Vision

Image Recognition and Understanding

- Almost all modern image understanding systems use ConvNets.
- Google, Facebook, Microsoft, IBM, Baidu, Yahoo/Flickr, Adobe, Yandex, Wechat, NEC, NVIDIA, MobilEye, Qualcomm..... Everyone uses ConvNets
- Each of the 700 Million photos uploaded on Facebook every day goes through two ConvNets:
  - 1 for object recognition, 1 for face recognition.
- The Tesla autopilot uses a ConvNet
- All the hardware companies are tuning their chips for running ConvNets
  - NVIDIA, Intel, MobilEye, Qualcomm, Samsung.....

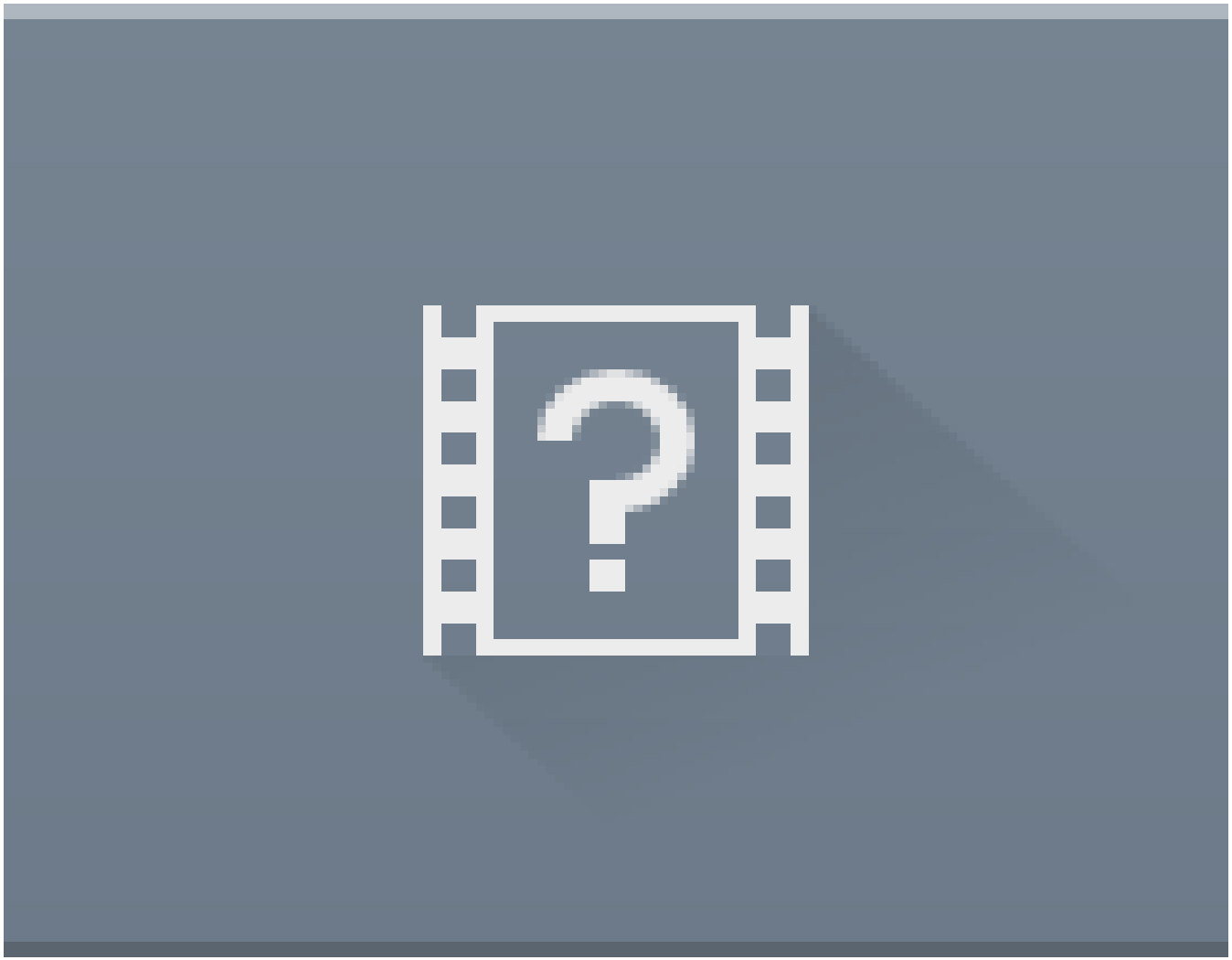
Simultaneous face detection and pose estimation (2003)

Y LeCun

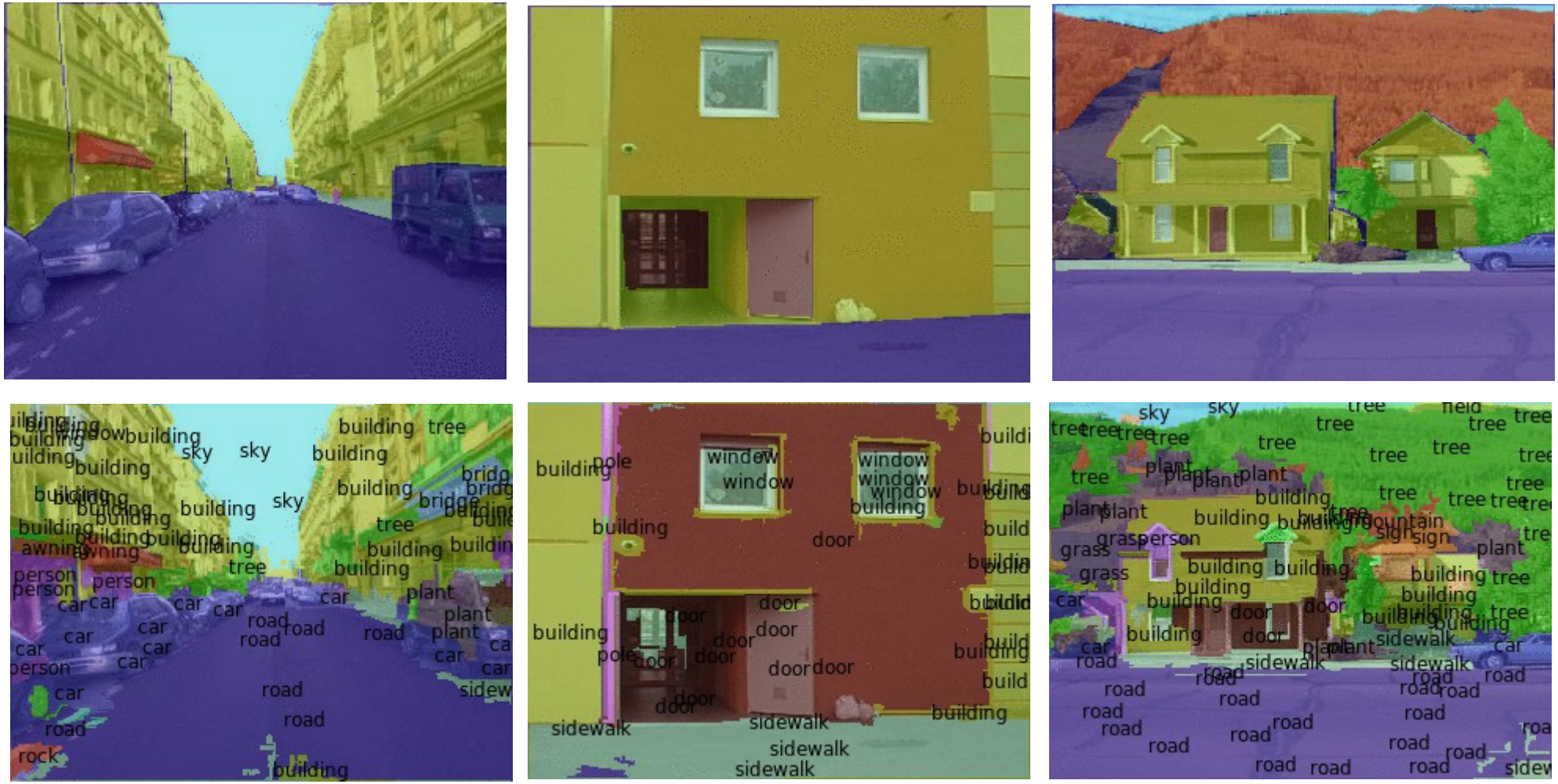
Pedestrian Detection

ConvNet in Connectomics [Jain, Turaga, Seung 2007-present]

- 3D ConvNet
  - Volumetric
  - Images
- Each voxel labeled as "membrane" or "non-membrane using a 7x7x7 voxel neighborhood
- Has become a standard method in connectomics



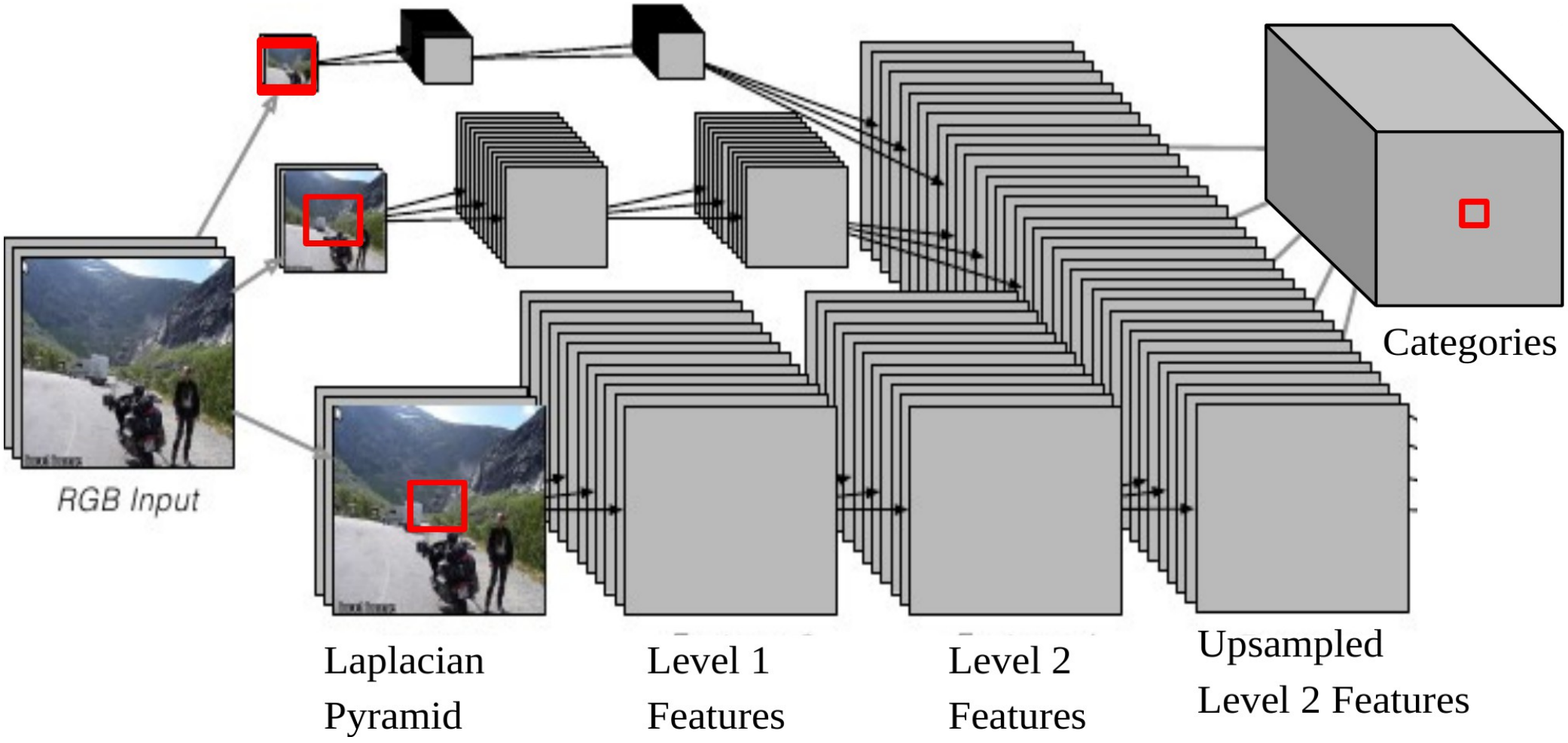
Scene Parsing/Labeling



[Farabet et al. ICML 2012, PAMI 2013]

