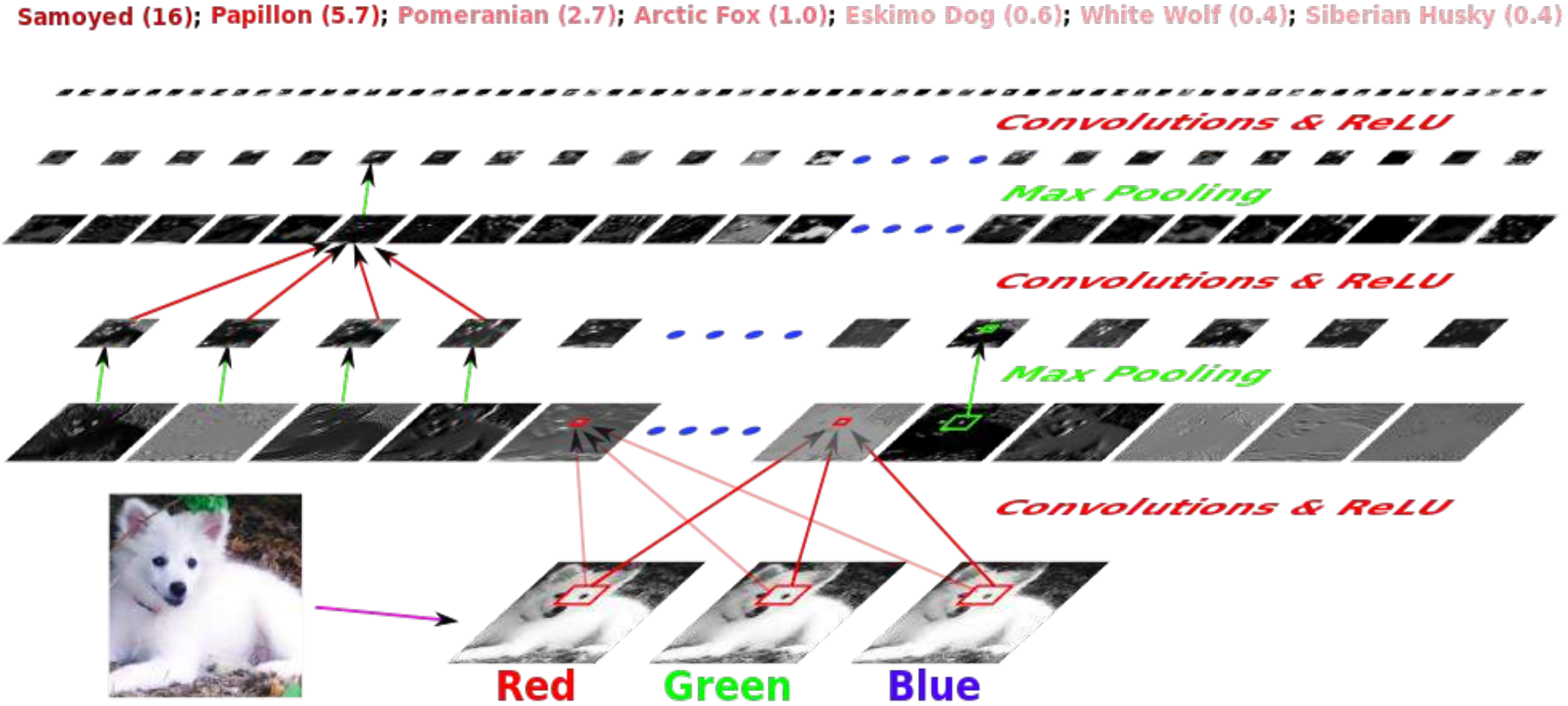
Convolutional Nets
Part 3

Yann LeCun
Facebook AI Research,
Center for Data Science, NYU
Courant Institute of Mathematical Sciences, NYU
http://yann.lecun.com

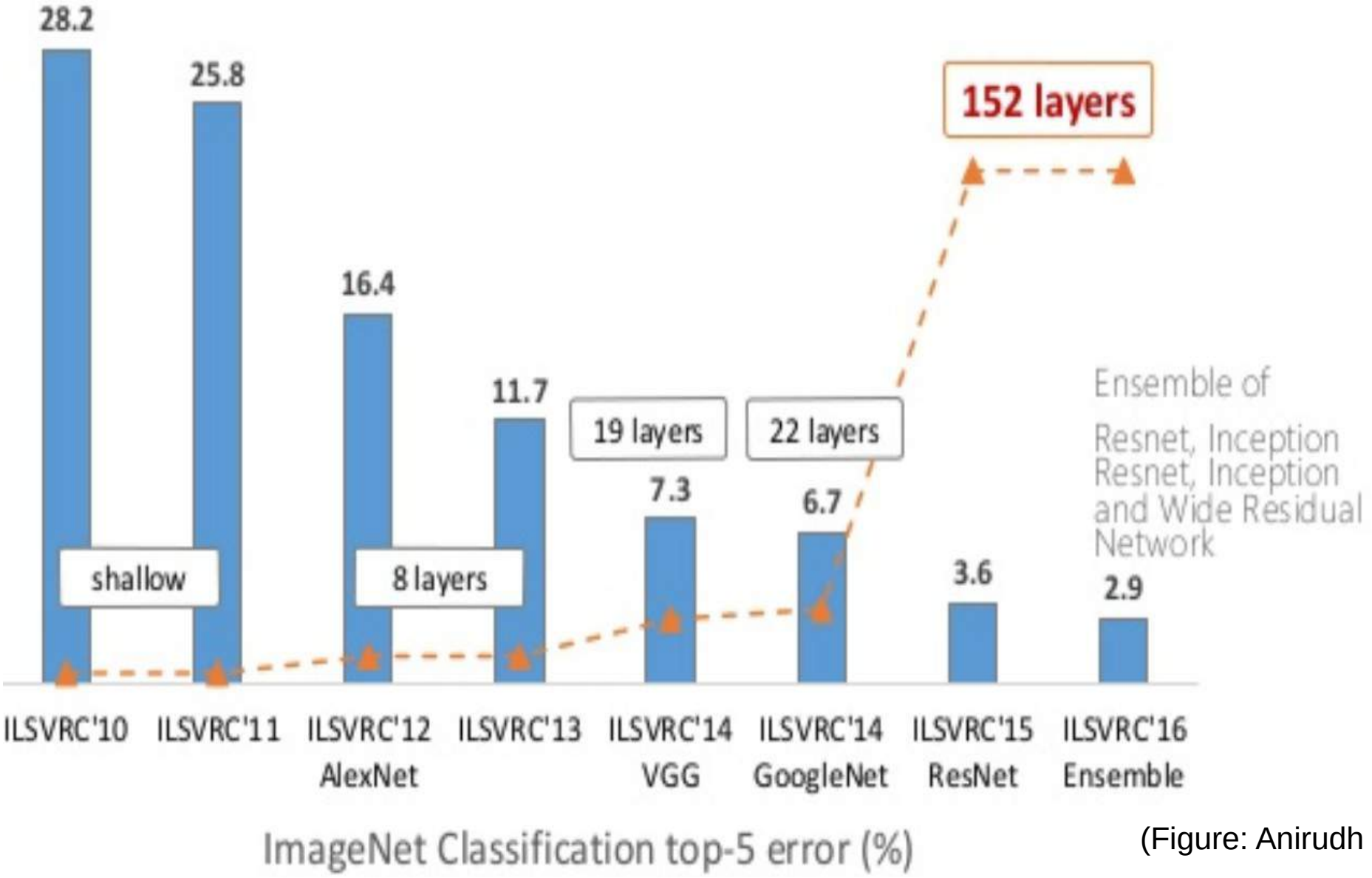
Deep ConvNets for Object Recognition (on GPU)

- AlexNet [Krizhevsky et al. NIPS 2012], OverFeat [Sermanet et al. 2013]
- 1 to 10 billion connections, 10 million to 1 billion parameters, 8 to 20 layers.



Error Rate on ImageNet

- **Depth inflation**

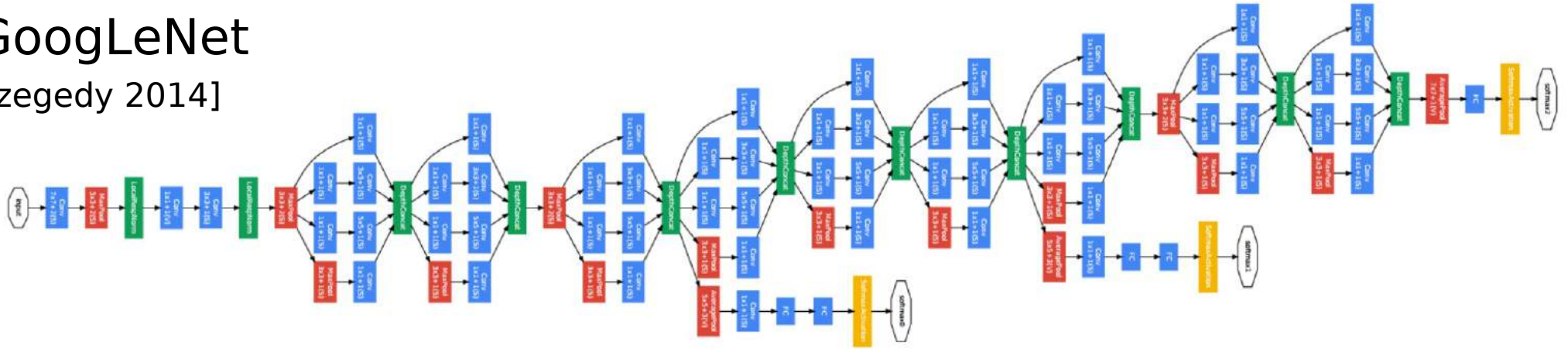


(Figure: Anirudh Koul)

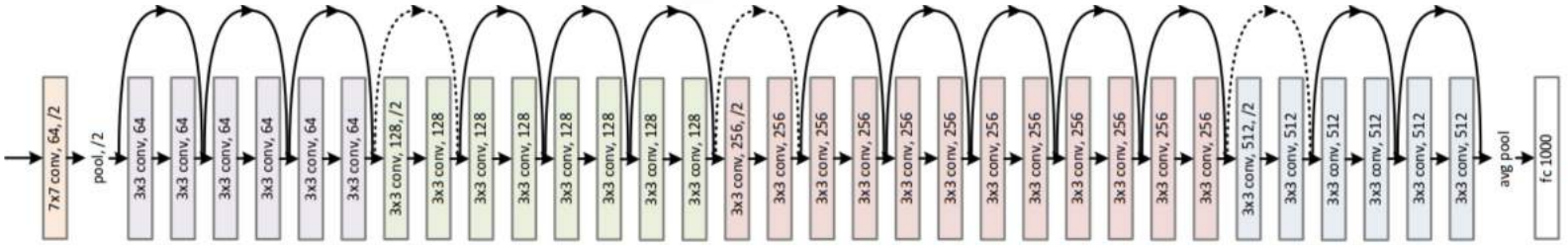
Deep ConvNets (depth inflation)

VGG
[Simonyan 2013]

