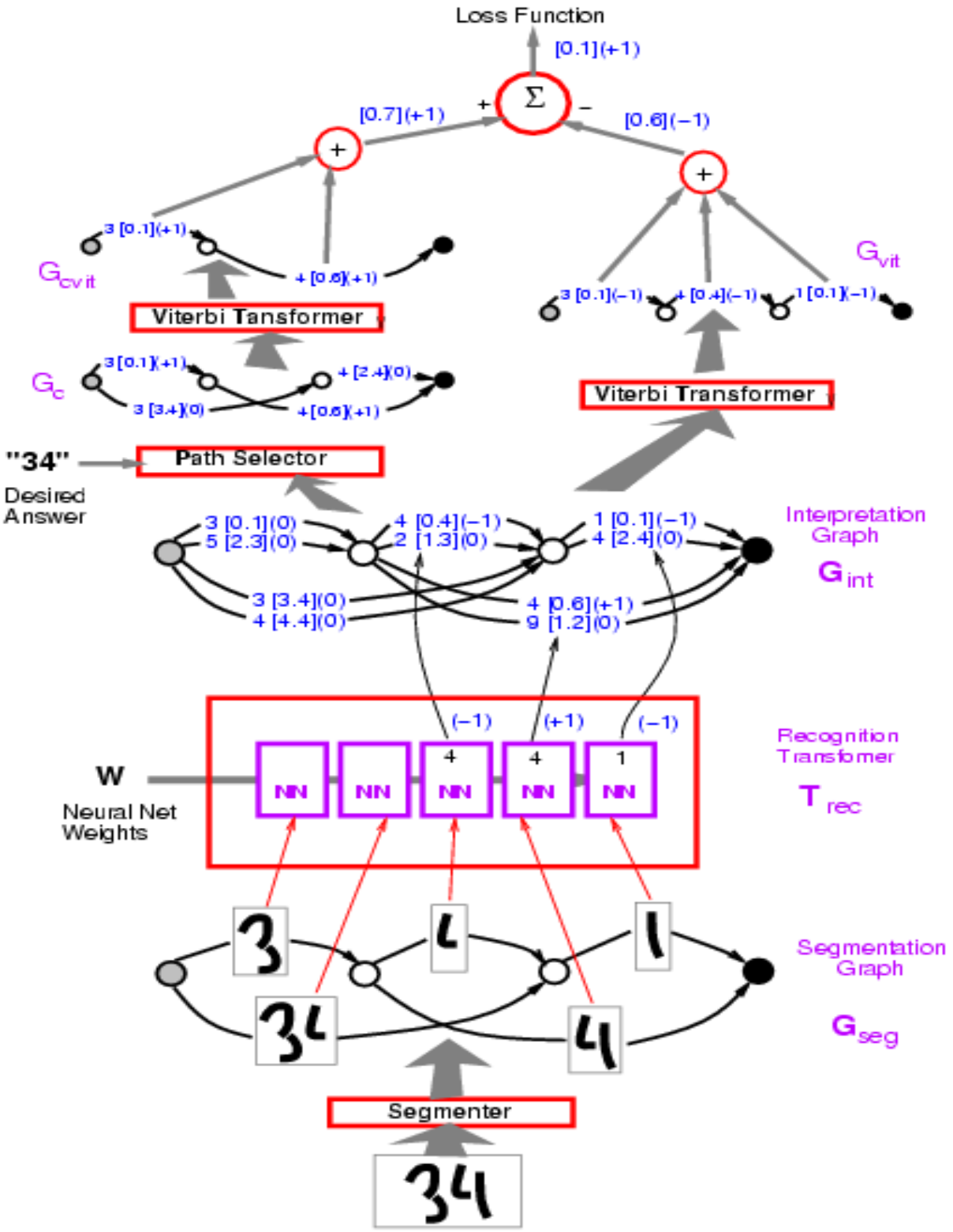
Graph Transformer Networks

Structured Prediction on top of Deep Learning

This example shows the structured perceptron loss.

In practice, we used negative log-likelihood loss.

Deployed in 1996 in check reading machines.



Check Reader (Bell Labs, 1995)