Scene Parsing/Labeling: Multiscale ConvNet Architecture

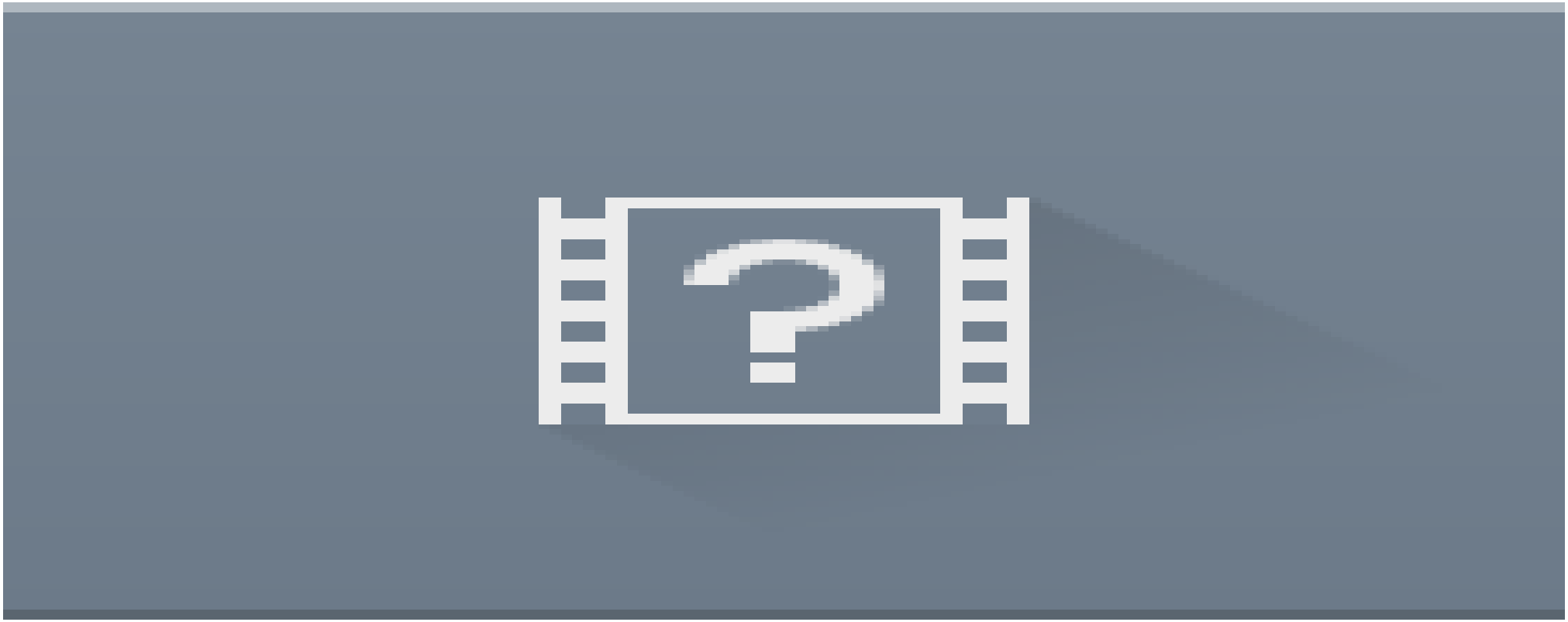
- Each output sees a large input context:
  - **46x46** window at full rez; **92x92** at ½ rez; **184x184** at ¼ rez
  - [7x7conv]->[2x2pool]->[7x7conv]->[2x2pool]->[7x7conv]->
  - Trained supervised on fully-labeled images



Scene Parsing/Labeling

[Farabet et al. ICML 2012, PAMI 2013]

Scene Parsing/Labeling

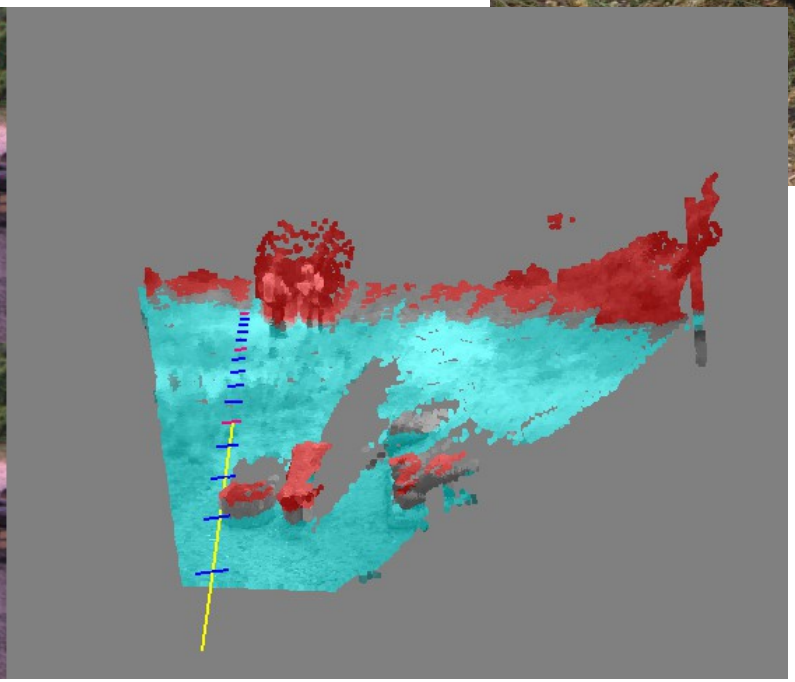


- No post-processing
- Frame-by-frame
- ConvNet runs at 50ms/frame on Virtex-6 FPGA hardware
  - But communicating the features over ethernet limits system performance

vision-based navigation for off-road robot (LAGR Project 2005-2009)

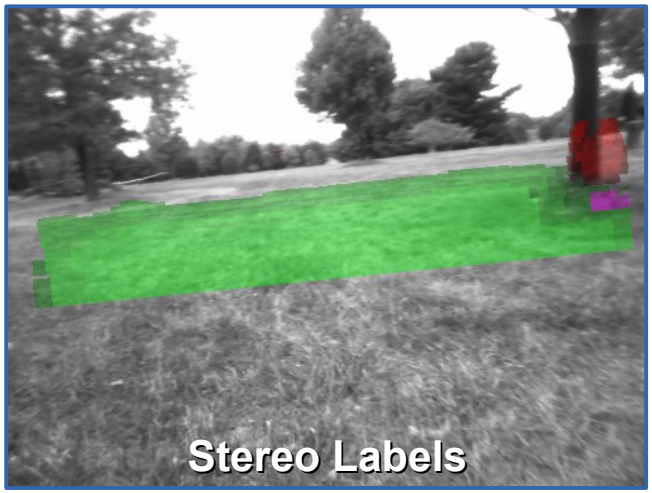
Y LeCun

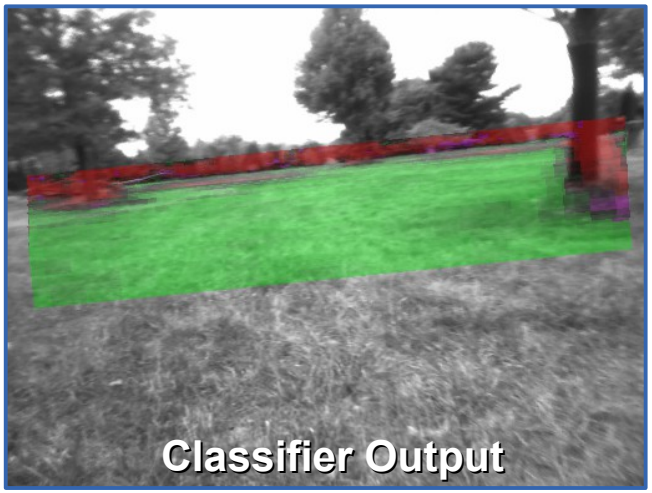
- **Getting a robot to drive autonomously in unknown terrain**
  - solely from vision (camera input).

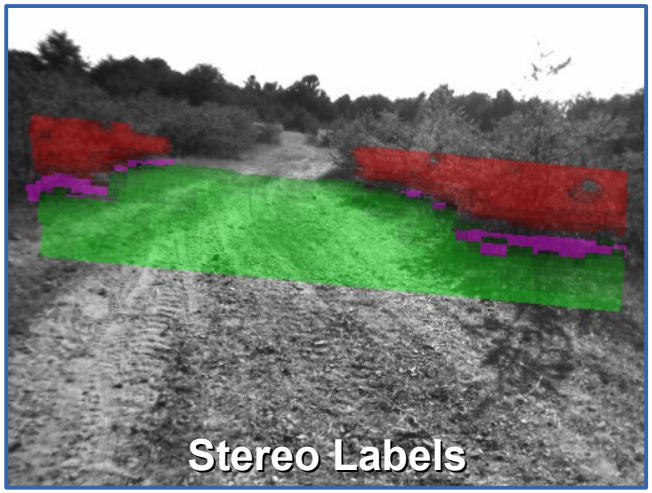


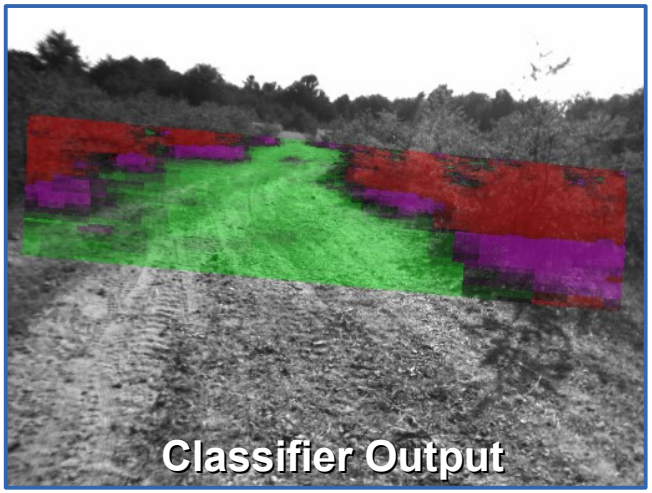
ConvNet for Long Range Adaptive Robot Vision (DARPA LAGR program 2005-2008)