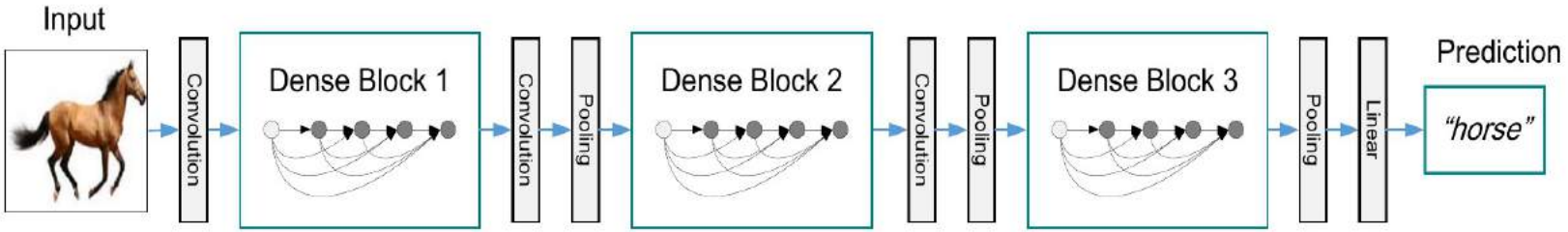
GoogLeNet
Szegedy 2014]



ResNet
[He et al. 2015]

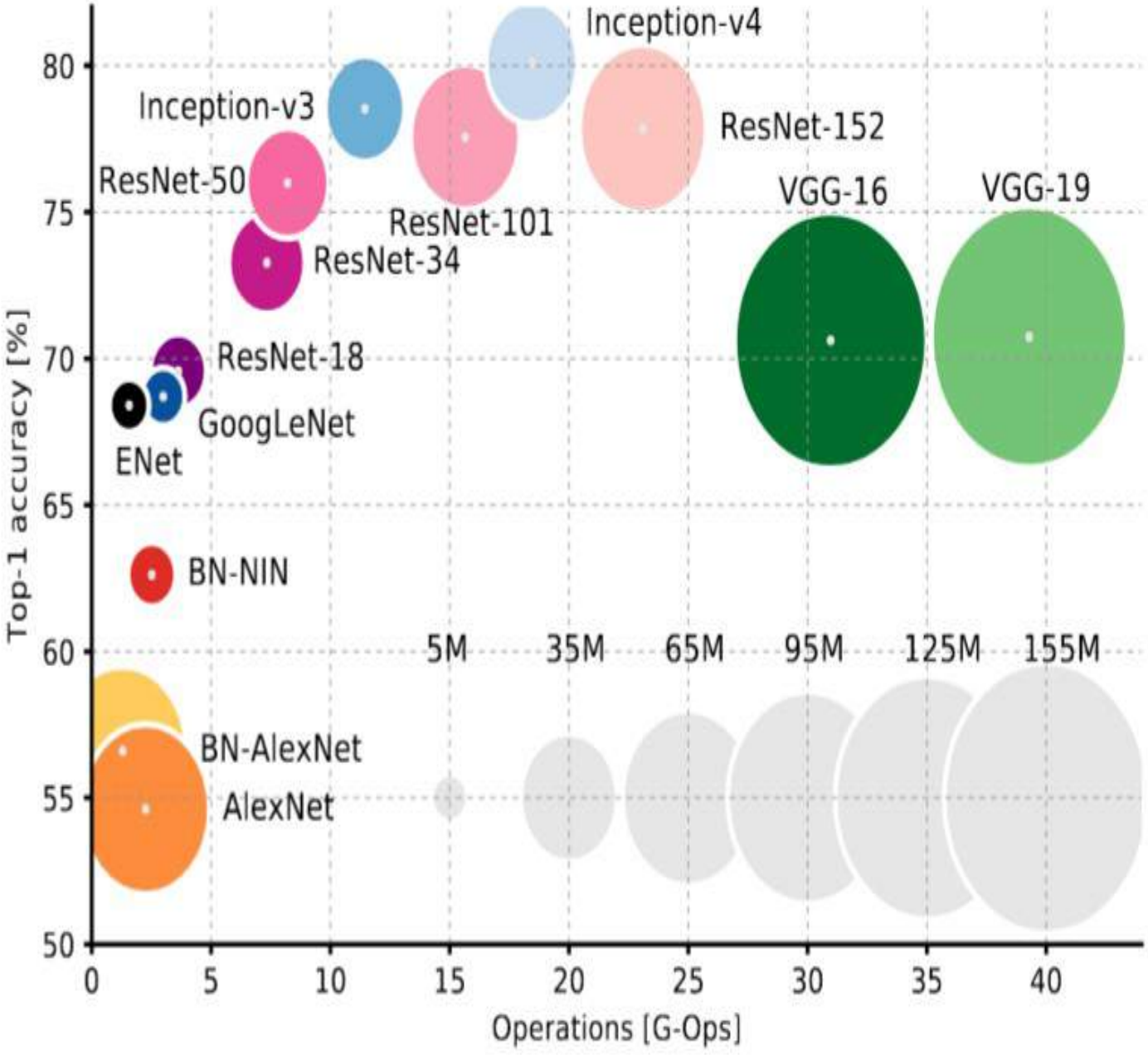


DenseNet
[Huang et al 2017]



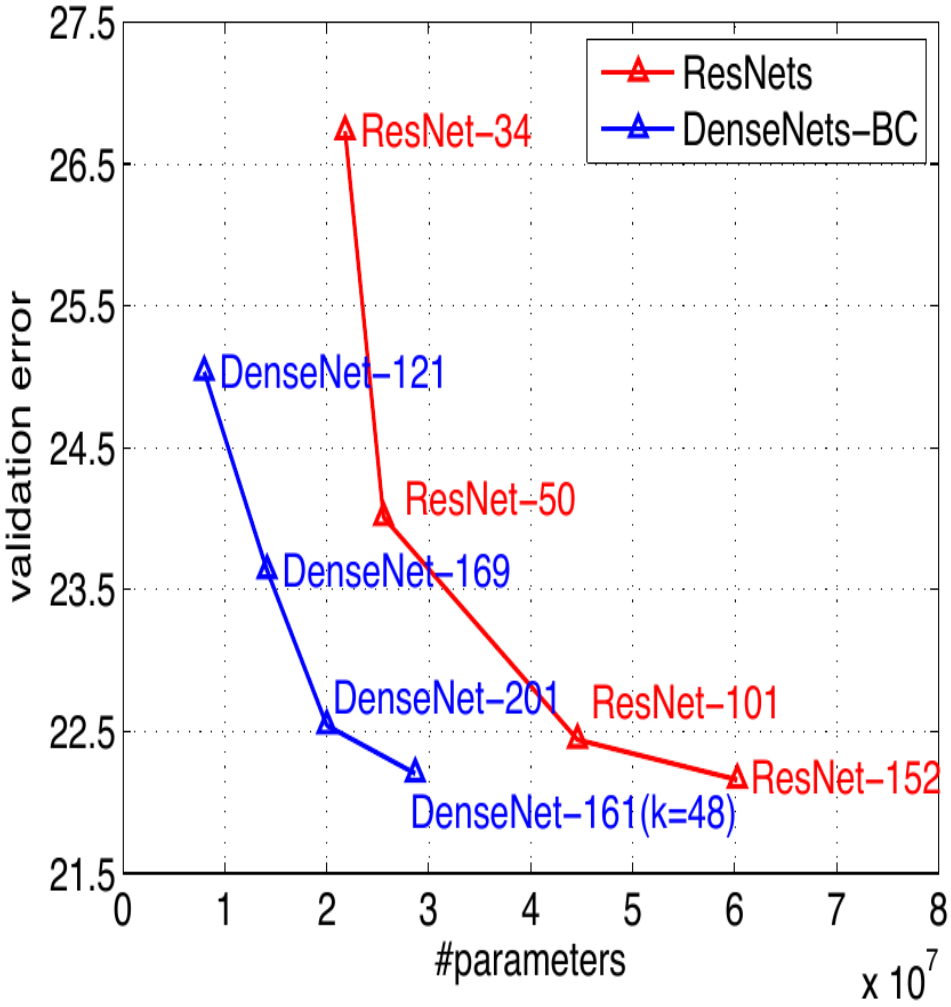
GOPS vs Accuracy on ImageNet vs #Parameters

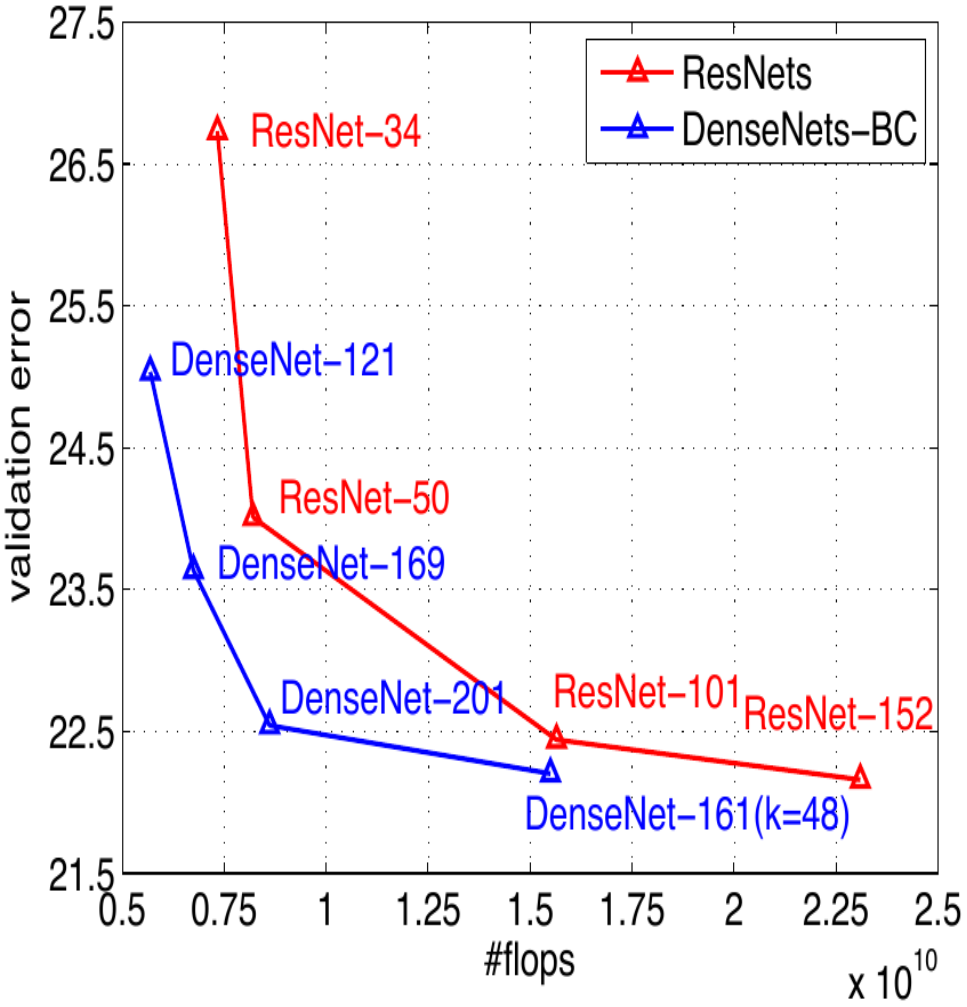
- [Canziani 2016]
- ResNet50 and ResNet 100 are used routinely in production.