Graph transformer network trained to read check amounts

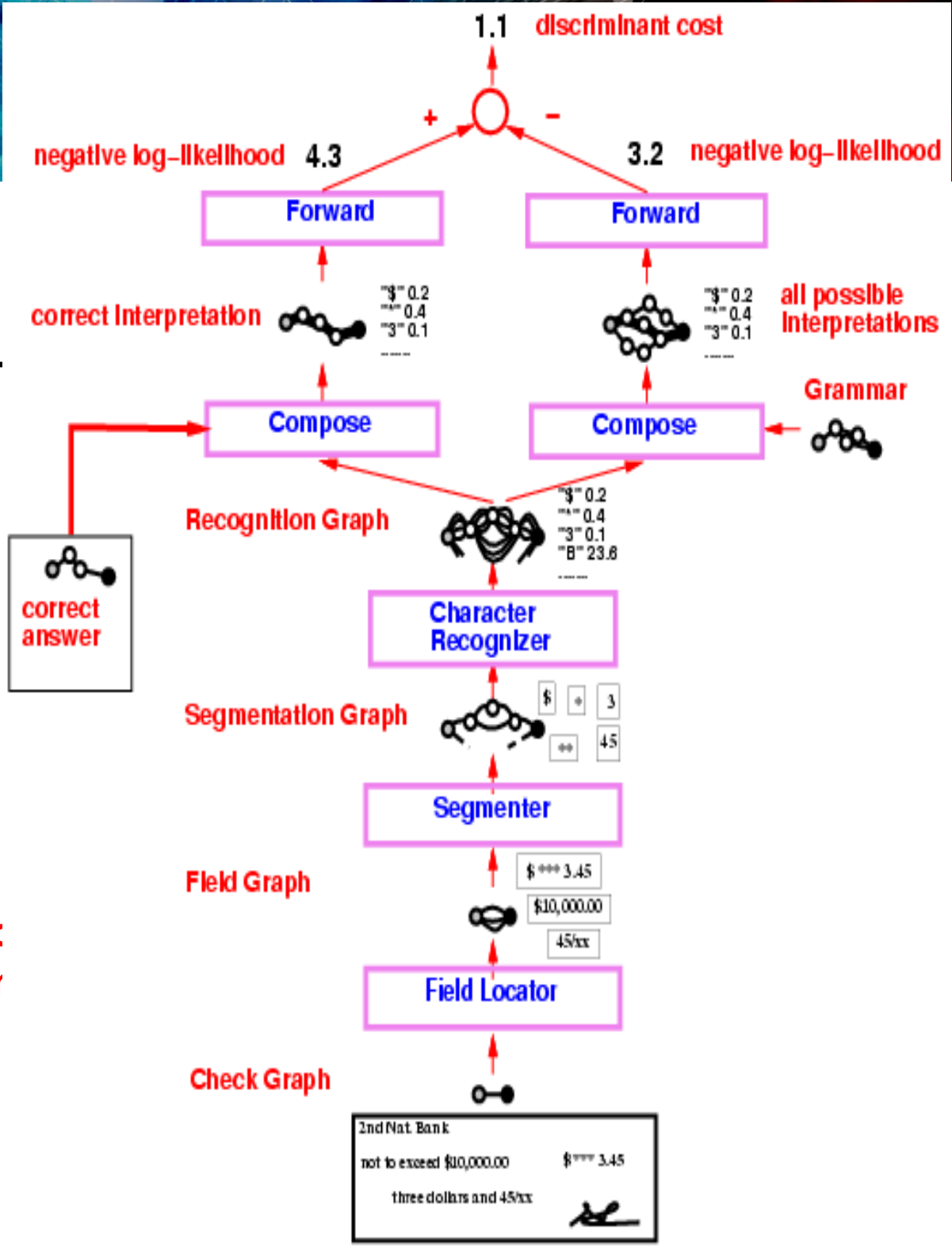
Trained globally with Negative-Log-Likelihood loss.

50% percent correct, 49% reject, 1% error (detectable later in the process).

Fielded in 1996, used in many banks in the US and Europe.

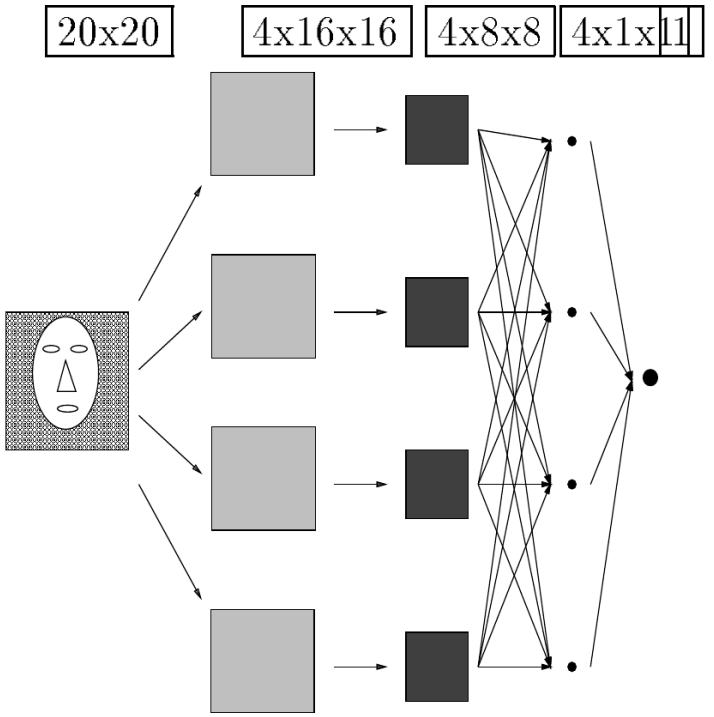
Processed an estimated 10% to 20% of all the checks written in the US in the early 2000s.

[LeCun, Bottou, Bengio, Haffner 1998]



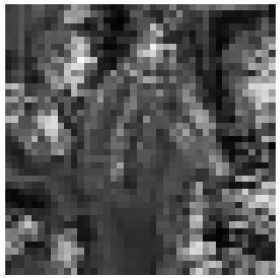
Face Detection [Vaillant et al. 93]

- ConvNet applied to large images
- Heatmaps at multiple scales
- Non-maximum suppression for candidates
- 6 second on a Sparcstation for 256x256