Y LeCun



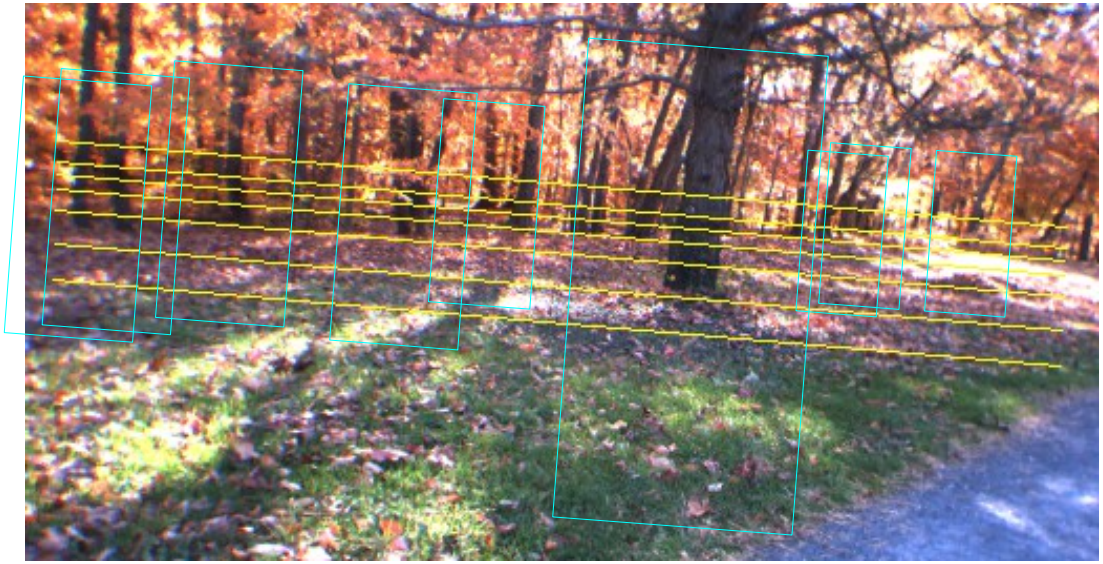






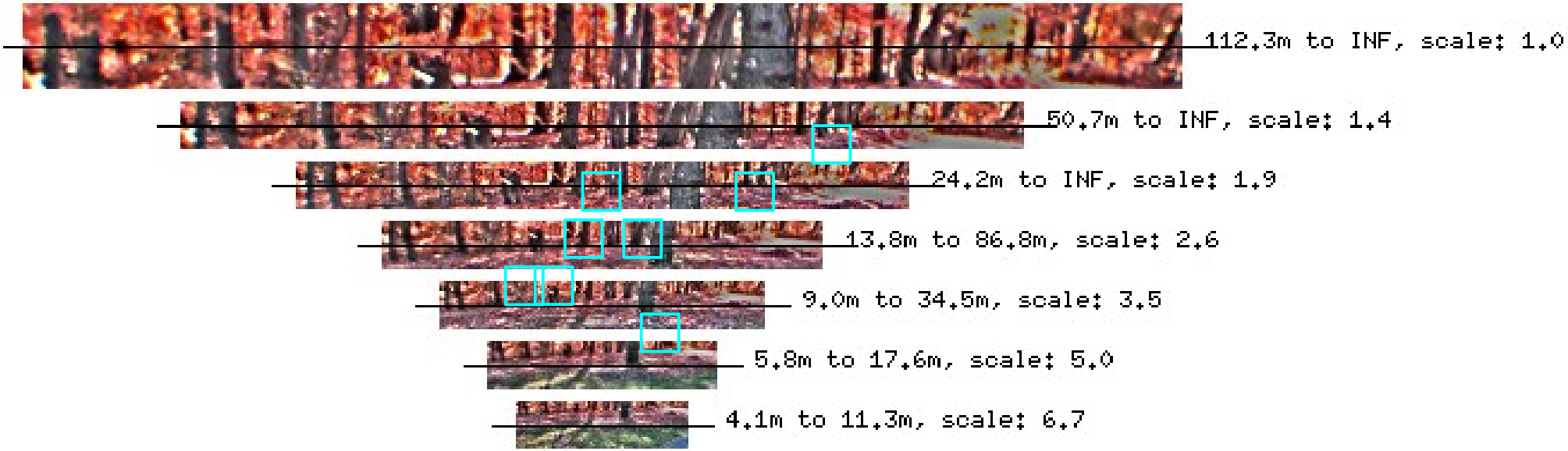
Long Range Vision with a Convolutional Net

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

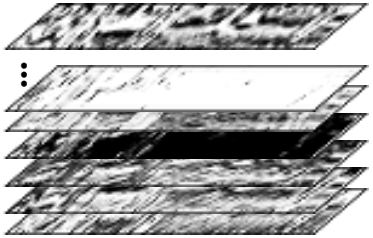
100 features per
3x12x25 input window

100@25x121

CONVOLUTIONS (6x5)

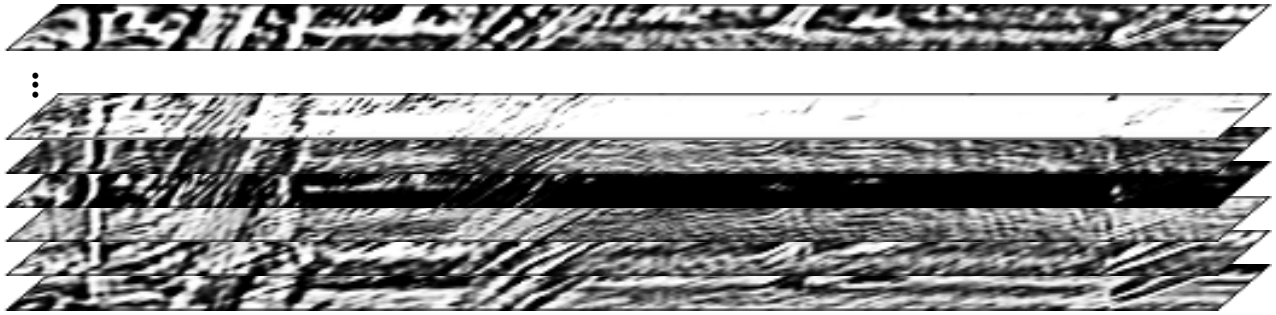
YUV image band
20-36 pixels tall,
36-500 pixels wide

20@30x125



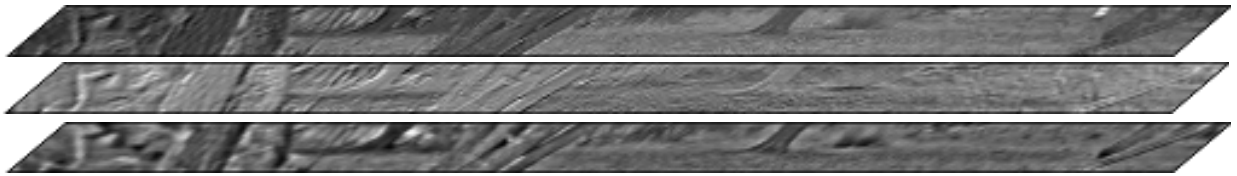
MAX SUBSAMPLING (1x4)

20@30x484



CONVOLUTIONS (7x6)

3@36x484



YUV input

Then in 2011, two things happened...

- The ImageNet dataset [Fei-Fei et al. 2012]
  - 1.2 million training samples
  - 1000 categories

- Fast Graphical Processing Units (GPU)
  - Capable of over 1 trillion operations/second

Matchstick

Flute

Backpack

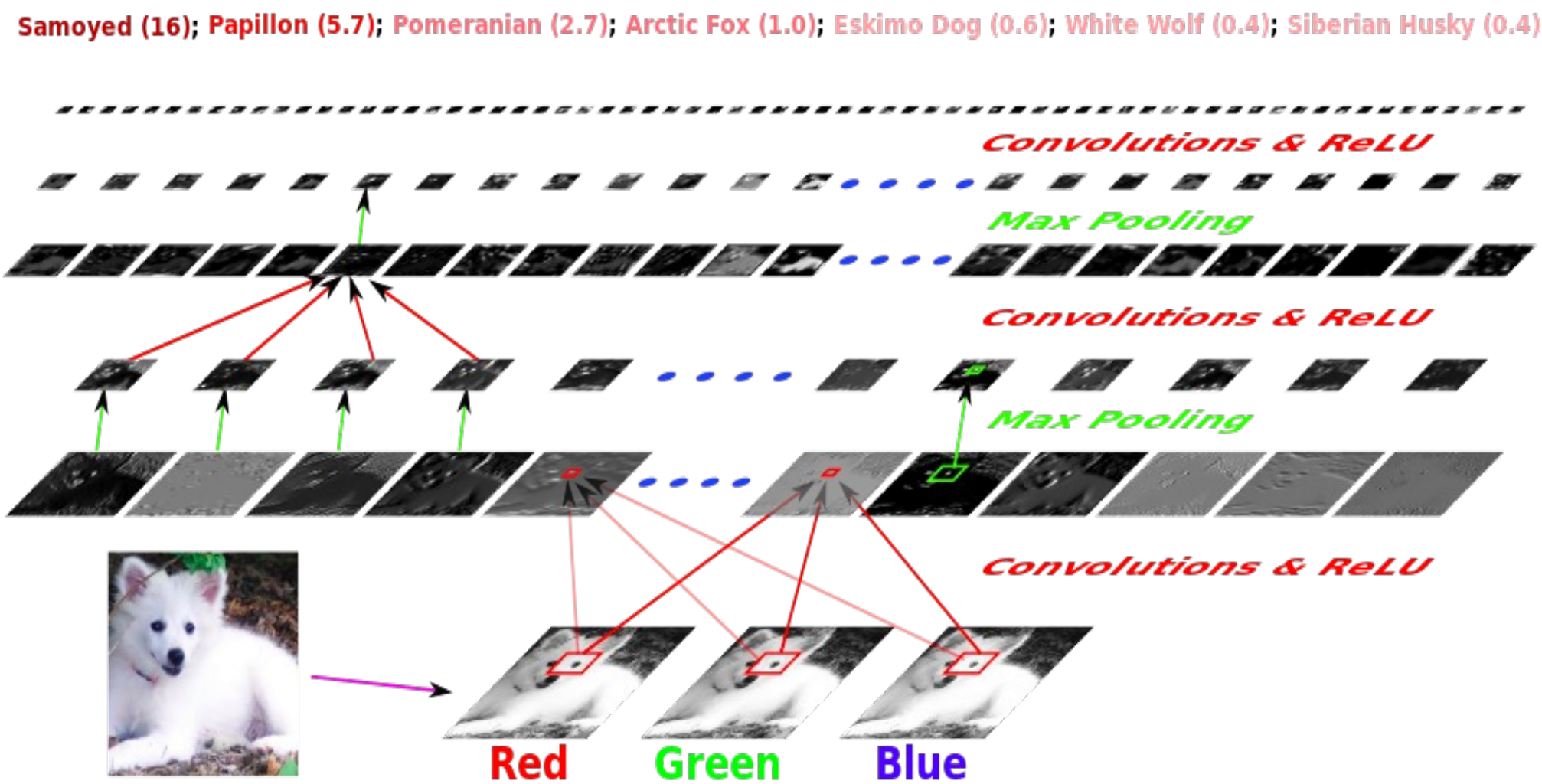
Sea lion

Strawberry

Bathing cap

Racket

Very Deep ConvNet for Object Recognition



ImageNet: Classification

- Give the name of the dominant object in the image
- Top-5 error rates: if correct class is not in top 5, count as error
  - Red:ConvNet, blue: no ConvNet