DenseNet Results on ImageNet 1K

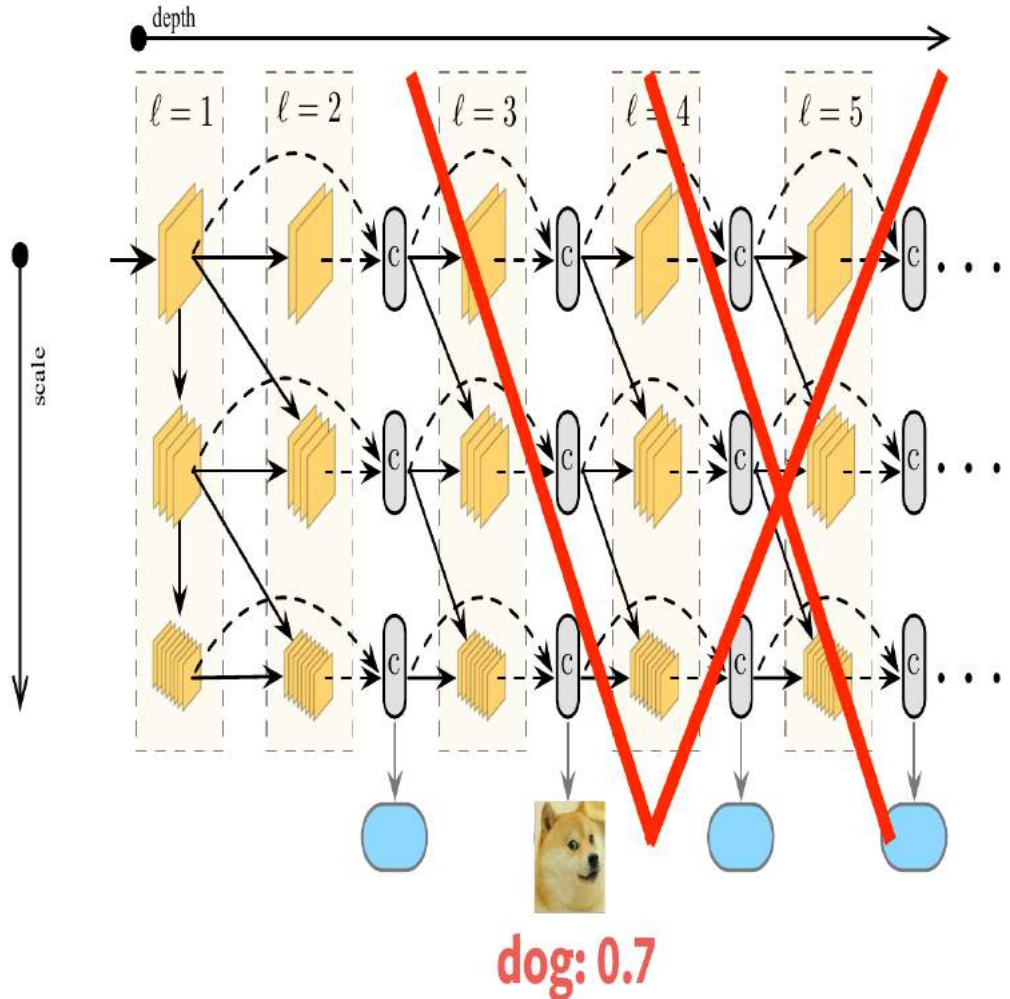
- **Good performance per FLOPS.**

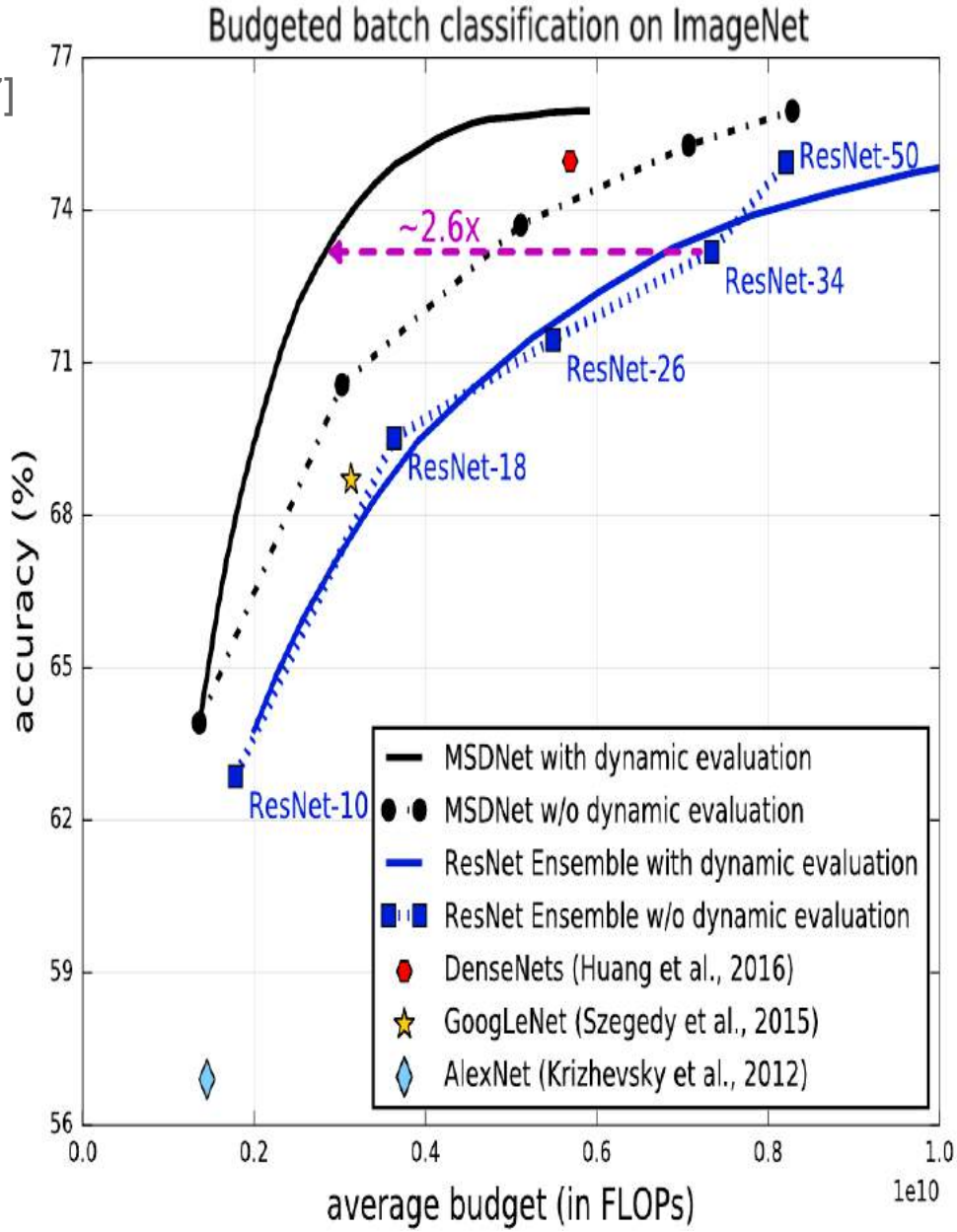




Multiscale DenseNet: fast conditional computation

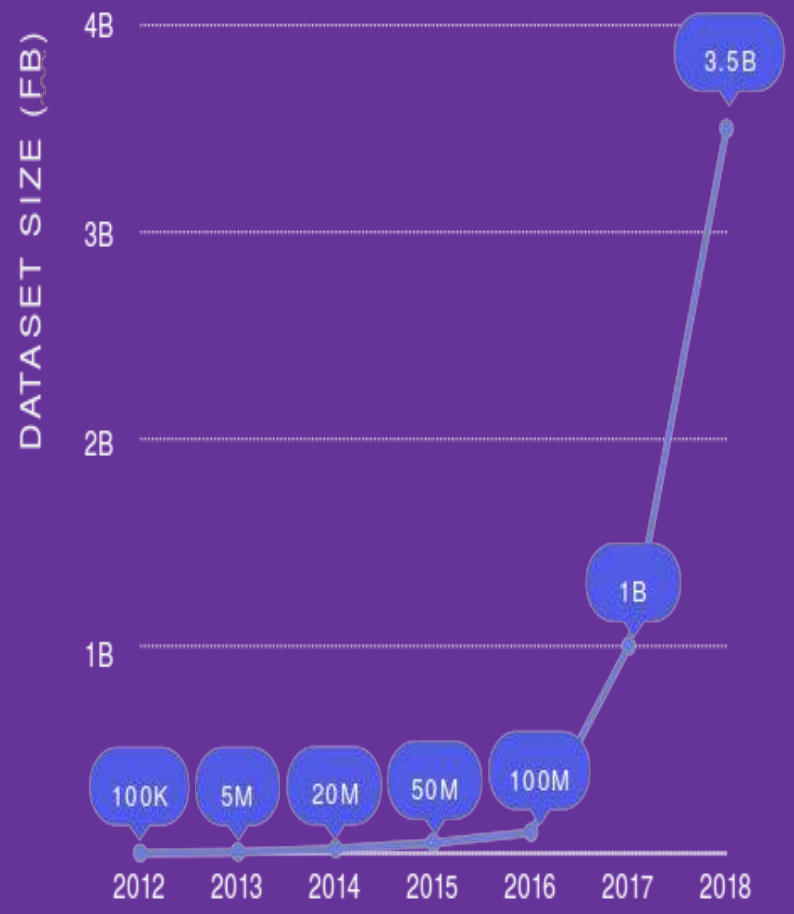
- **2.6x acceleration**
  - [Huang, Li, Weinberger, van der Maaten CVPR 2017]
  - [Huang et al. ArXiv:1703:09844]

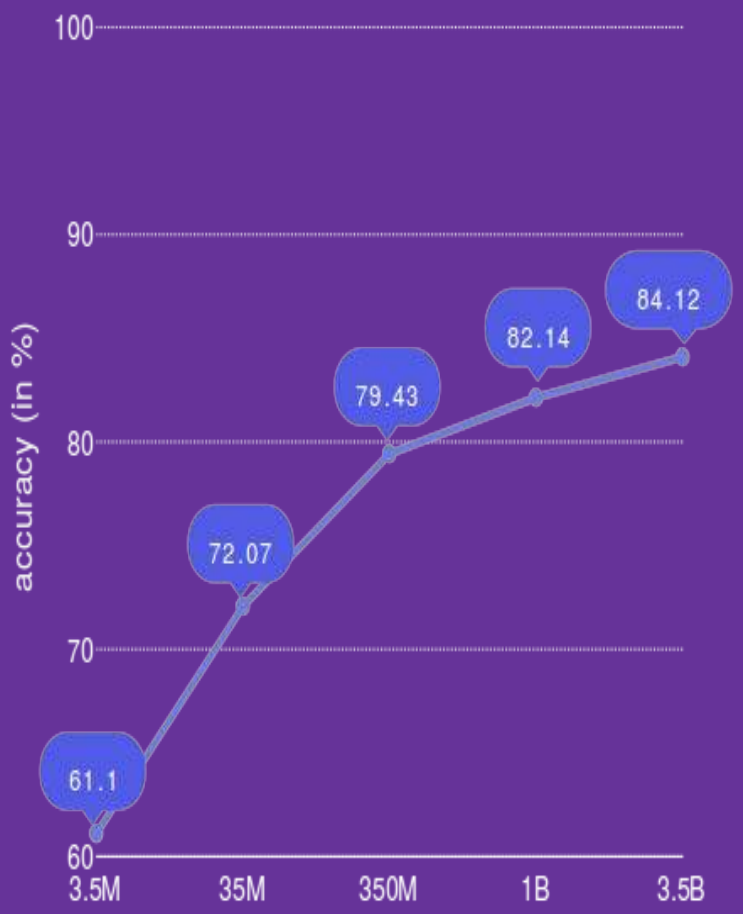




Future: weakly/self-supervised learning on massive datasets

- **Pretraining on 3.5b instagram images with 17k hashtags. Training/test on ImageNet**