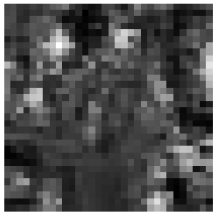


Scale 3

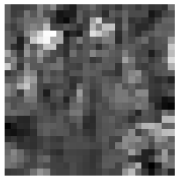
Scale 4



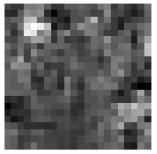
Scale 5



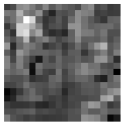
Scale 6



Scale 7



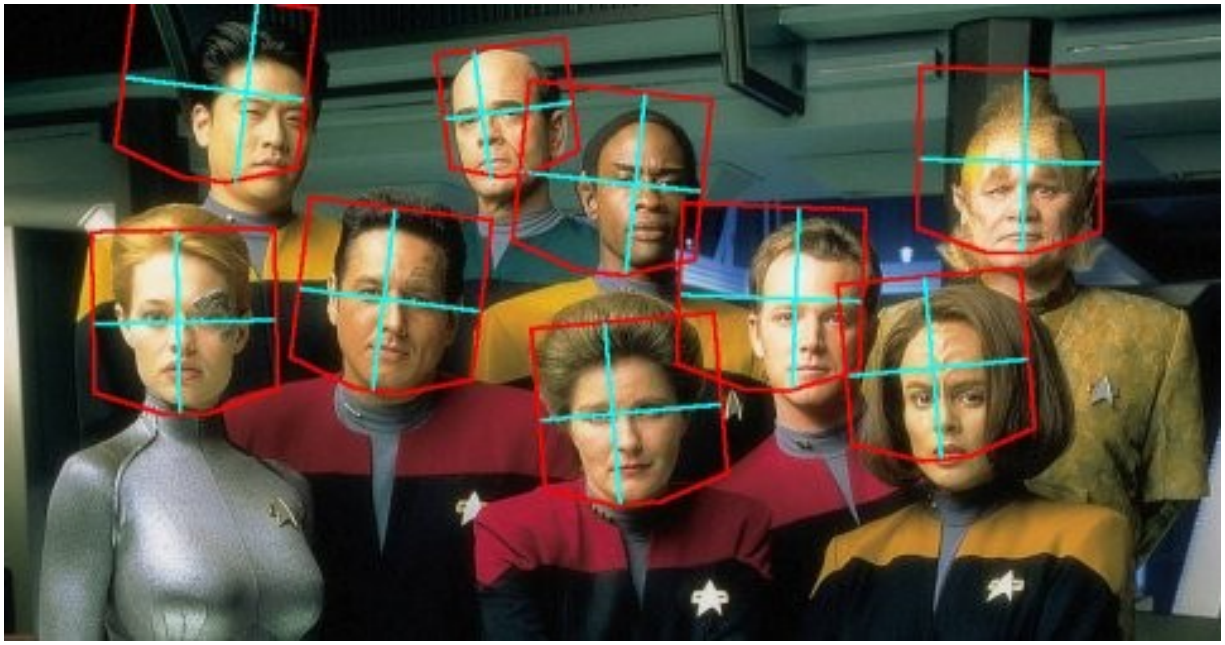
Scale 8

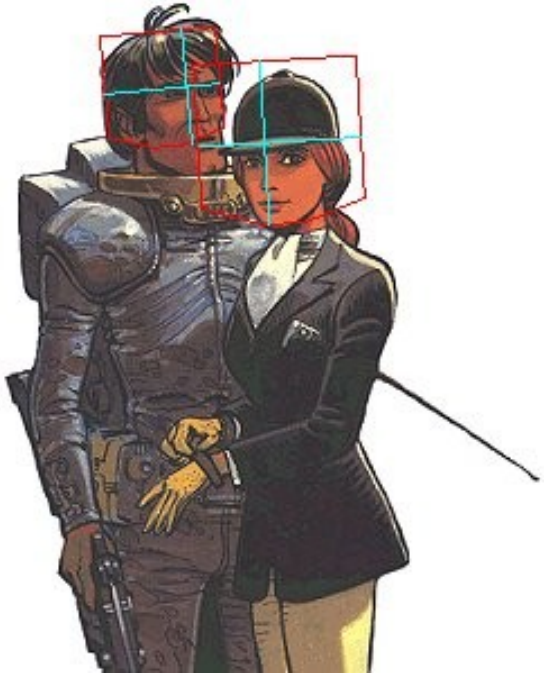


Scale 9

mid 2000s: state of the art results on face detection

Data Set->	TILTED		PROFILE		MIT+CMU	
False positives per image->	4.42	26.9	0.47	3.36	0.5	1.28
Our Detector	90%	97%	67%	83%	83%	88%
Jones & Viola (tilted)	90%	95%	x		x	
Jones & Viola (profile)	x		70%	83%	x	





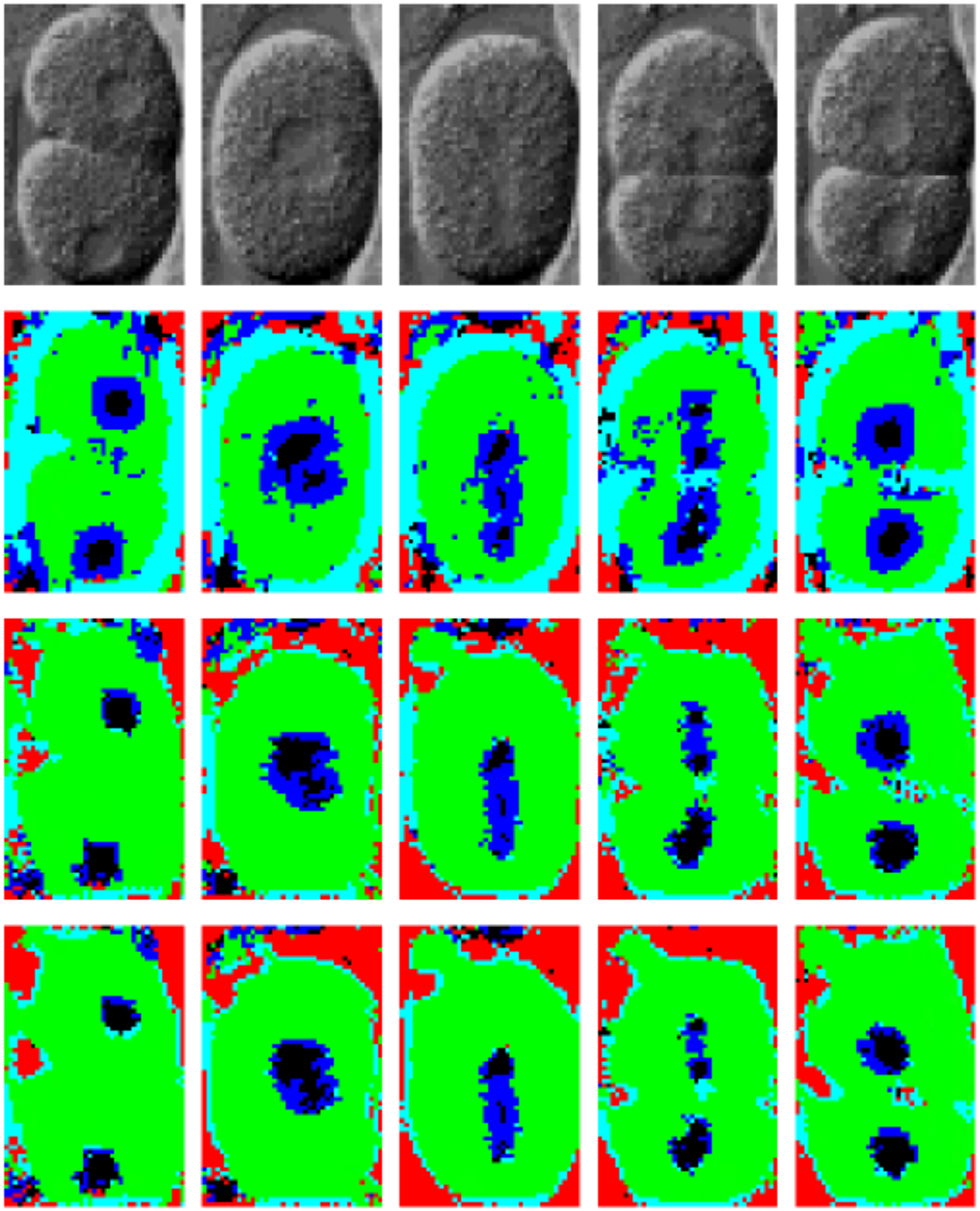
[Garcia & Delakis 2003][Osadchy et al. 2004] [Osadchy et al, JMLR 200