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

Learning in Action

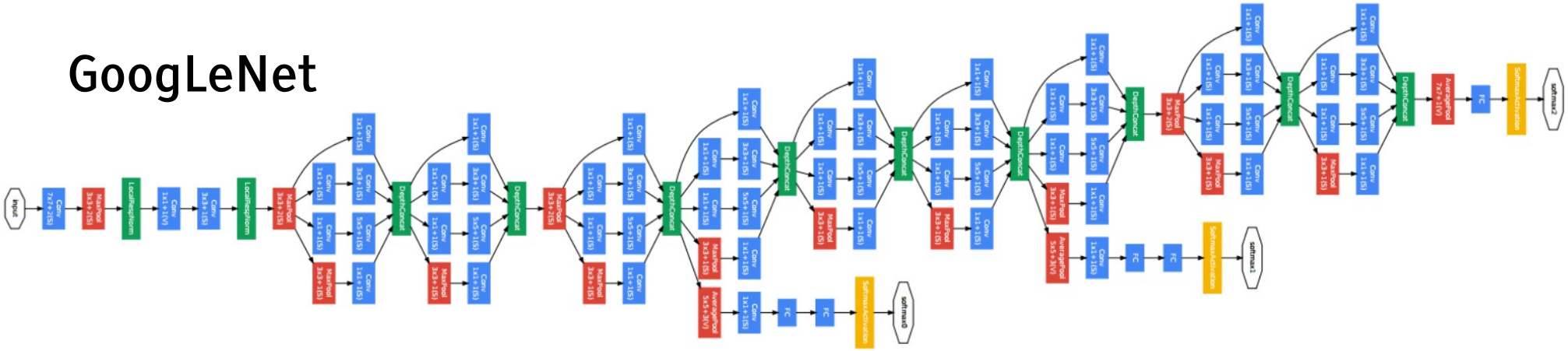
- **How the filters in the first layer learn**

Very Deep ConvNet Architectures

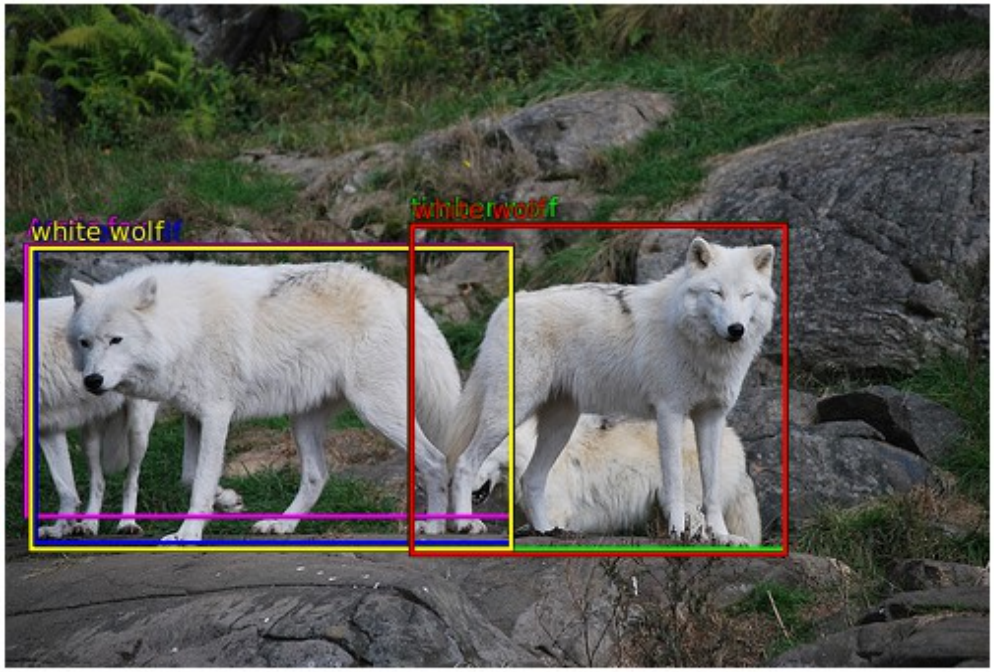
- Small kernels, not much subsampling (fractional subsampling).

VGG

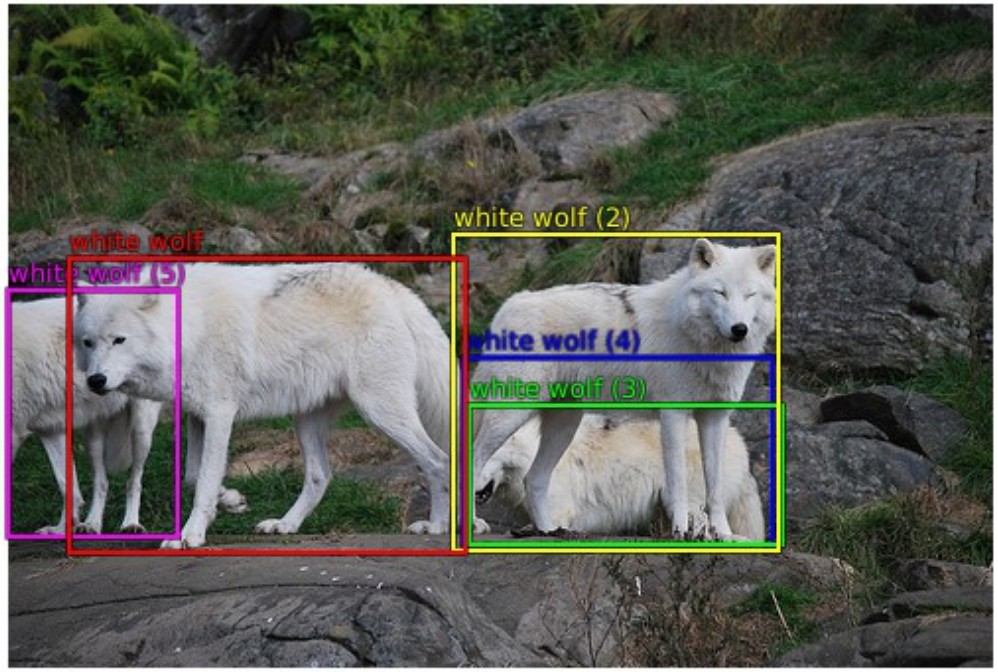
GoogLeNet



Classification+Localization. Results



Top 5:
white wolf
white wolf
timber wolf
timber wolf
Arctic fox

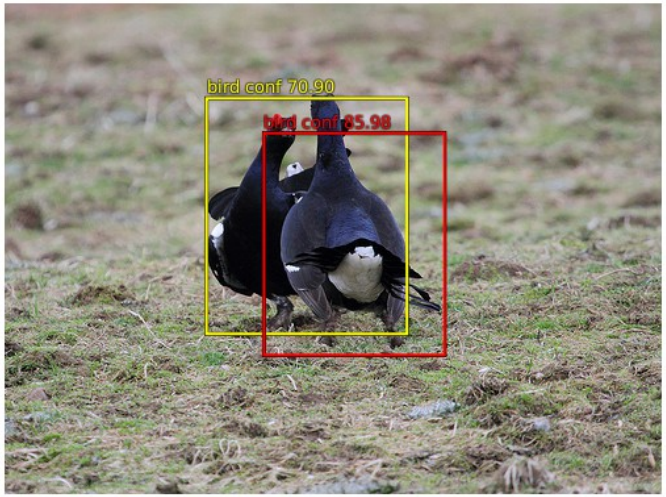


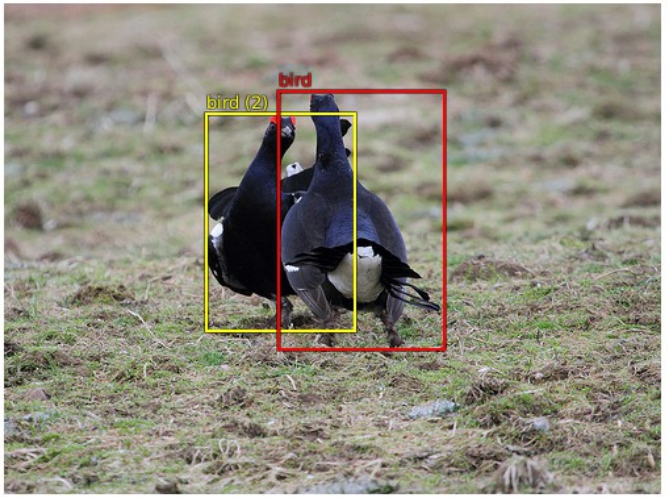
Groundtruth:
white wolf
white wolf (2)
white wolf (3)
white wolf (4)
white wolf (5)

ILSVRC2012_val_00000027.JPEG

Detection: Examples

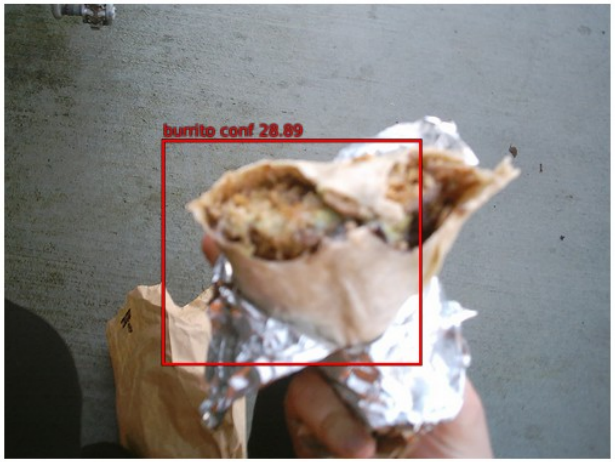
- 200 broad categories
- There is a penalty for false positives
- Some examples are easy some are impossible/ambiguous
- Some classes are well detected
  - Burritos?



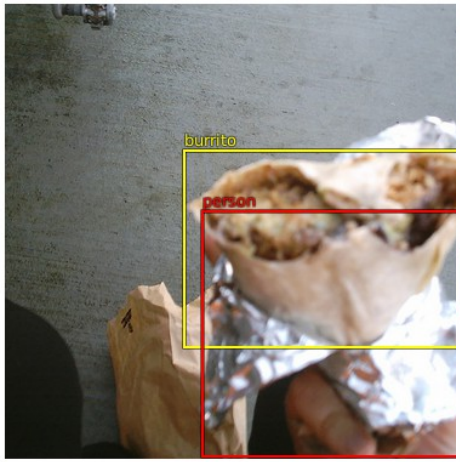


Top predictions:
bird (confidence 86.(
bird (confidence 70.9
ILSVRC2012_val_00001136.JPEG

Groundtruth:



Top predictions:
burrito (confidence 28.9)
ILSVRC2012_val_00000572.JPEG



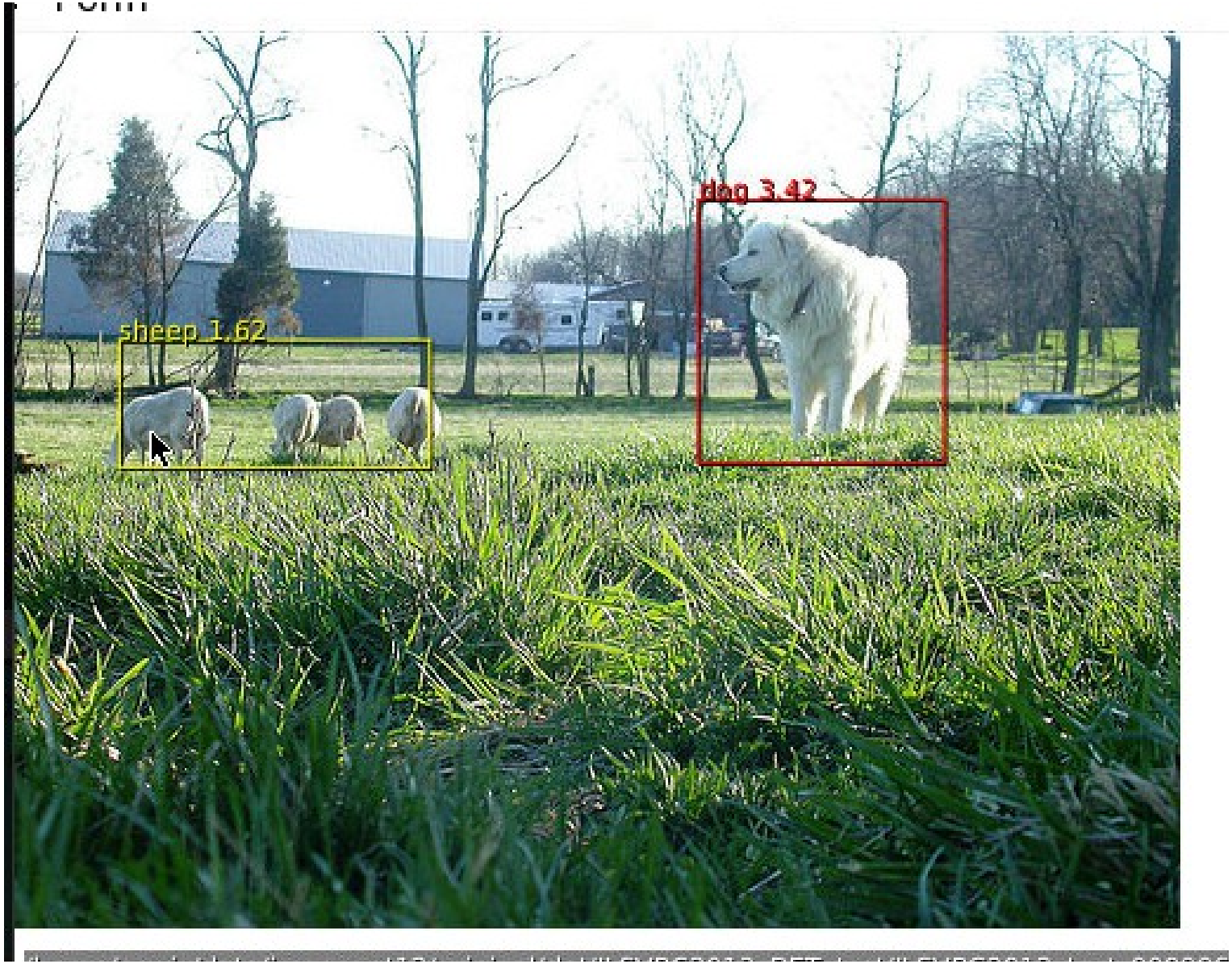
Groundtruth:
person
burrito

Top predictions:
burrito (confidence 17.4)
ILSVRC2012_val_00000606.JPEG

Groundtruth:
burrito
burrito (2)

Detection Examples

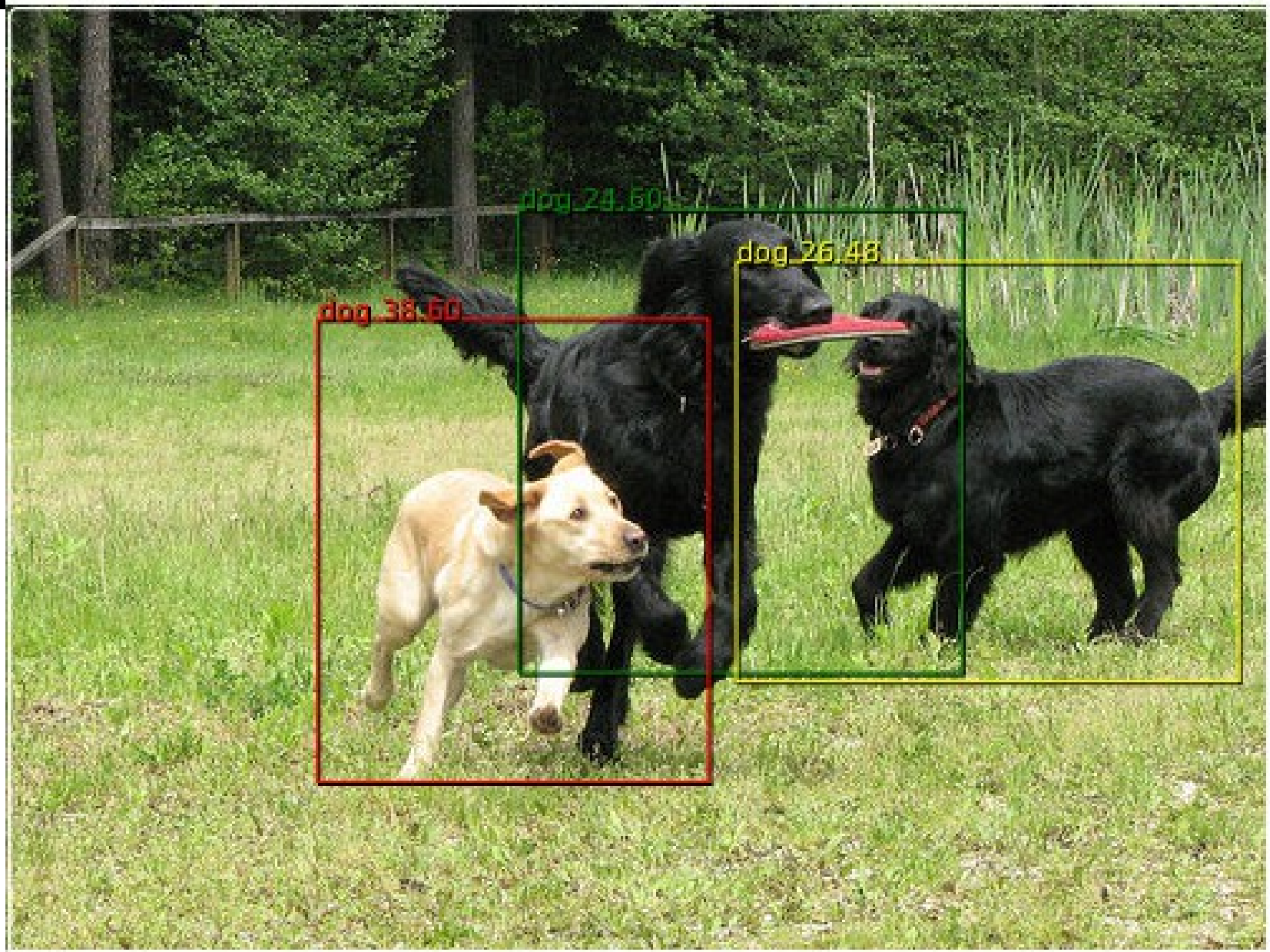
Detection Examples



Detection Examples

/home/snwiz/data/imagenet12/original/det/ILSVRC2013_DET_test/ILSVRC2012_test_00091048.JPEG
person conf 17.898635
bow conf 15.628116

Detection Examples

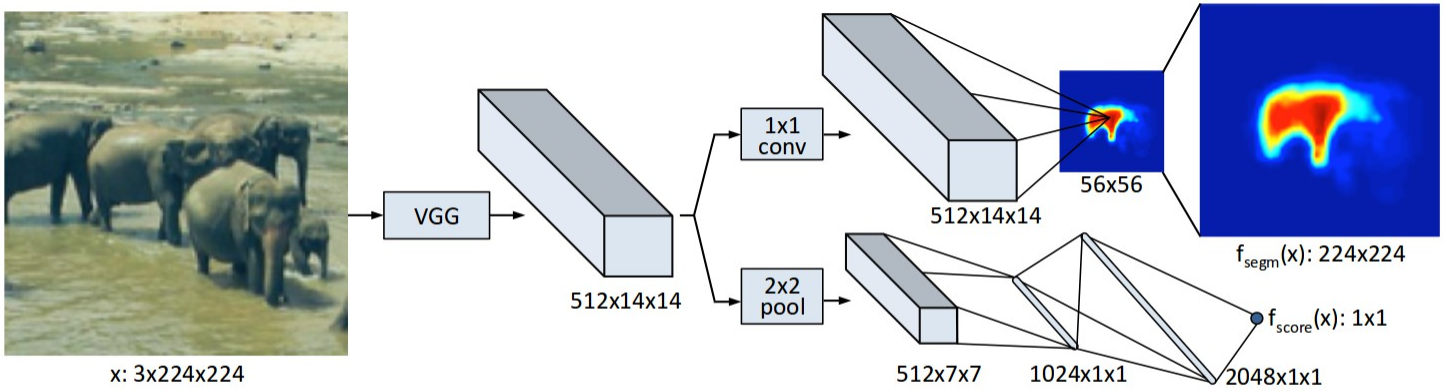


/home/snwiz/data/imagenet12/original/det/ILSVRC2013_DET_test/ILSVRC2012_test_00000172.JPEG
dog conf 38.603936