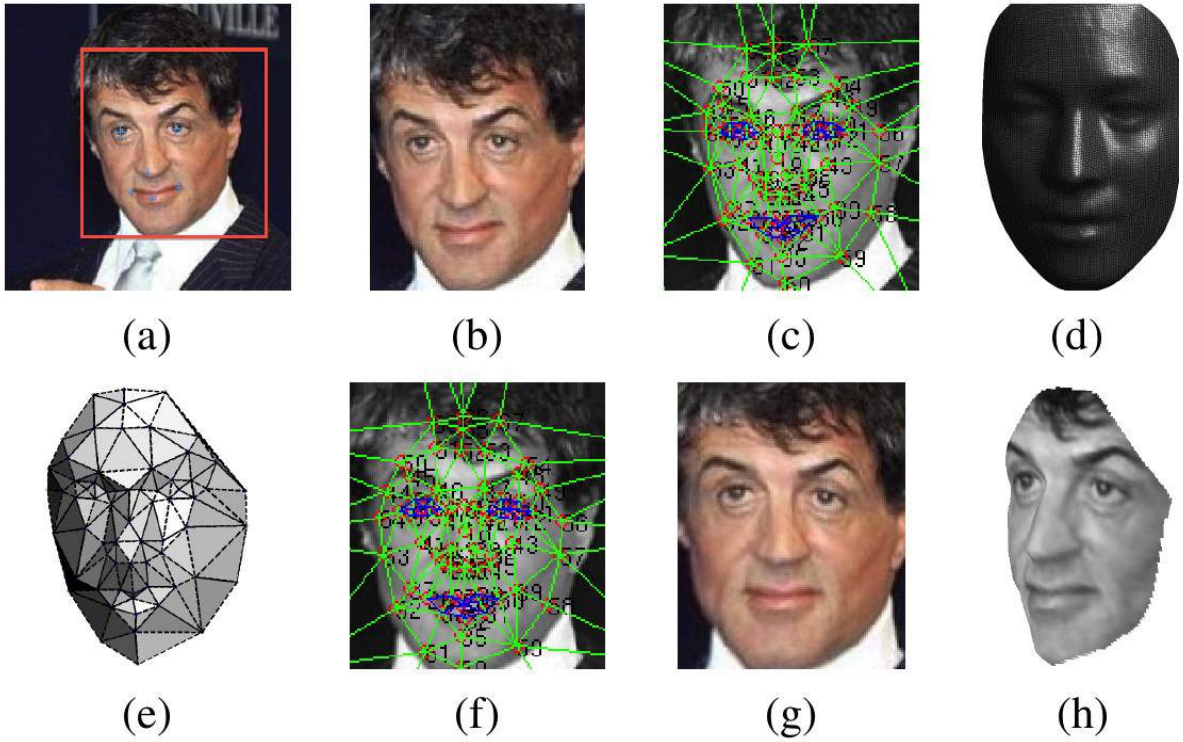


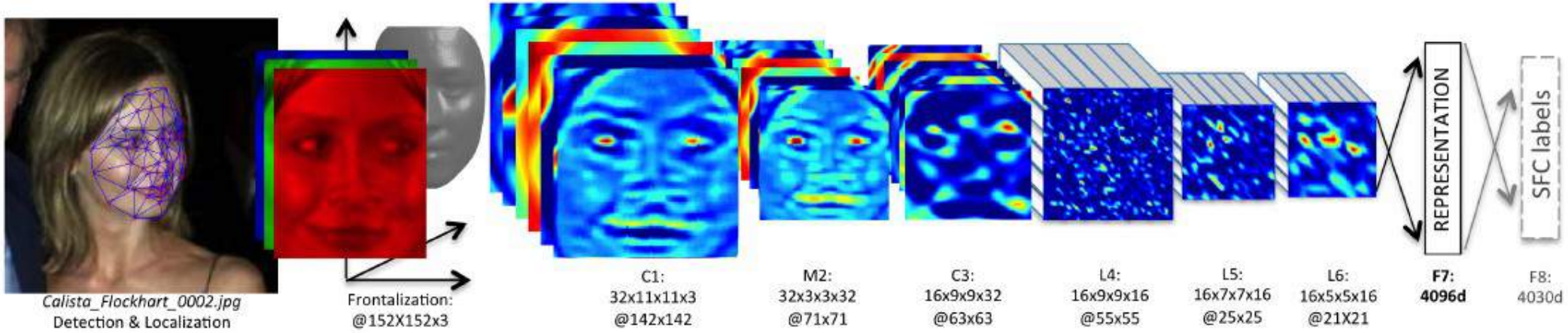


[Mahajan et al. CVPR 2018]

Deep Face

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

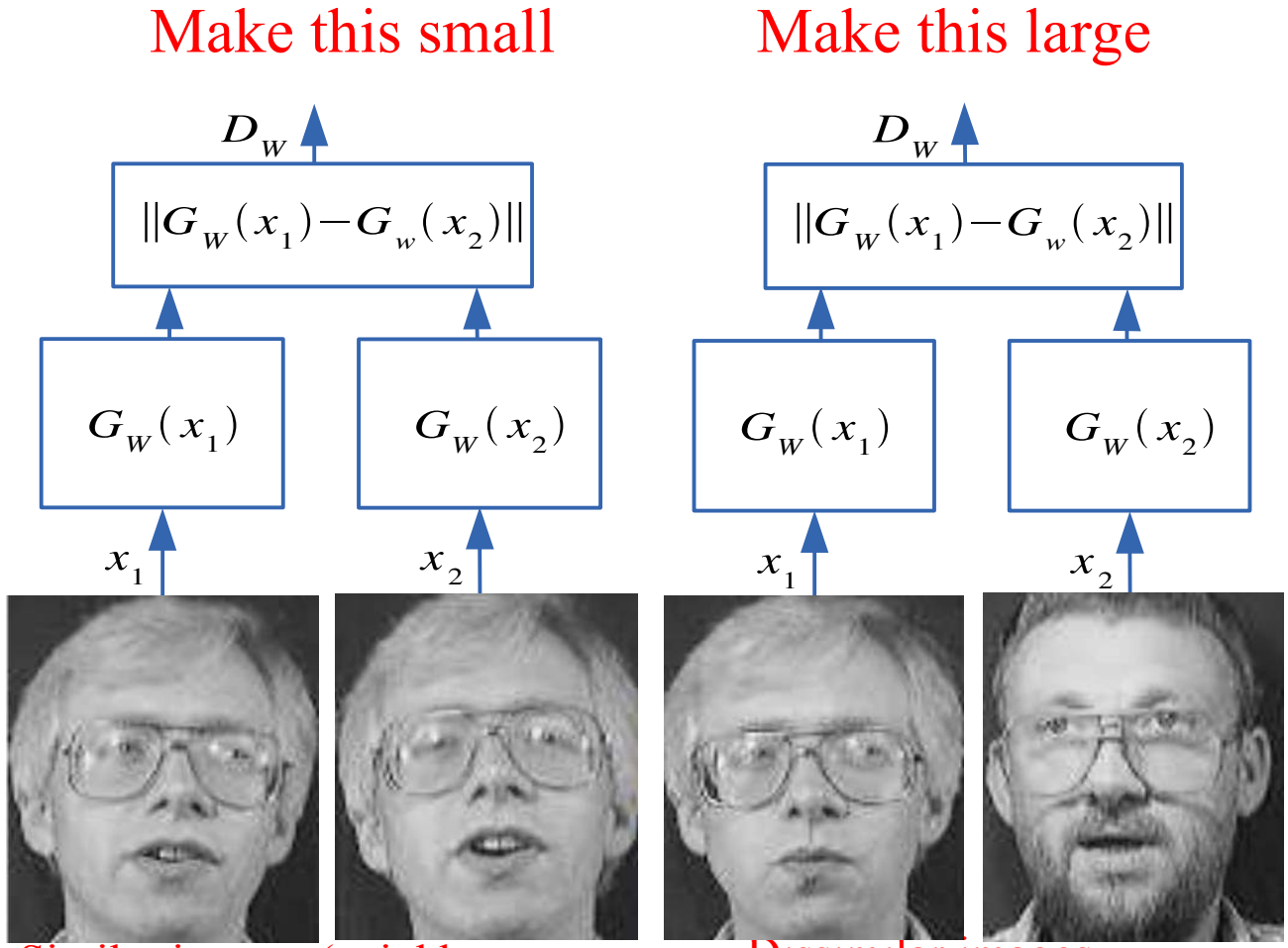




Metric Learning with a Siamese Architecture

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]



Object Detection And Localization With ConvNets

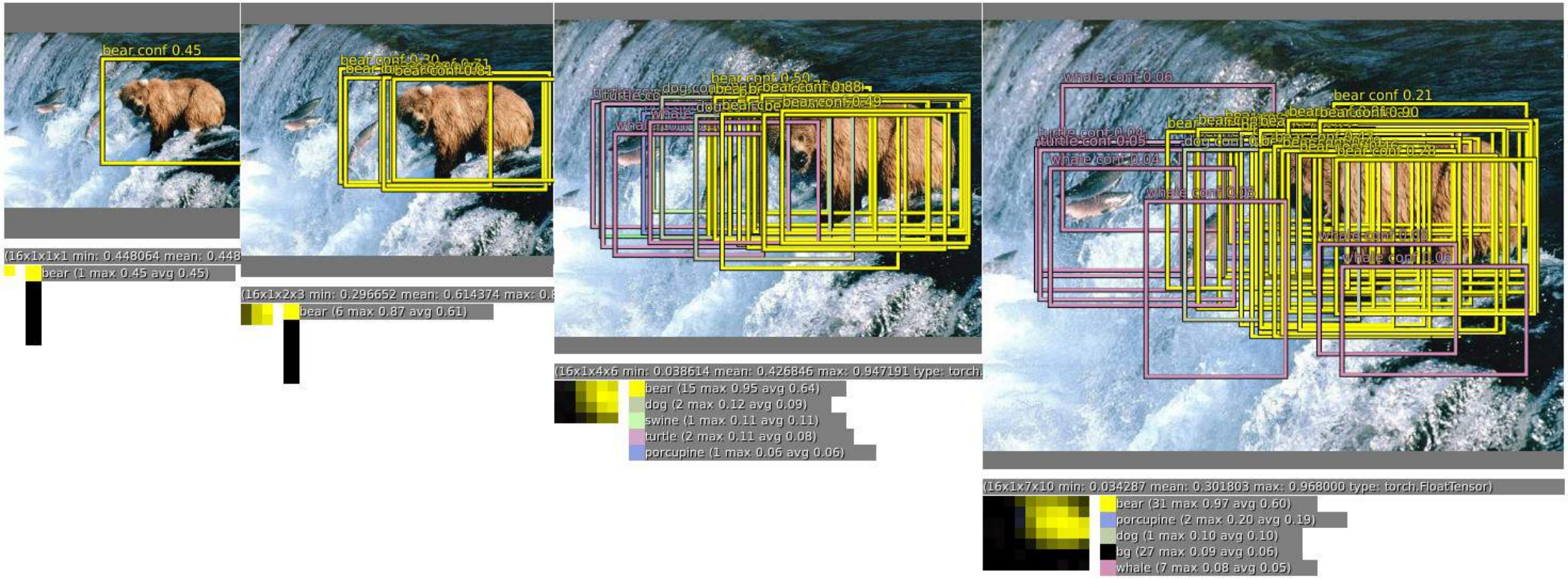
Classification + Localization: multiscale sliding window