Simultaneous face detection and pose estimation

Semantic Segmentation

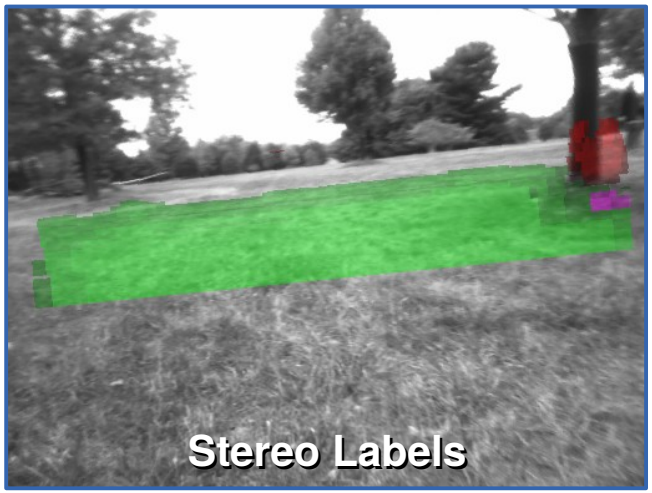
ConvNets for Biological Image Segmentation

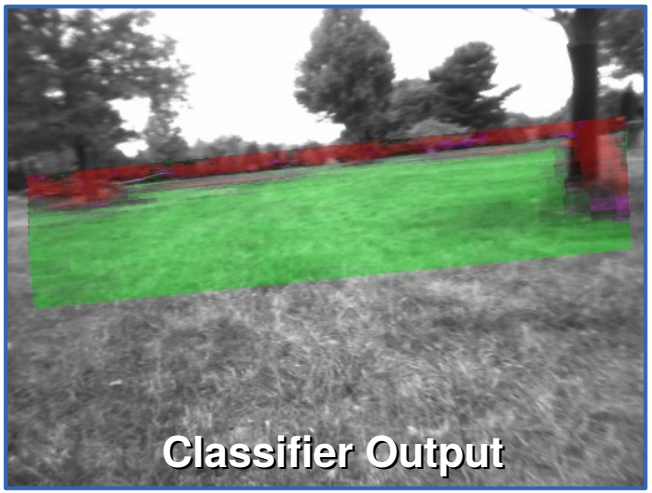
- **Biological Image Segmentation**
  - [Ning et al. IEEE-TIP 2005]
- **Pixel labeling with large context using a convnet**
- **ConvNet takes a window of pixels and produces a label for the central pixel**
- **Cleanup using a kind of conditional random field (CRF)**
  - Similar to a field of expert, but conditional.
- **3D version for connectomics**
  - [Jain et al. 2007]

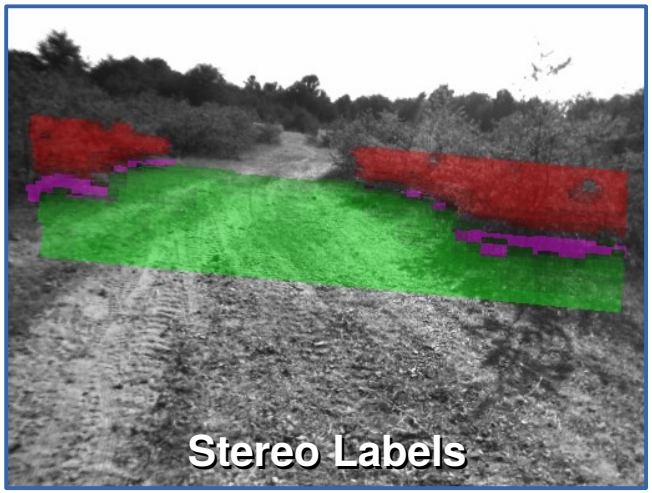


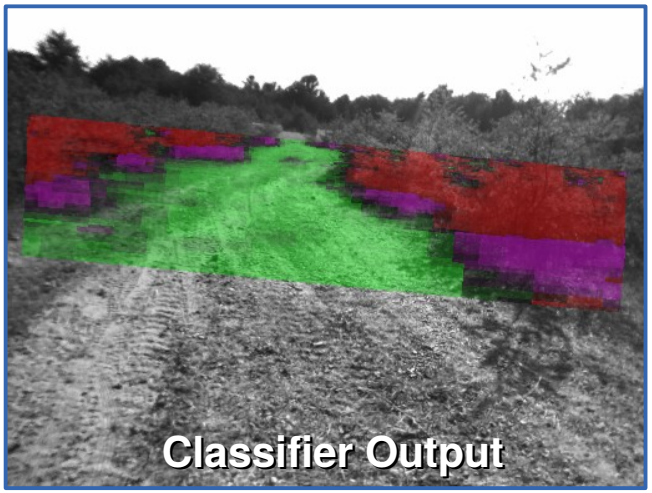
ConvNet for Long Range Adaptive Robot Vision (DARPA LAGR program 2005-2008)