WE PROVIDI
THIS SERVICE
AT YOUR
SERVICE
person 9.06
french horn 21.89

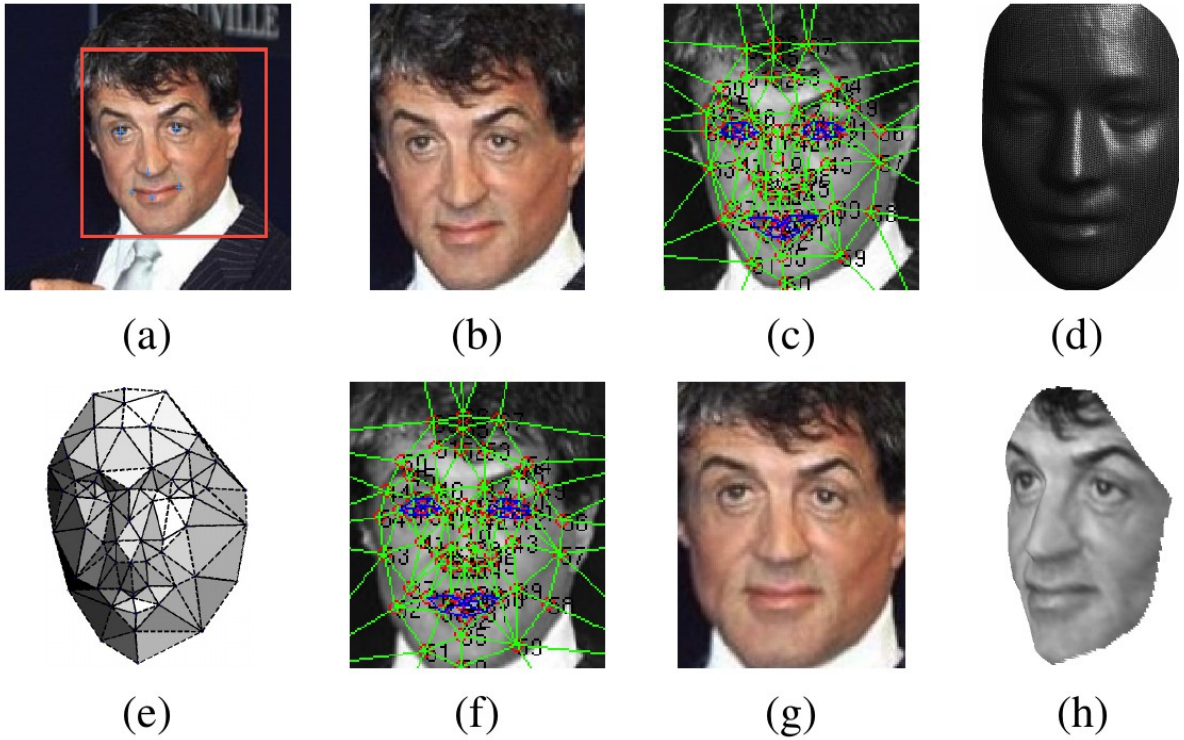
Segmenting and Localizing Objects

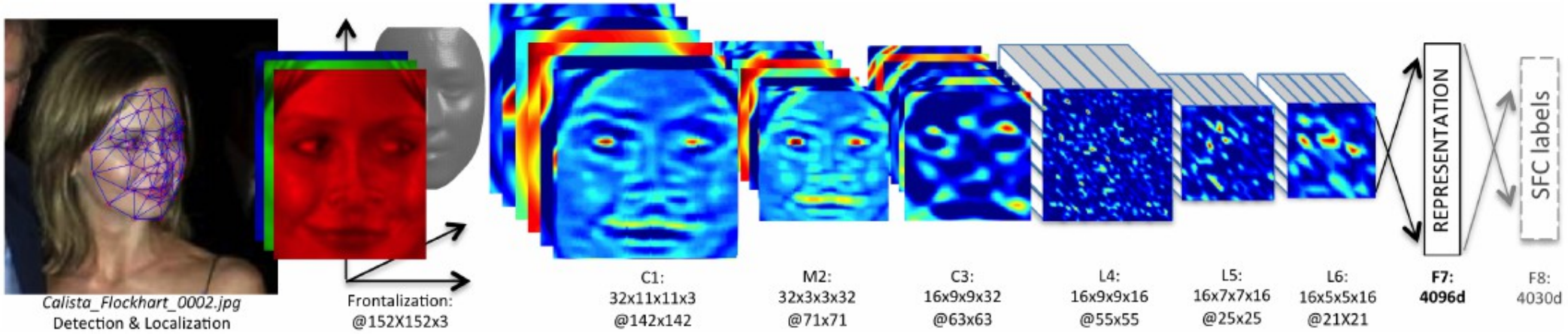
- [Pinheiro, Collobert, Dollar ICCV 2015]
  - ConvNet produces object masks



Deep Face

- **[Taigman et al. CVPR 2014]**
  - Alignment
  - ConvNet
  - Metric Learning
- **Deployed at Facebook for Aut**
  - 600 million photos per day



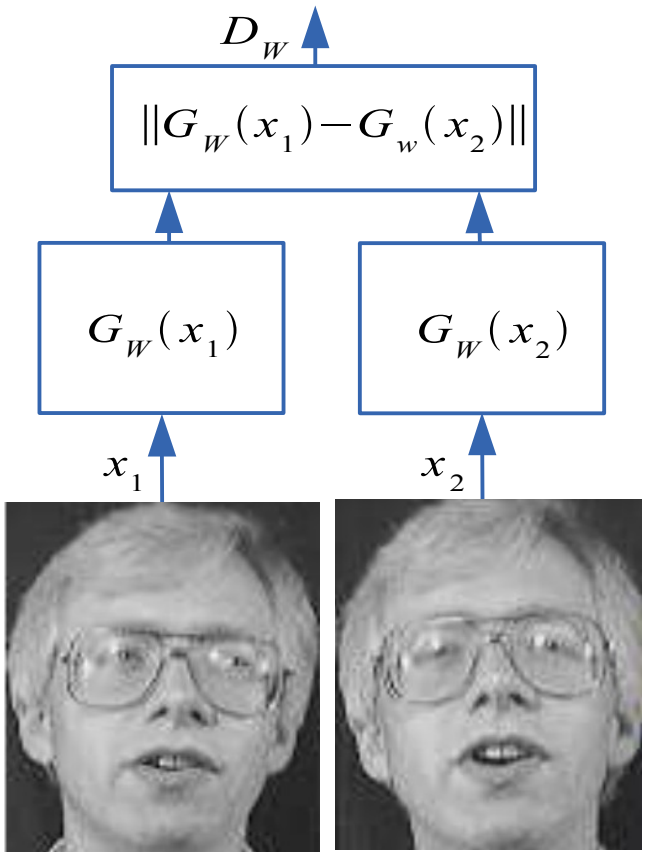


Siamese Architecture and loss function

Contrative Obective Function

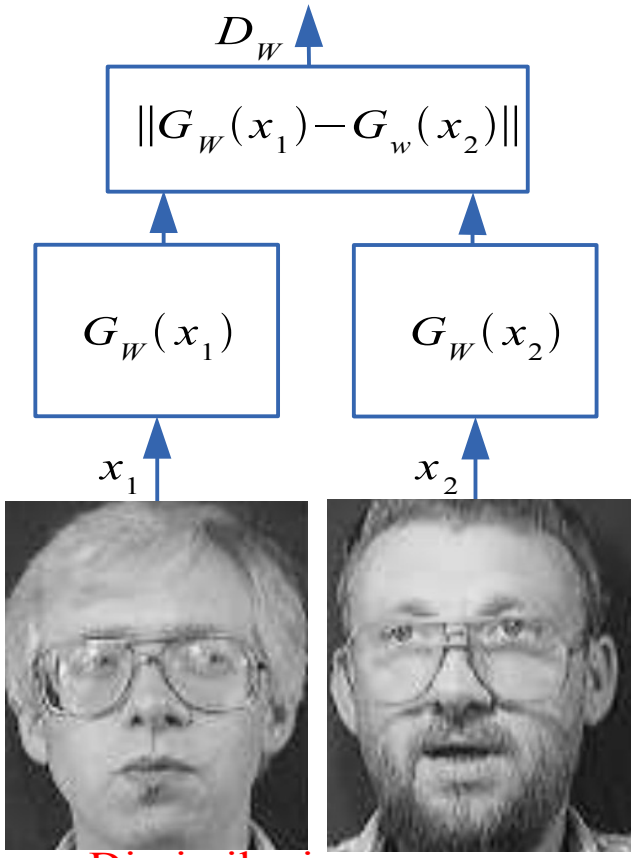
- Similar objects should produce outputs that are nearby
- Dissimilar objects should produce output that are far apart.
- DrLIM: Dimensionality Reduction by Learning and Invariant Mapping
- [Chopra et al. CVPR 2005]
- [Hadsell et al. CVPR 2006]

Make this small



Similar images (neighbors in the neighborhood graph)

Make this large



Dissimilar images (non-neighbors in the neighborhood graph)

Pose Estimation and Attribute Recovery with ConvNets

Y LeCun

Pose-Aligned Network for Deep Attribute Modeling

[Zhang et al. CVPR 2014] (Facebook AI Research)

(a) Highest scoring results for people wearing glasses.

(b) Highest scoring results for people wearing a hat.

Real-time hand pose recovery

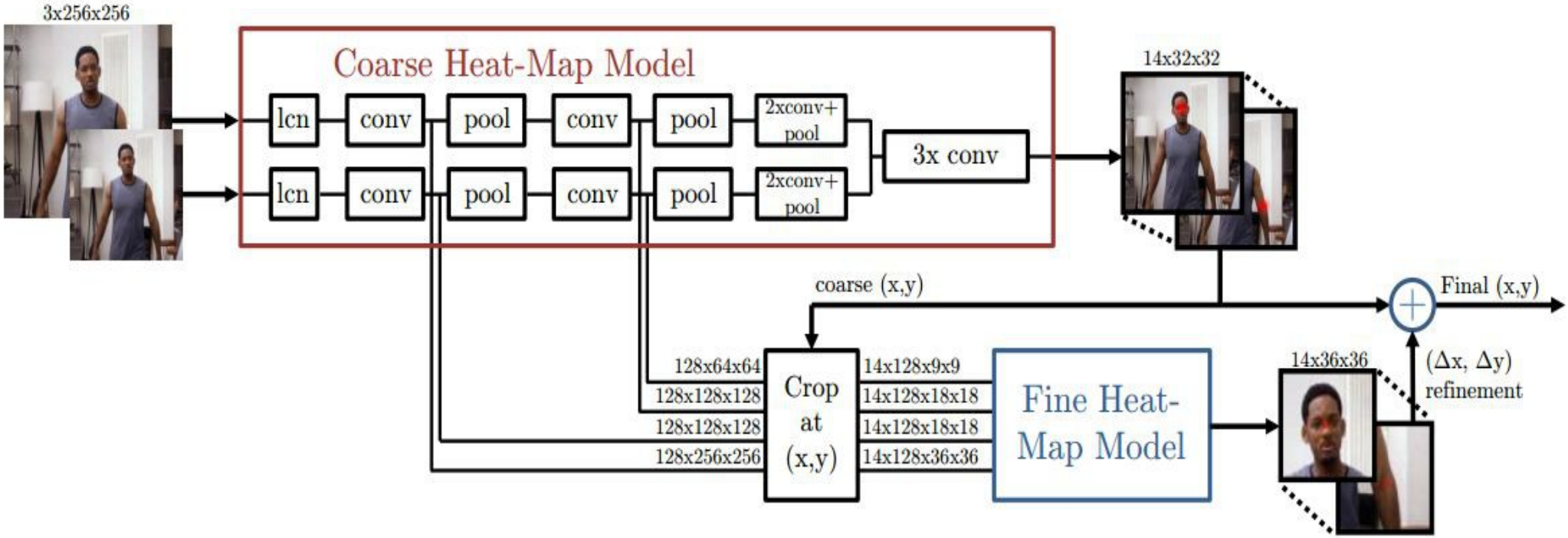
[Tompson et al. Trans. on Graphics 14]

Body pose estimation [Tompson et al. ICLR, 2014]

HAND POSE V IDEO

Person Detection and Pose Estimation

Tompson, Goroshin, Jain, LeCun, Bregler arXiv:1411.4280 (2014)