- **Apply convnet with a sliding window over the image at multiple scales**
- **Important note: it's very cheap to slide a convnet over an image**
  - Just compute the convolutions over the whole image and replicate the fully-connected layers

Classification + Localization: sliding window + bounding box regression

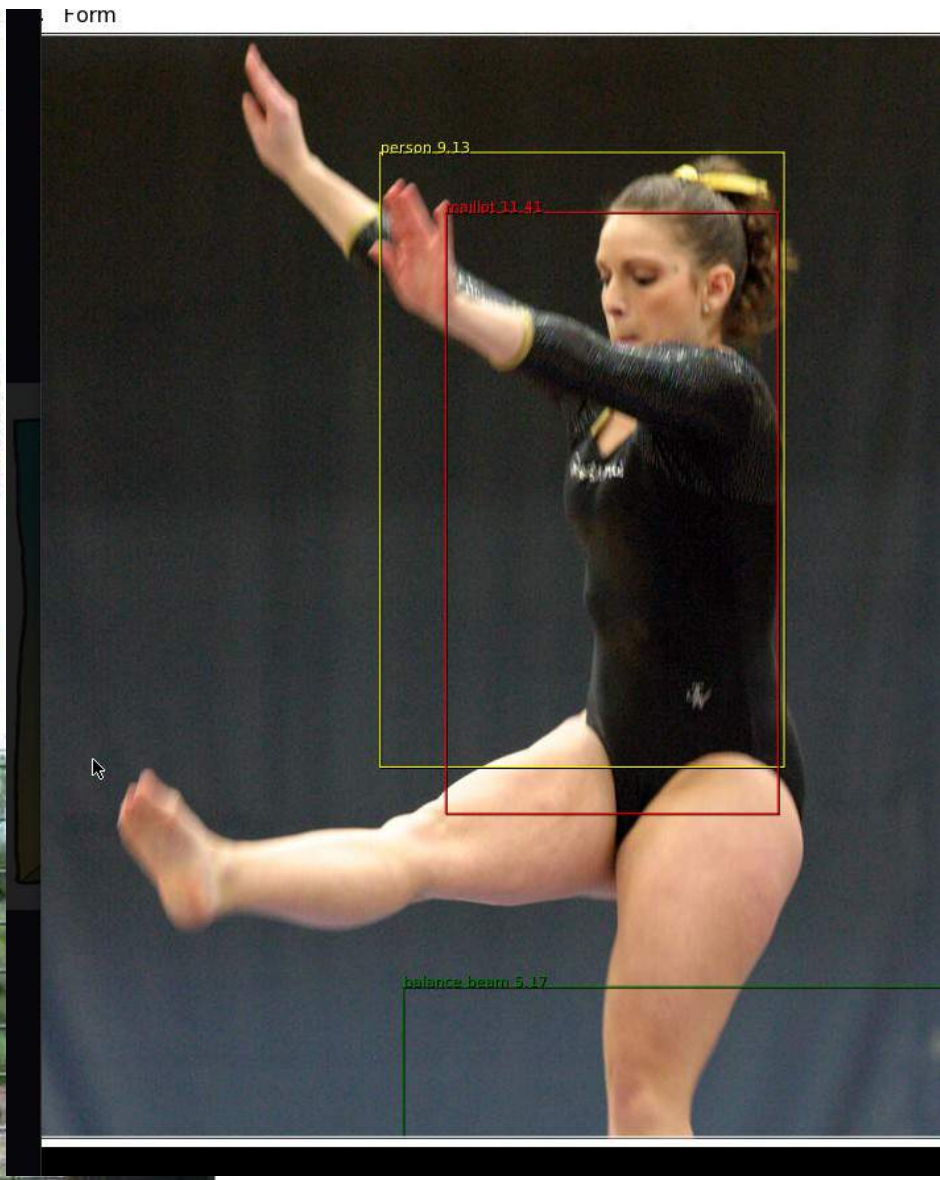
Y LeCun

- **Apply convnet with a sliding window over the image at multiple scales**
- **For each window, predict a class and bounding box parameters**
  - Evenif the object is not completely contained in the viewing window, the convnet can predict where it thinks the object is.