Y LeCun



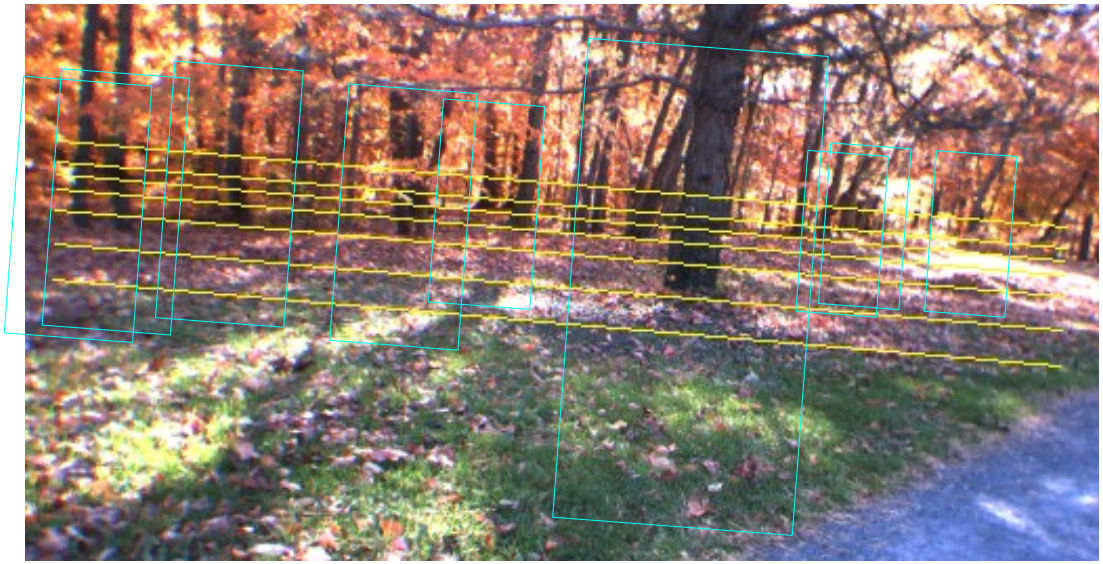






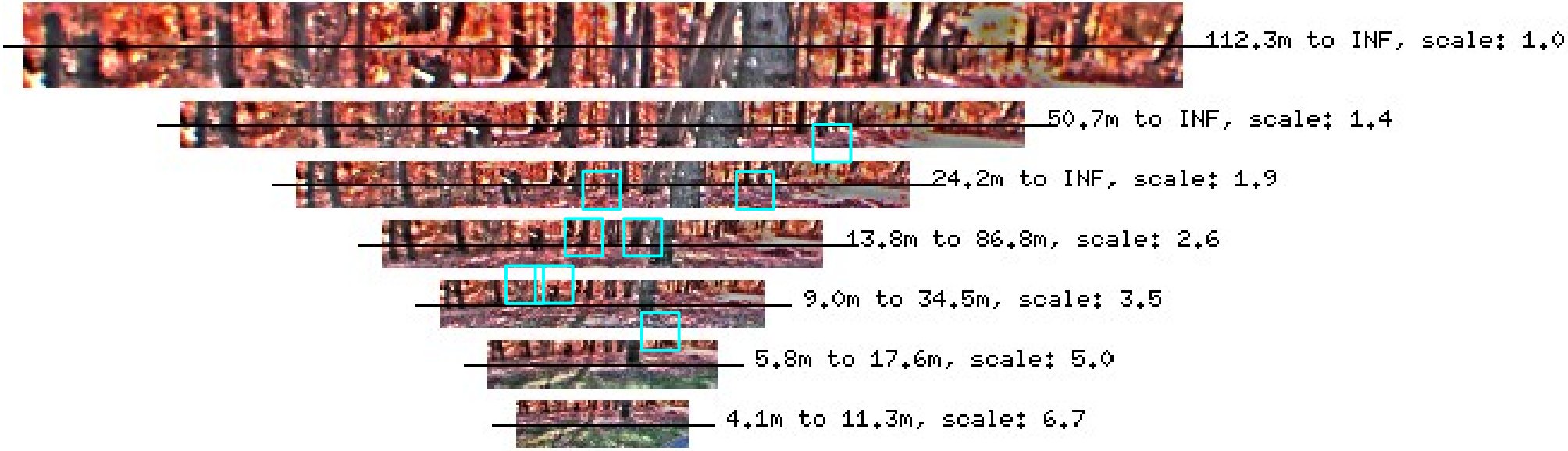
Long Range Vision with a Convolutional Net

Y LeCun



Pre-processing (125 ms)

- Ground plane estimation
- Horizon leveling
- Conversion to YUV + local contrast normalization
- Scale invariant pyramid of distance-normalized image "bands"



Convolutional Net Architecture

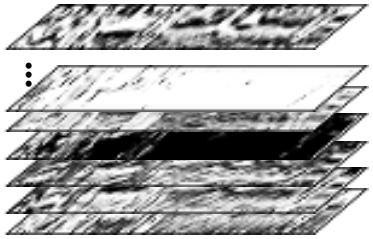
100 features per
3x12x25 input window

100@25x121

CONVOLUTIONS (6x5)