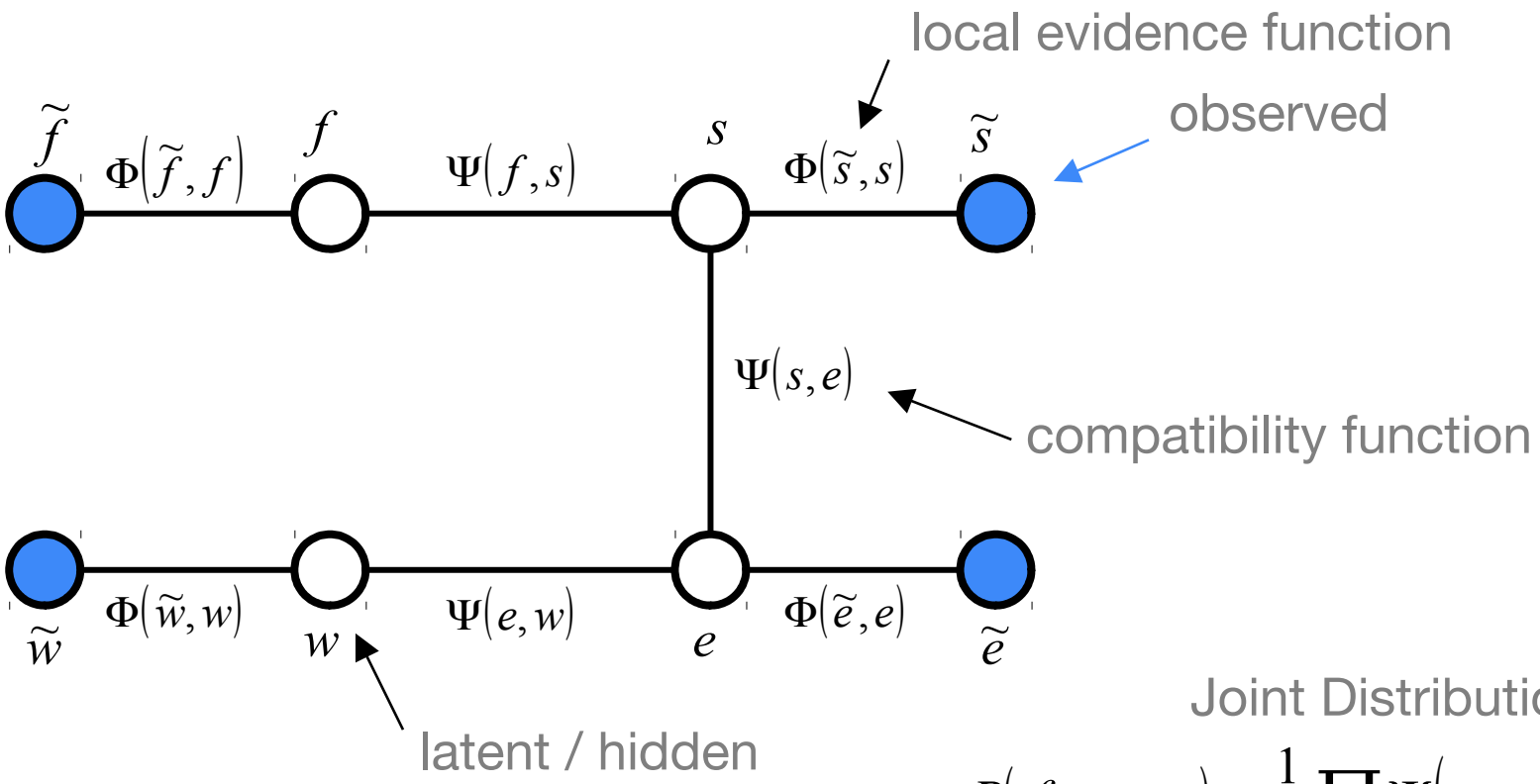
Person Detection and Pose Estimation

Tompson, Goroshin, Jain, LeCun, Bregler arXiv:1411.4280 (2014)

SPATIAL MODEL

Start with a tree graphical model

MRF over spatial locations



Joint Distribution:

$$P(f,s,e,w) = \frac{1}{Z}\prod_{i,j}\Psi(x_i,x_j)\prod_i \Phi(x_i,\tilde{x}_i)$$

SPATIAL MODEL

Start with a tree graphical model

… And approximate it

$$b(f) = \Phi(f) \prod_i \left(\Phi(x_i) * \Psi(f \mid x_i) + c(f \mid x_i) \right)$$

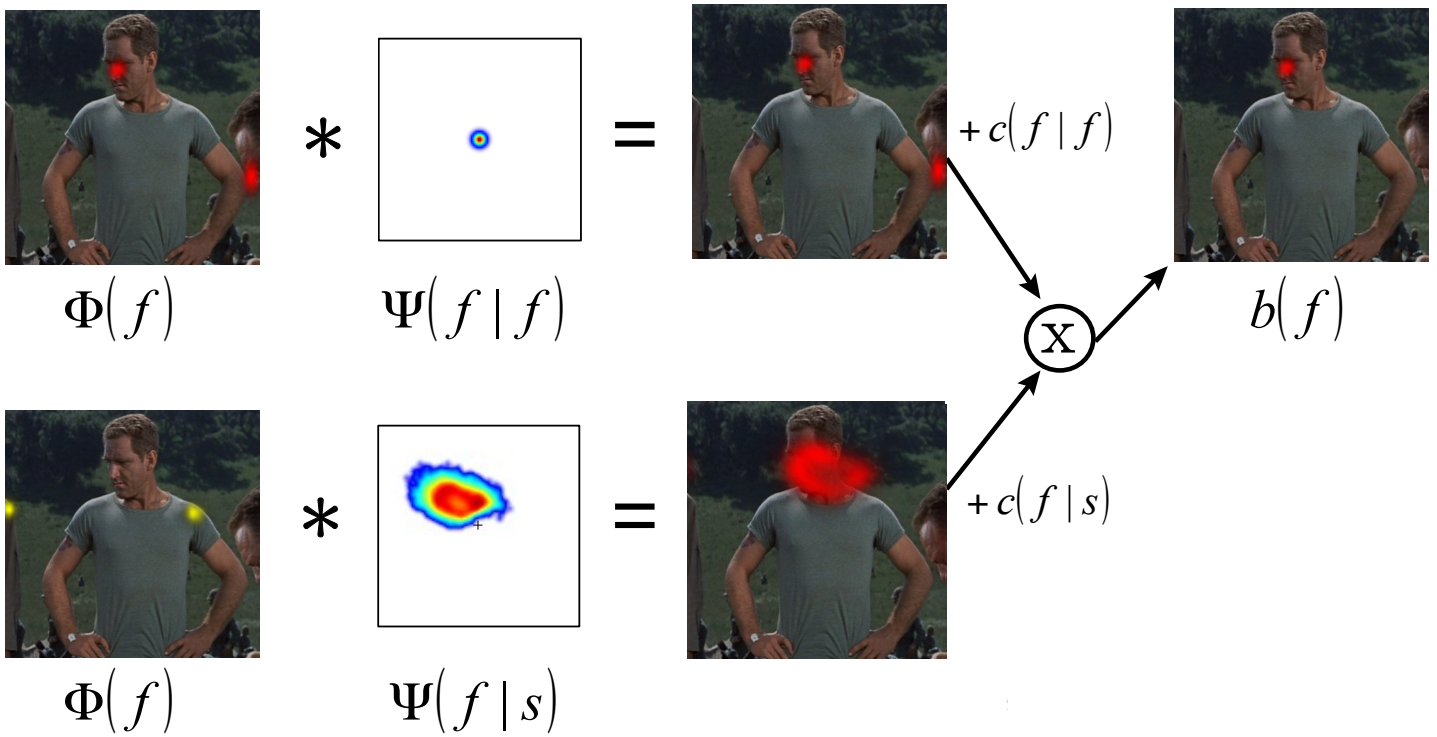


Image captioning: generating a descriptive sentence

Y LeCun

[Lebret, Pinheiro, Collobert 2015] [Kulkarni 11] [Mitchell 12] [Vinyals 14] [Mao 14]

A man riding skis on a snow covered ski slope.

NP: a man, skis, the snow, a person, a woman, a snow covered slope, a slope, a snowboard, a skier, man.
VP: wearing, riding, holding, standing on, skiing down.
PP: on, in, of, with, down.

A man wearing skis on the snow.

A man is doing skateboard tricks on a ramp.

NP: a skateboard, a man, a trick, his skateboard, the air, a skateboarder, a ramp, a skate board, a person, a woman.
VP: doing, riding, is doing, performing, flying through.
PP: on, of, in, at, with.

A man riding a skateboard on a ramp.

The girl with blue hair stands under the umbrella.

NP: a woman, an umbrella, a man, a person, a girl, umbrellas, that, a little girl, a cell phone.
VP: holding, wearing, is holding, holds, carrying.
PP: with, on, of, in, under.

A woman is holding an umbrella.

A slice of pizza sitting on top of a white plate.

NP: a plate, a white plate, a table, pizza, it, a pizza, food, a sandwich, top, a close.
VP: topped with, has, is, sitting on, is on.
PP: of, on, with, in, up.

A table with a plate of pizza on a white plate.

A baseball player swinging a bat on a field.

NP: the ball, a game, a baseball player, a man, a tennis court, a ball, home plate, a baseball game, a batter, a field.
VP: swinging, to hit, playing, holding, is swinging.
PP: on, during, in, at, of.

A baseball player swinging a bat on a baseball field.

A bunch of kites flying in the sky on the beach.

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group.
VP: flying, flies, is flying, flying in, are.
PP: on, of, with, in, at.

People flying kites on the beach.


Video Classification

Learning Video Features with C3D

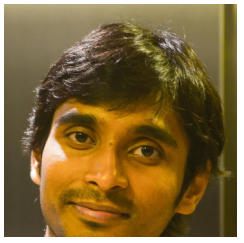
- C3D Architecture
  - 8 convolution, 5 pool, 2 fully-connected layers
  - 3x3x3 convolution kernels
  - 2x2x2 pooling kernels
- Dataset: Sports-1M [Karpathy et al. CVPR'14]
  - 1.1M videos of 487 different sport categories
  - Train/test splits are provided


Du Tran (1,2)


Lubomir Bourdev (2)

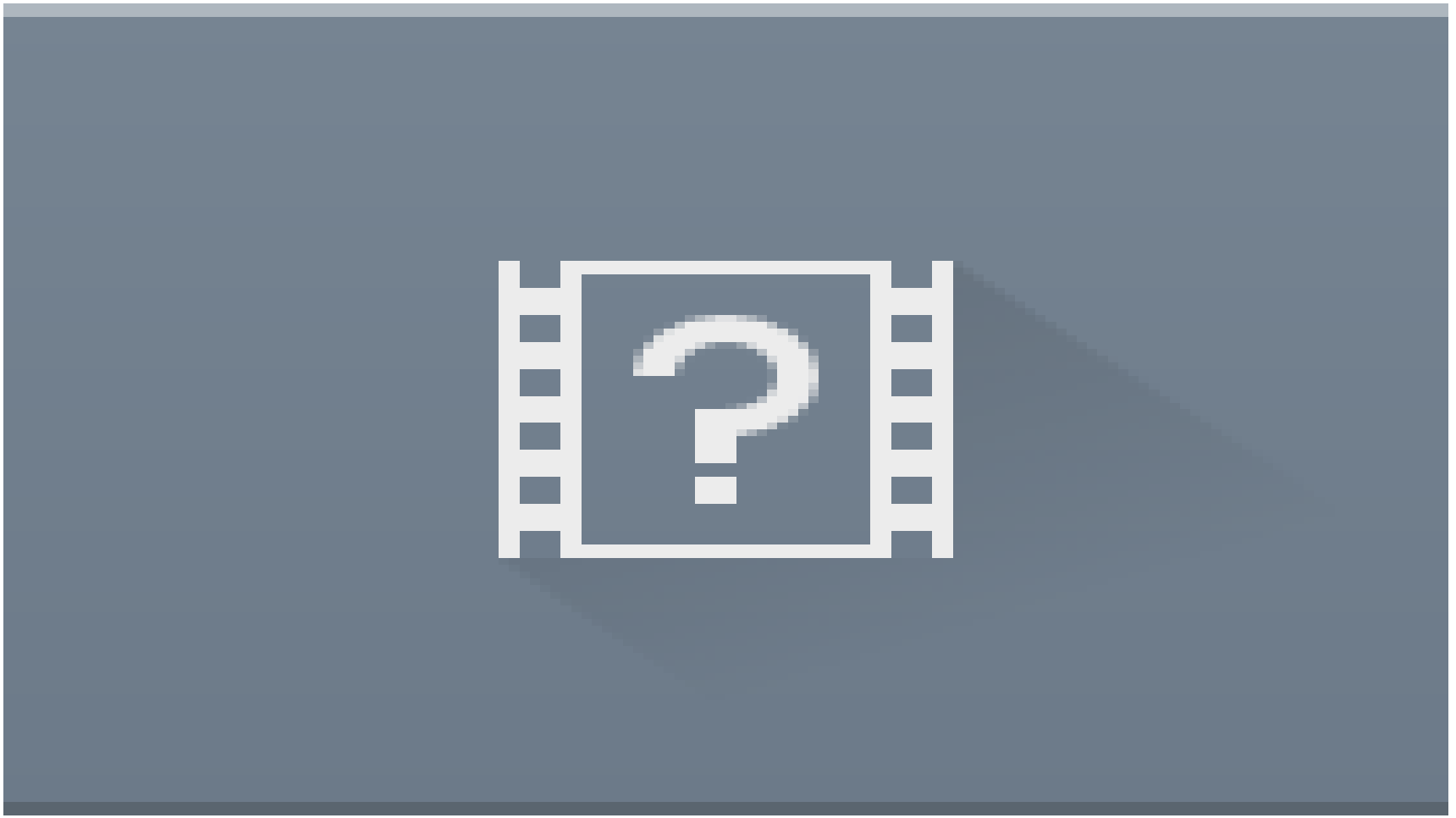

Rob Fergus (2,3)