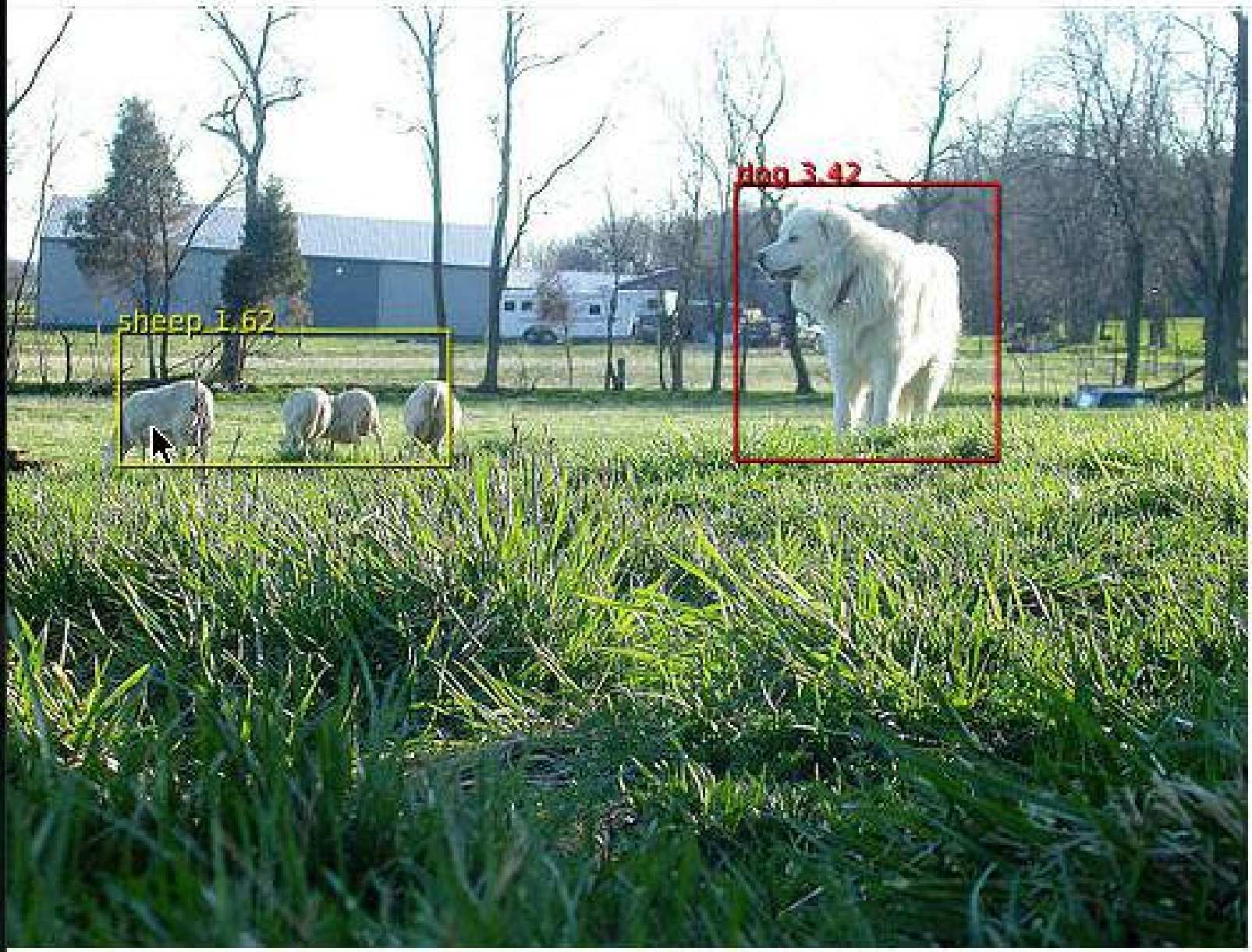


Results: pre-trained on ImageNet1K, fine-tuned on ImageNet Detection

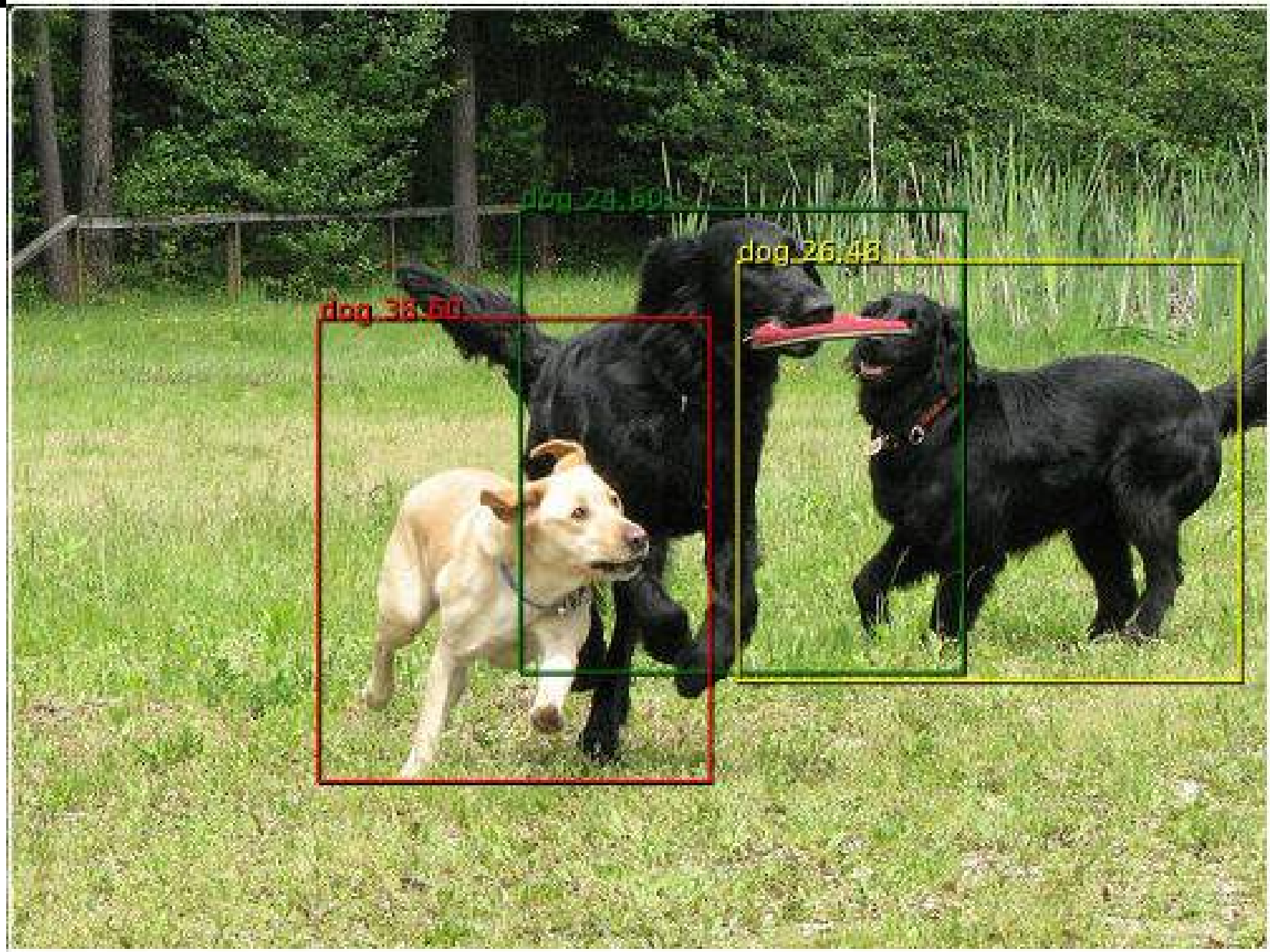
Y LeCun



Detection Examples



Detection Examples



Detection Examples

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

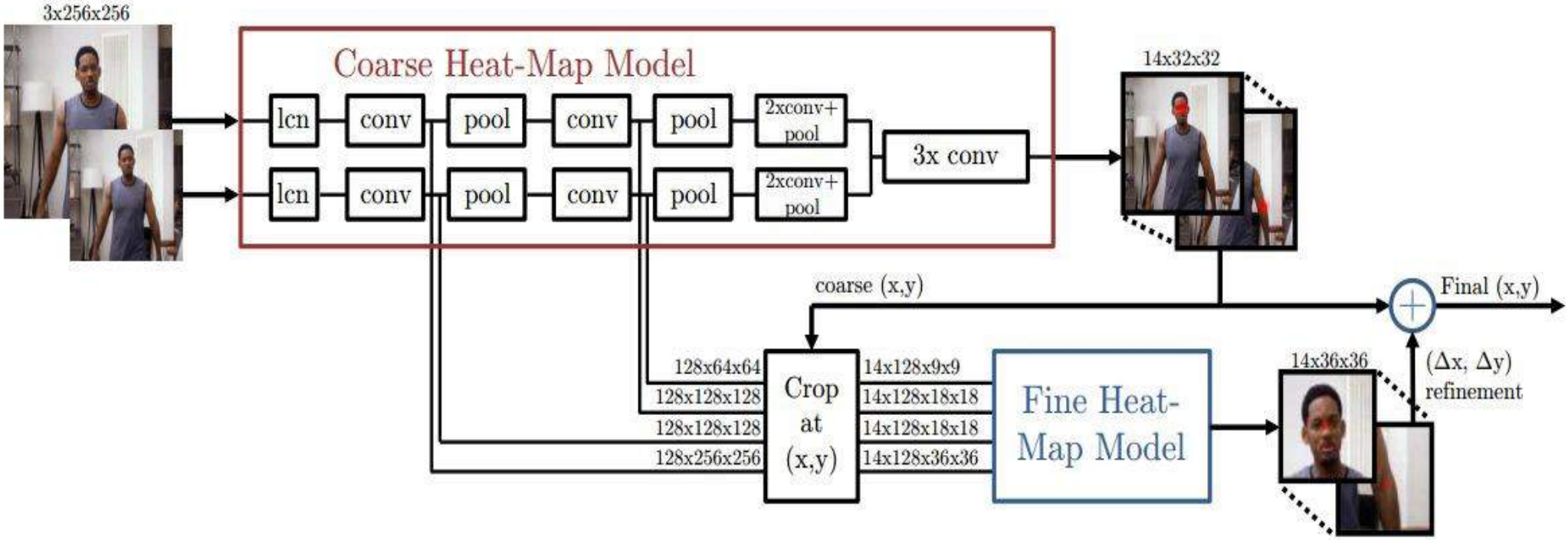
[Tompson et al. Trans. on Graphics 14]

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

HAND POSE VIDEO

Person Detection and Pose Estimation

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



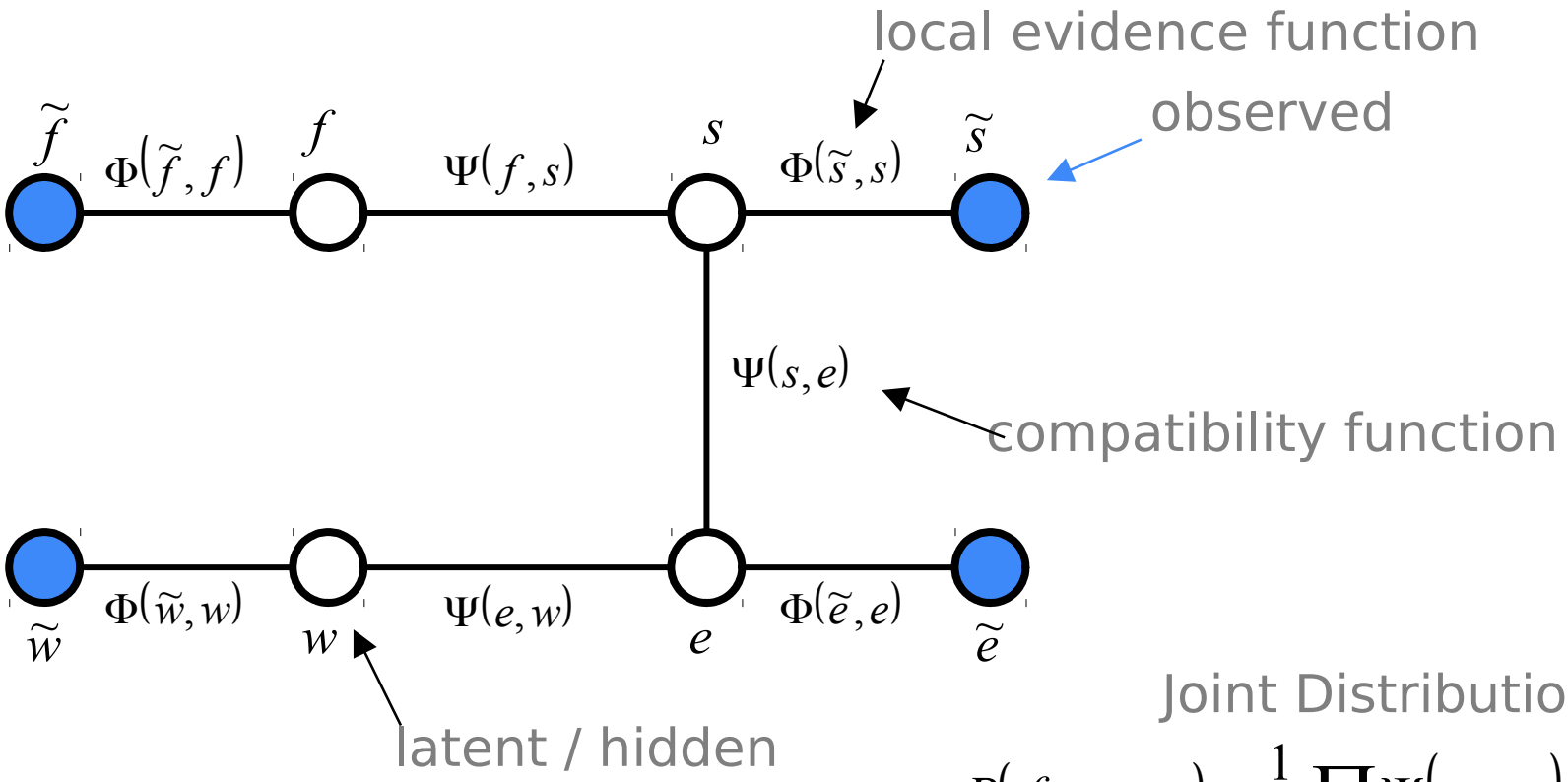
Person Detection and Pose Estimation

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

SPATIAL MODEL

Start with a tree graphical model

MRF over spatial locations



Joint Distribution:

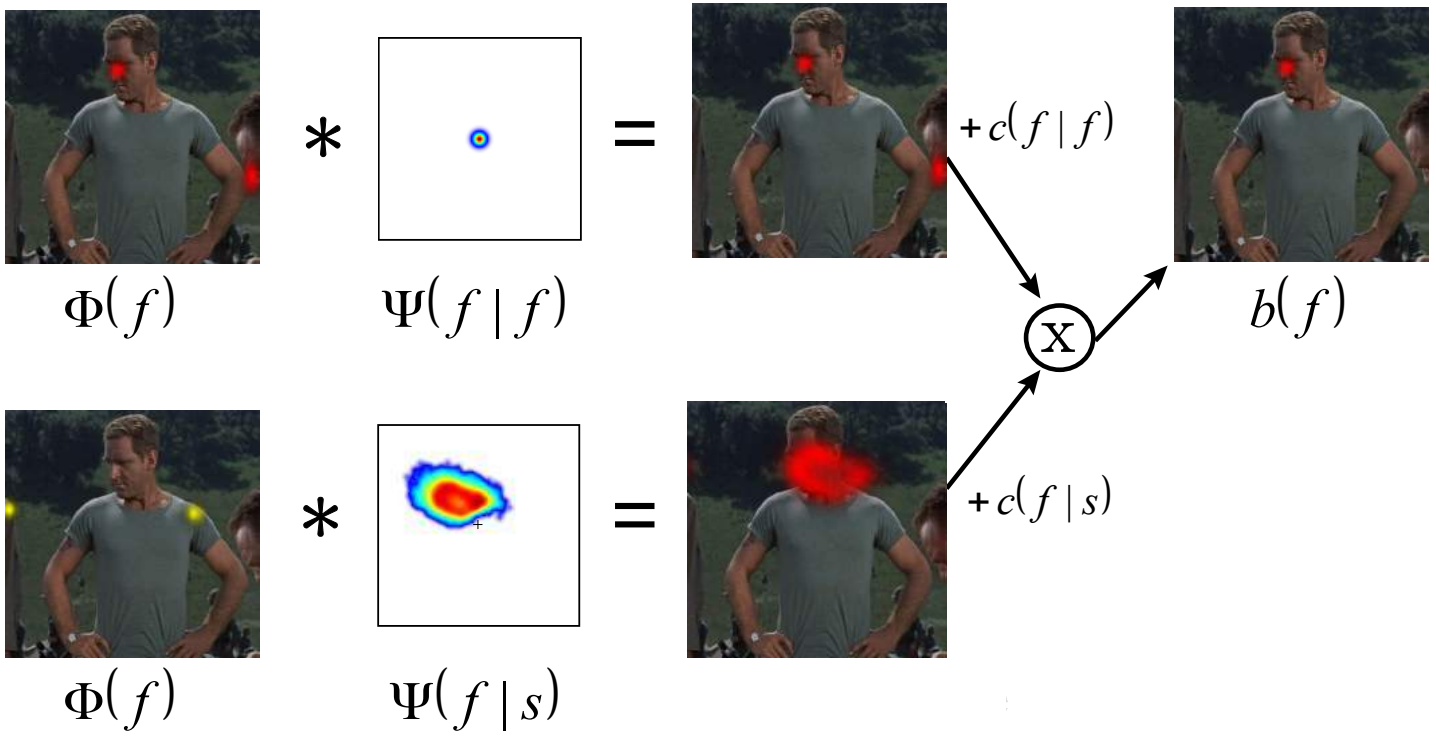
$$P(f,s,e,w) = \frac{1}{Z}\prod_{i,j}\Psi(x_i,x_j)\prod_{i}\Phi(x_i,\tilde{x}_i)$$