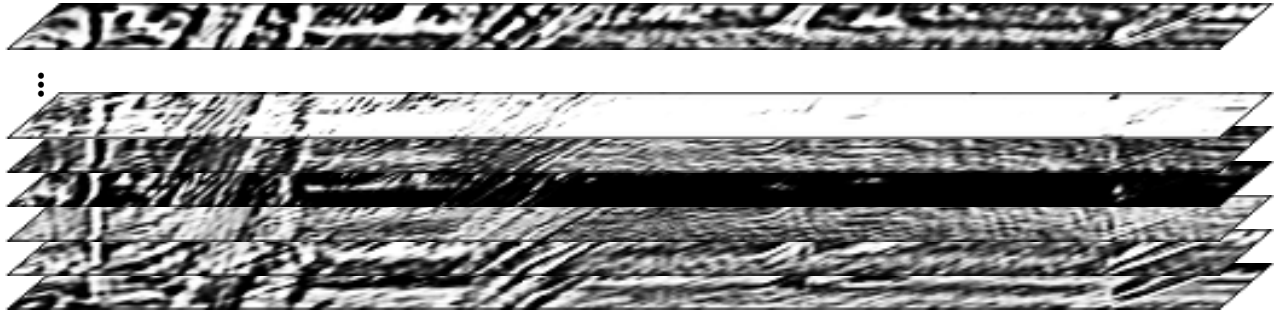
VIDEO: LAGR

20@30x125



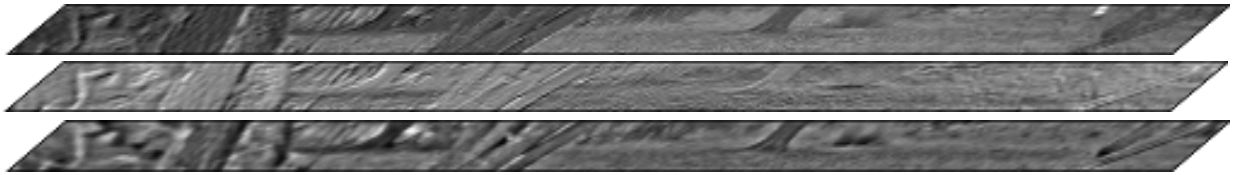
MAX SUBSAMPLING (1x4)

20@30x484



CONVOLUTIONS (7x6)