Lorenzo Torresani (1)

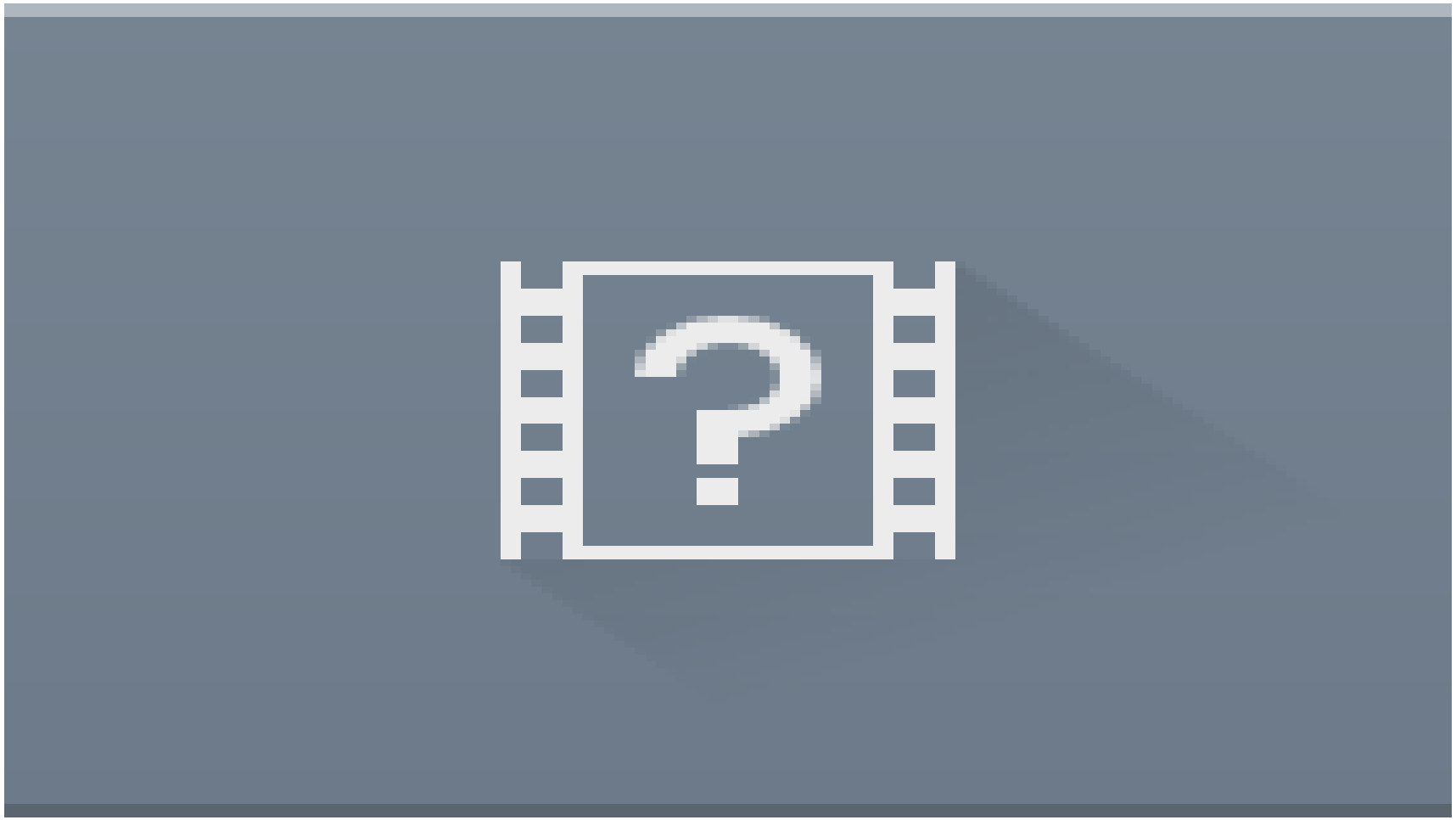

Manohar Paluri (2)

(1) Dartmouth College, (2) Facebook AI Research, (3) New York University

Learning Video Features with C3D



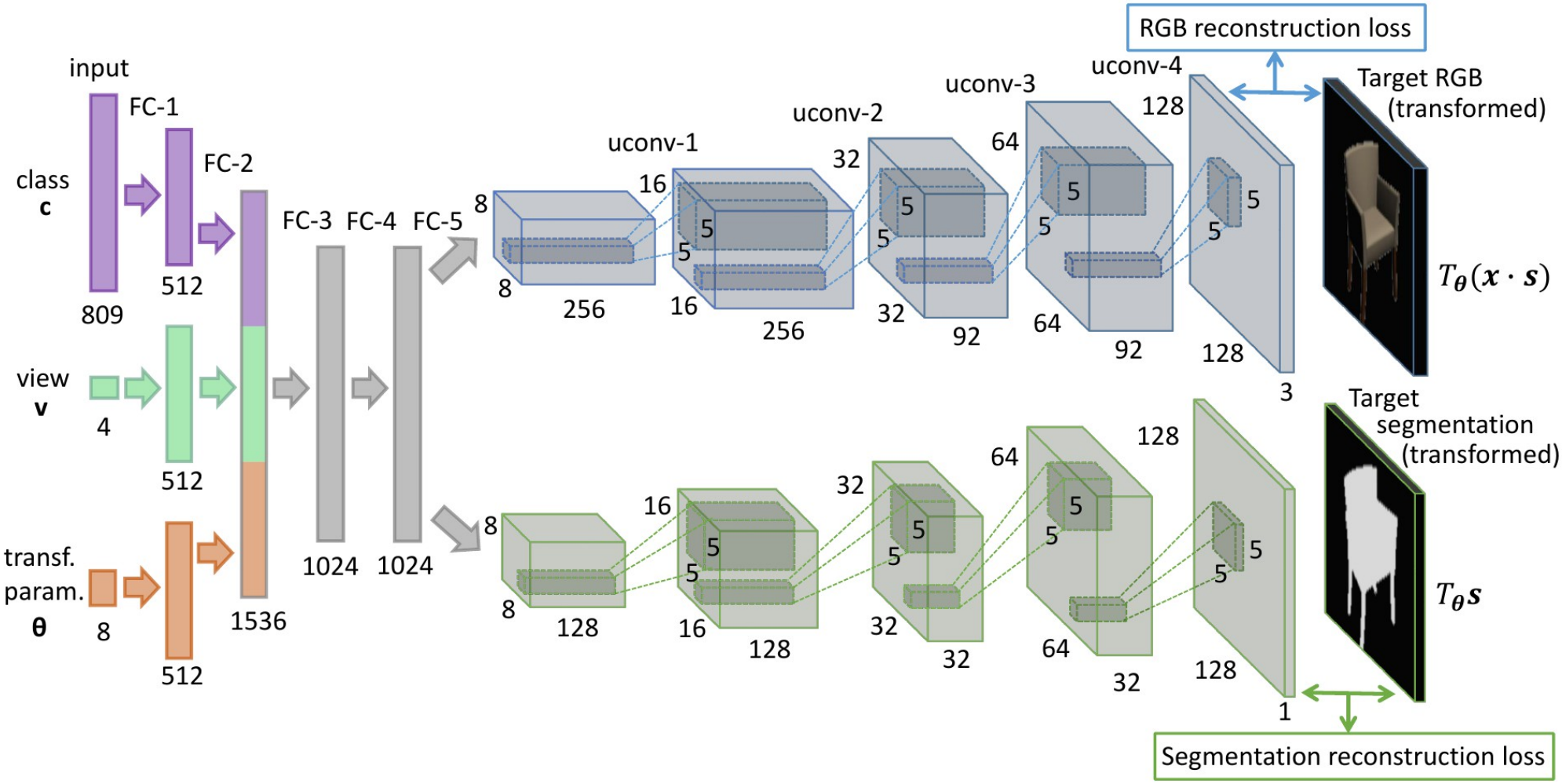
Learning Video Features with C3D



Supervised ConvNets that Draw Pictures

Y LeCun

- Using ConvNets to Produce Images
- [Dosovitskyi et al. Arxiv:1411:5928

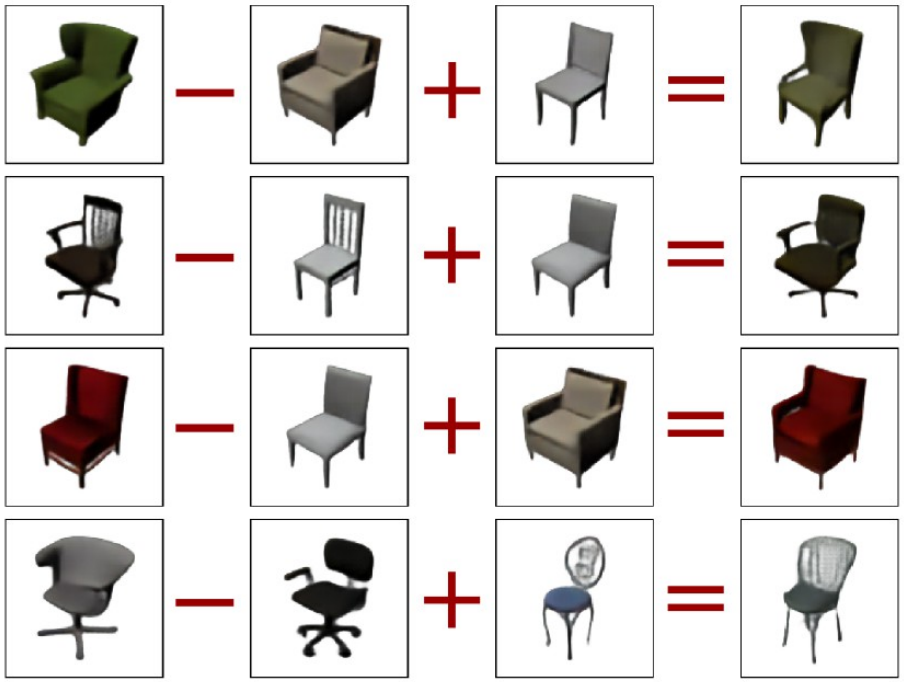


Supervised ConvNets that Draw Pictures

Y LeCun

Generating Chairs

Chair Arithmetic in Feature Space



Convolutional Encoder-Decoder

- Generating Faces
- [Kulkarni et al. Arxiv:1503:03167]