SPATIAL MODEL

Start with a tree graphical model

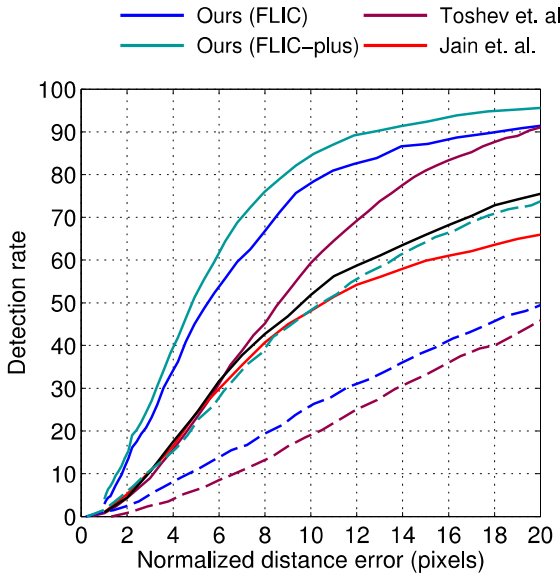
... And approximate it

$$b(f) = \Phi(f) \prod_i (\Phi(x_i) \star \Psi(f \mid x_i) + c(f \mid x_i))$$

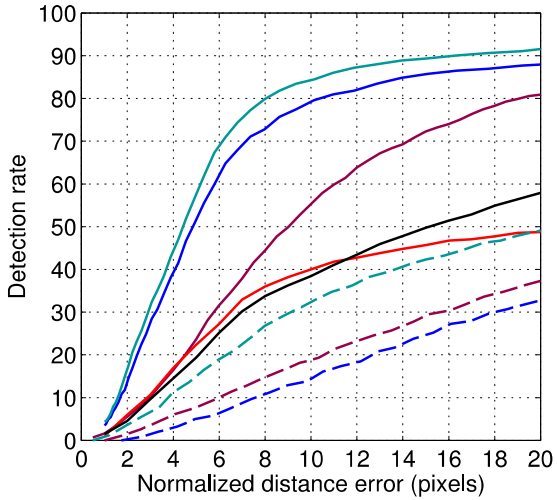


SPATIAL MODEL: RESULTS

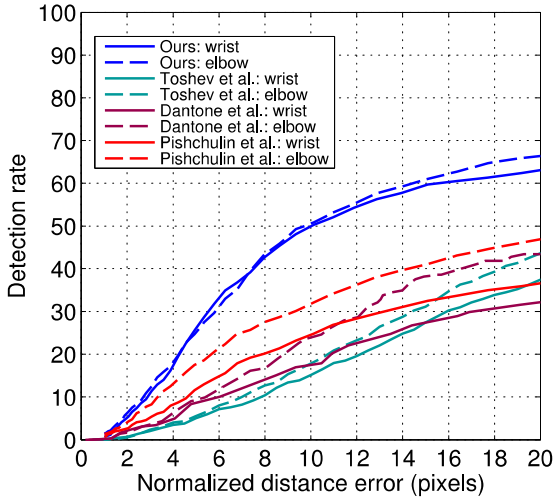
FLIC[1] Elbow



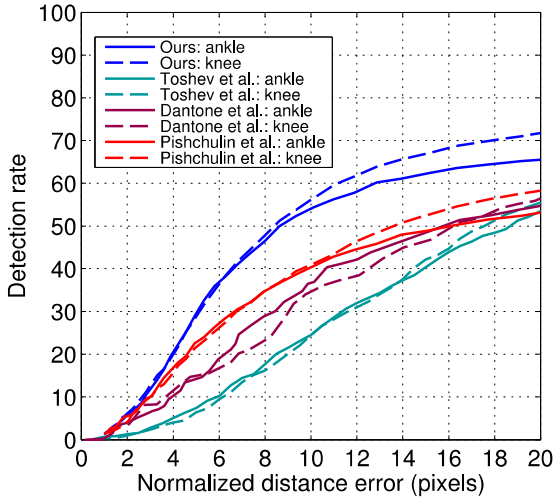
FLIC[1] Wrist



LSP[2] Arms



LSP[1] Legs



(1) B. Sapp and B. Taskar. MODEC: Multimodel decomposition models for human pose estimation. CVPR'13
(2) S. Johnson and M. Everingham. Learning Effective Human Pose Estimation for Inaccurate Annotation. CVPR'11

Image captioning: generating a descriptive sentence

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

C3D: Video Classification with 3D ConvNet

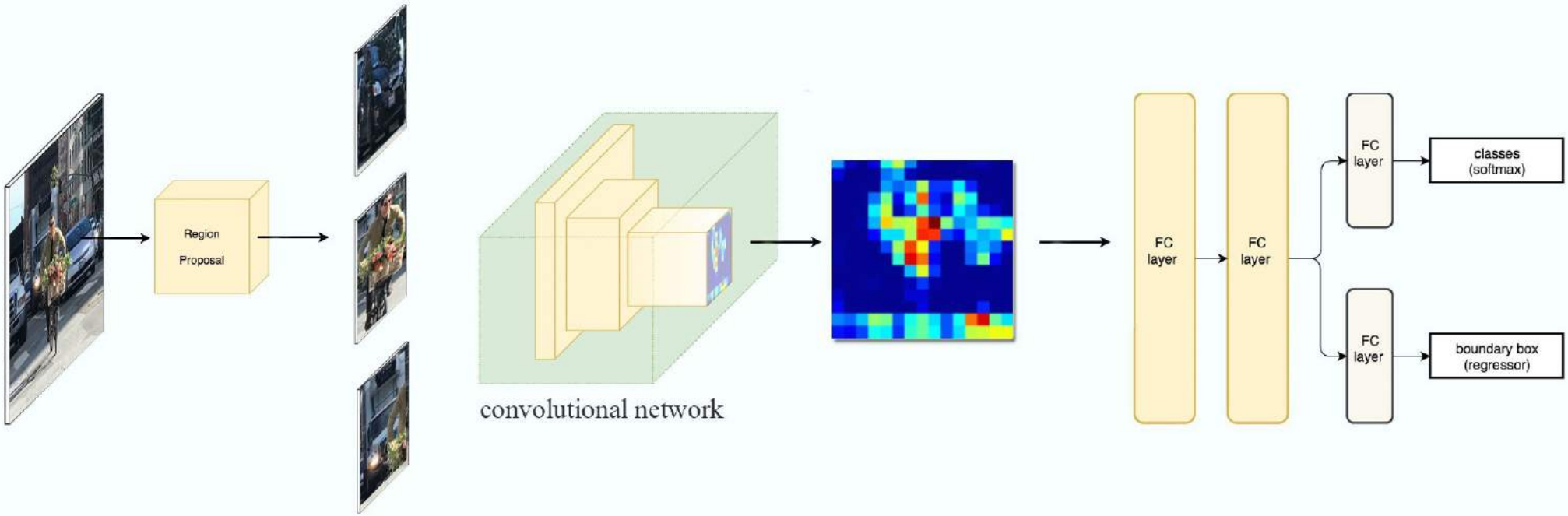
[Tran et al. 2015]

VIDEO: COMMON SPORTS

VIDEO: UNCOMMON SPORTS

R-CNN

[Girshick 2014]



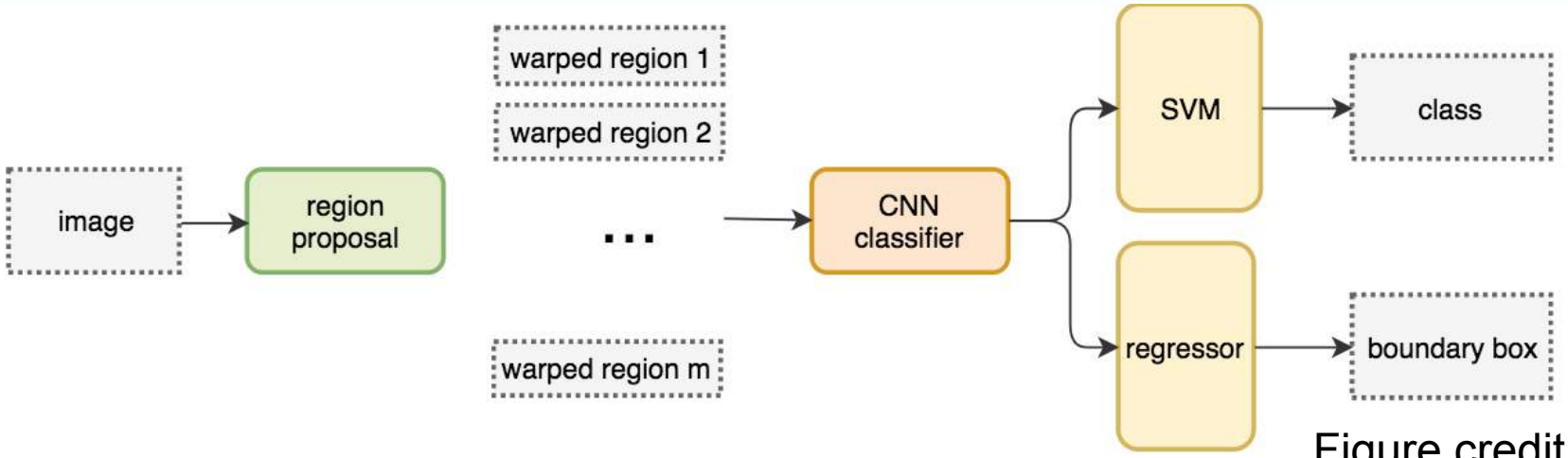
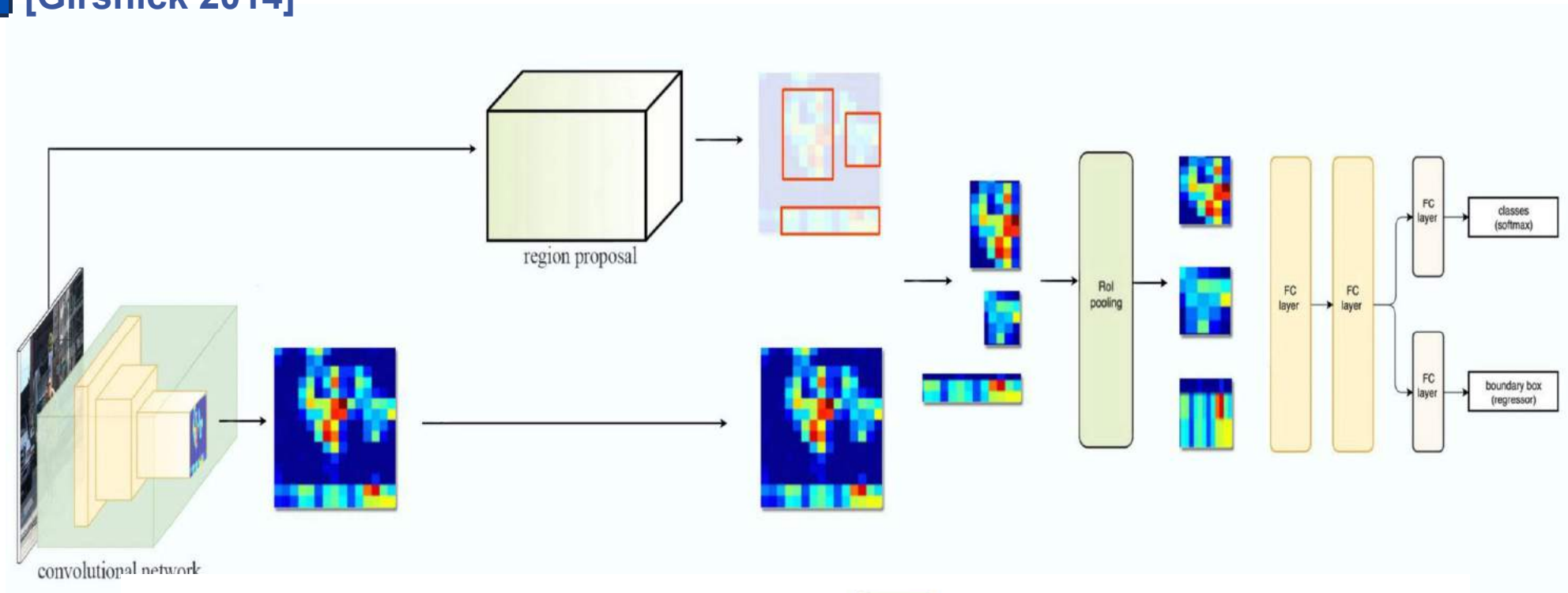