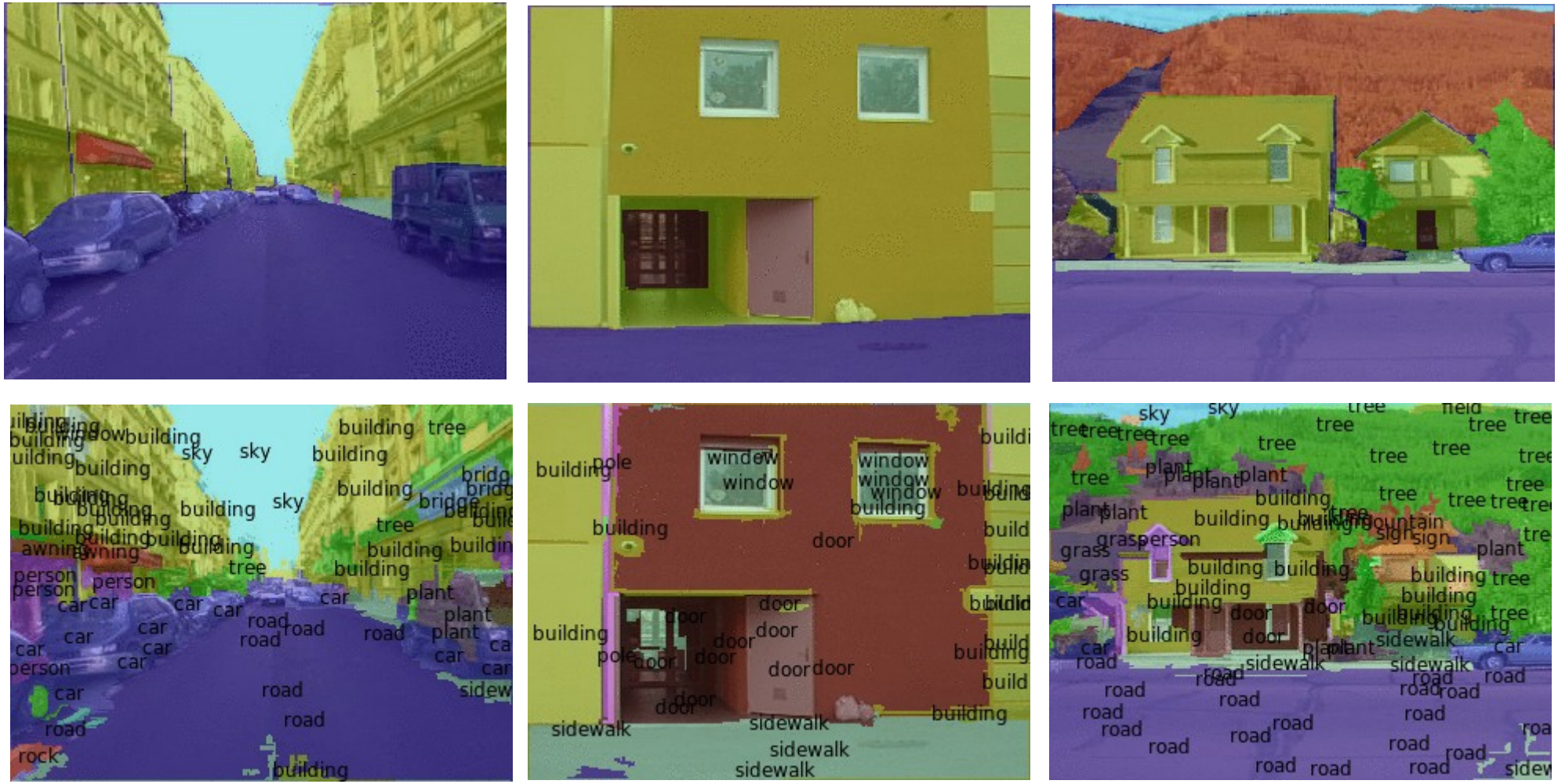
3@36x484



YUV image band
20-36 pixels tall,
36-500 pixels wide

YUV input

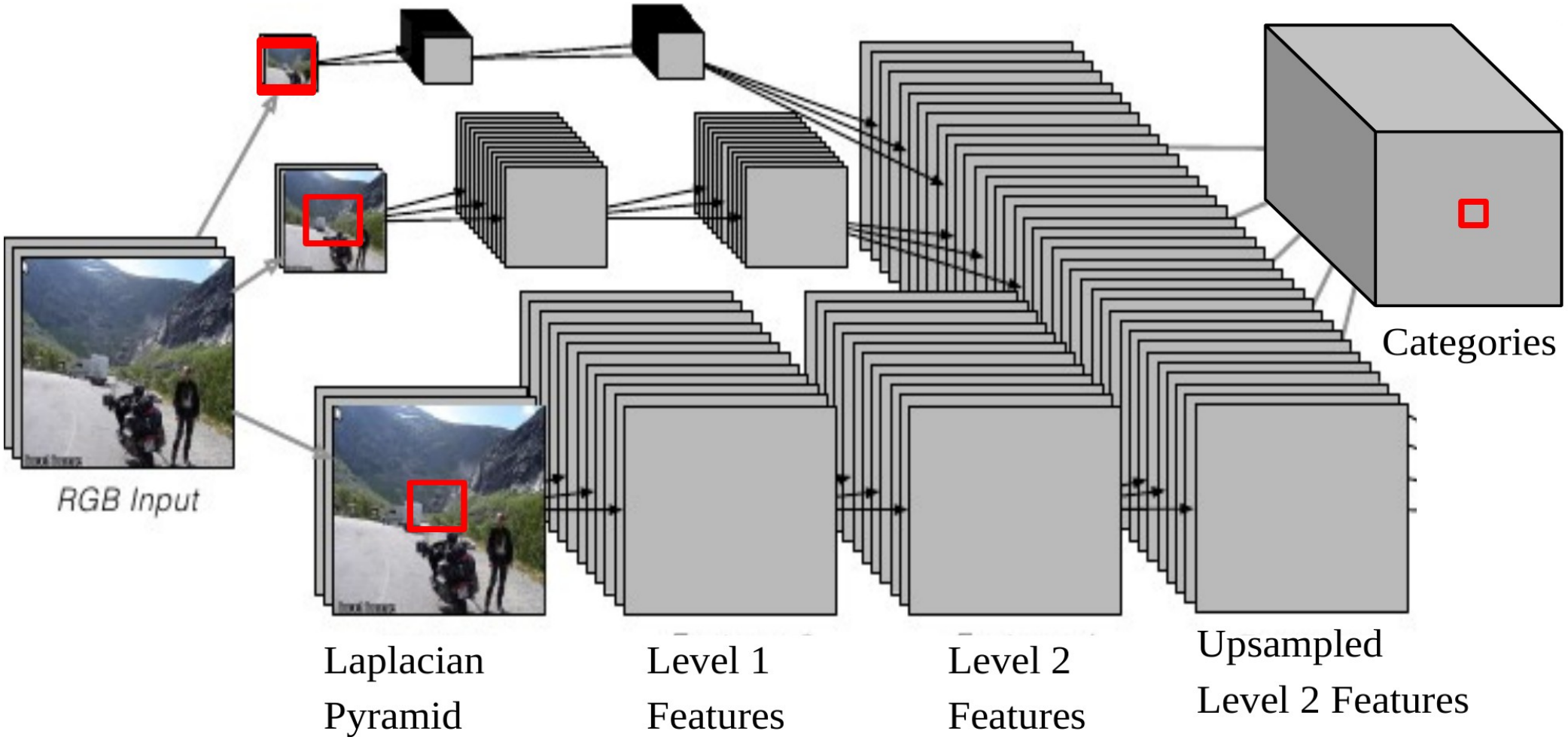
Scene Parsing/Labeling



[Farabet et al. ICML 2012, PAMI 2013]

Scene Parsing/Labeling: Multiscale ConvNet Architecture

- **Each output sees a large input context:**
  - **46x46** window at full rez; **92x92** at ½ rez; **184x184** at ¼ rez
  - [7x7conv]->[2x2pool]->[7x7conv]->[2x2pool]->[7x7conv]->
  - Trained supervised on fully-labeled images



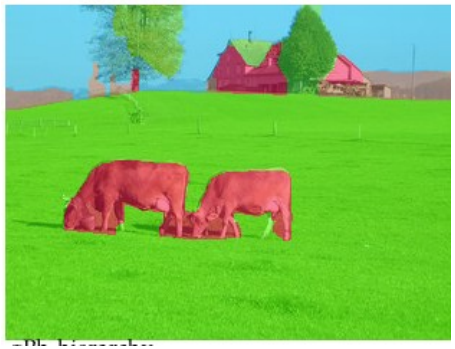
Method 1: majority over super-pixel regions

Input image

Super-pixel boundary hypetheses

Superpixel boundaries

Majority Vote Over Superpixels



Categories aligned With region boundaries

Multi-scale ConvNet

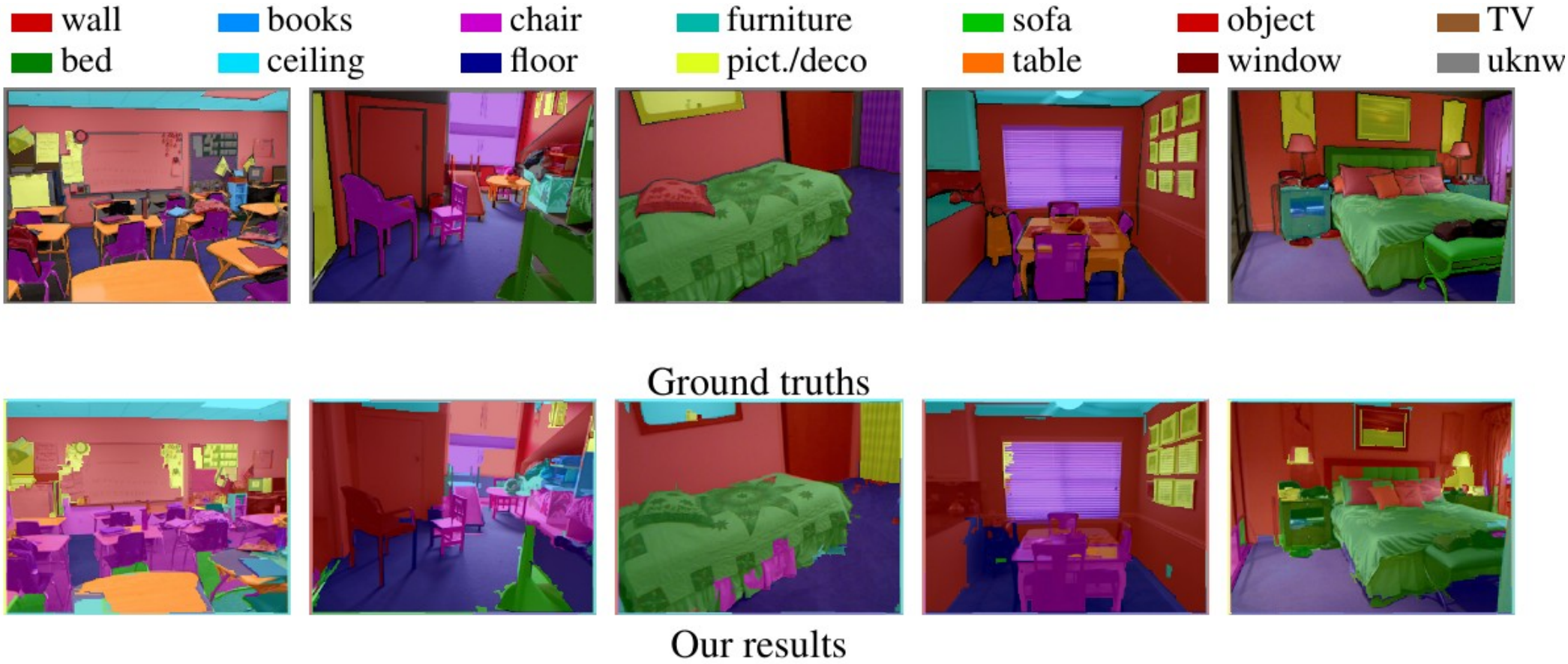
Features from Convolutional net (d=768 per pixel)

Convolutional classifier

“soft” categories scores

[Farabet et al. IEEE T. PAMI 2013]

Scene Parsing/Labeling on RGB+Depth Images



[Couprie, Farabet, Najman, LeCun ICLR 2013, ICIP 2013]

Scene Parsing/Labeling

- **No post-processing**
- **Frame-by-frame**
- **ConvNet runs at 50ms/frame on Virtex-6 FPGA hardware**
  - But communicating the features over ethernet limits system performance

VIDEO: SCENE PARSING

[Farabet et al. ICML 2012, PAMI 2013]

Scene Parsing/Labeling: Performance

- **Stanford Background Dataset [Gould 1009]: 8 categories**

	Pixel Acc.	Class Acc.	CT (sec.)
Gould *et al.* 2009 [14]	76.4%	-	10 to 600s
Munoz *et al.* 2010 [32]	76.9%	66.2%	12s
Tighe *et al.* 2010 [46]	77.5%	-	10 to 300s
Socher *et al.* 2011 [45]	78.1%	-	?
Kumar *et al.* 2010 [22]	79.4%	-	$< 600s$
Lempitzky *et al.* 2011 [28]	**81.9%**	72.4%	$> 60s$
singlescale convnet	66.0 %	56.5 %	0.35s
multiscale convnet	78.8 %	72.4%	0.6s
multiscale net + superpixels	80.4%	74.56%	**0.7s**
multiscale net + gPb + cover	80.4%	75.24%	61s
multiscale net + CRF on gPb	81.4%	**76.0%**	60.5s

[Rejected from CVPR 2012]

[Farabet et al. ICML 2012][Farabet et al. IEEE T. PAMI 2013]

Scene Parsing/Labeling: Performance

	Pixel Acc.	Class Acc.
Liu *et al.* 2009 [31]	74.75%	-
Tighe *et al.* 2010 [44]	76.9%	29.4%
raw multiscale net[1]	67.9%	45.9%
multiscale net + superpixels[1]	71.9%	**50.8%**
multiscale net + cover[1]	72.3%	**50.8%**
multiscale net + cover[2]	**78.5%**	29.6%

- **SIFT Flow Dataset**
- **[Liu 2009]:**
- **33 categories**

- Barcelona dataset
- [Tighe 2010]:
- 170 categories.

	Pixel Acc.	Class Acc.
Tighe *et al.* 2010 [44]	66.9%	7.6%
raw multiscale net[1]	37.8%	**12.1%**
multiscale net + superpixels[1]	44.1%	**12.4%**
multiscale net + cover[1]	46.4%	**12.5%**
multiscale net + cover[2]	**67.8%**	**9.5%**

[Farabet et al. IEEE T. PAMI 2012]