Figure credit: Jonathan Hui

Fast R-CNN

[Girshick 2014]



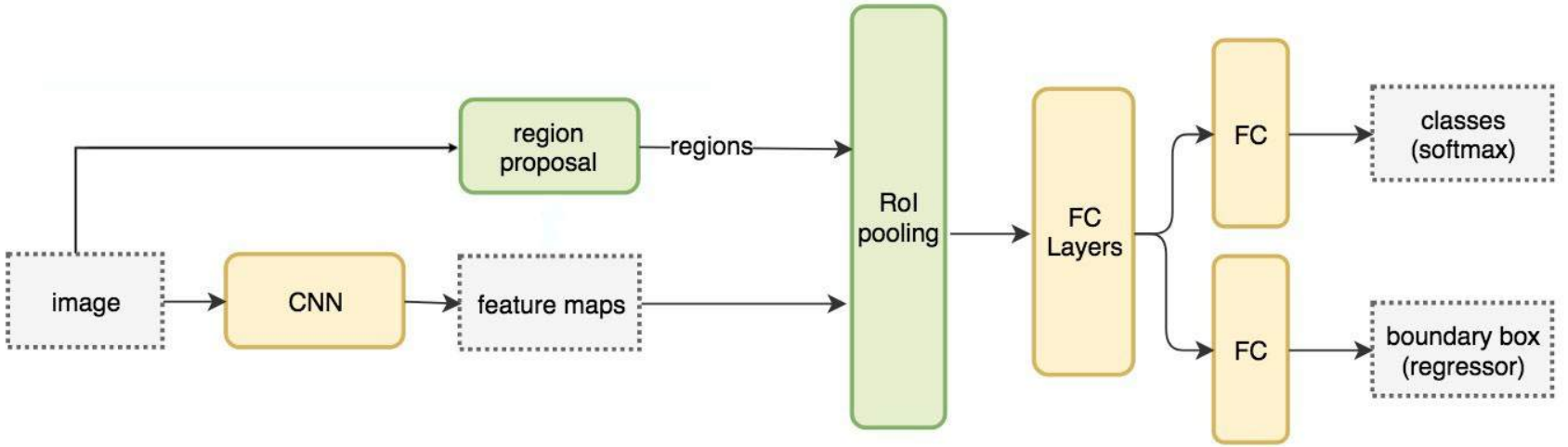
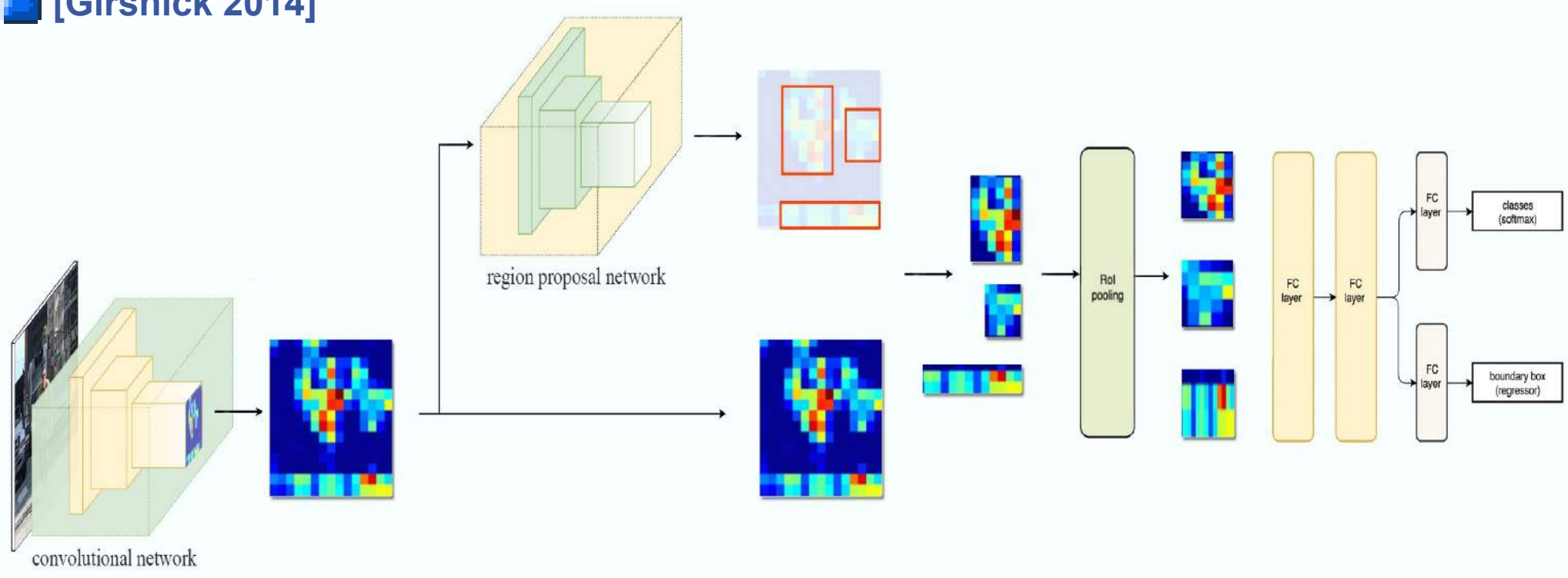


Figure credit: Jonathan Hui

Faster R-CNN

[Girshick 2014]



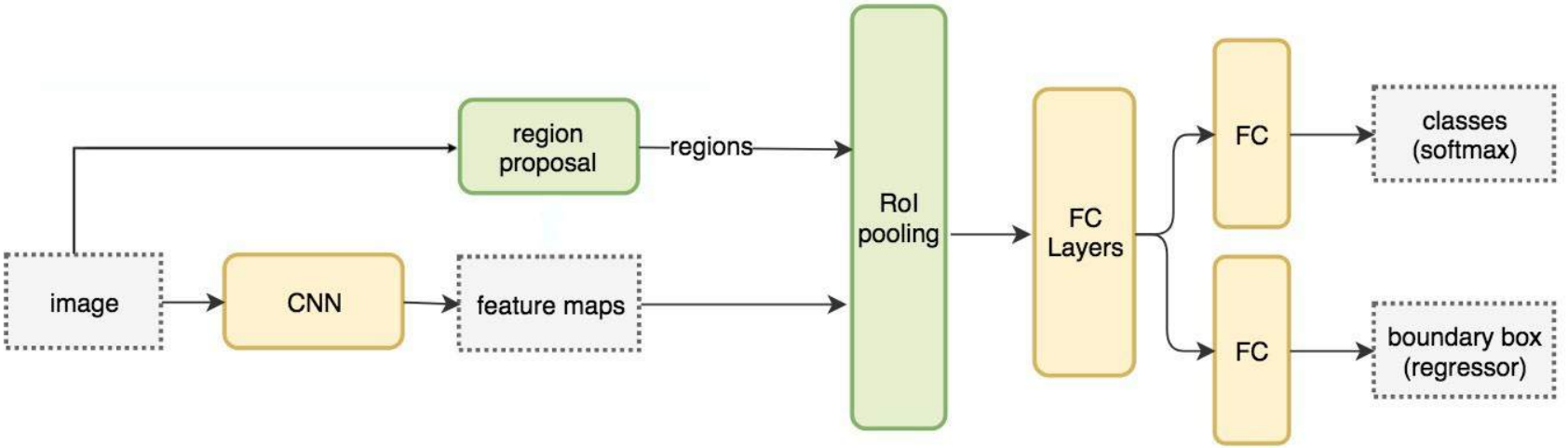
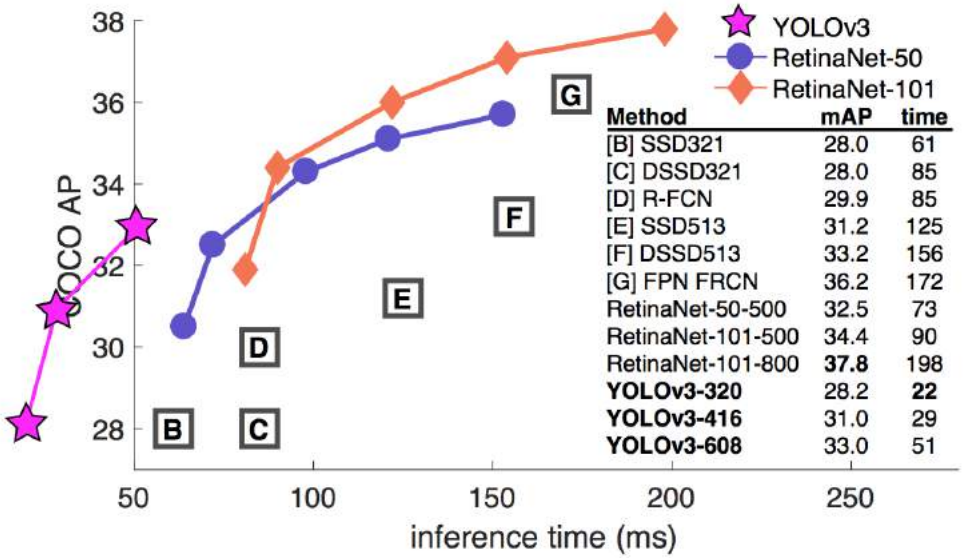


Figure credit: Jonathan Hui

YOLOv1, v2, v3, 9000

[Redmon 2015]

	backbone	AP	AP_{50}
Two-stage methods			
Faster R-CNN+++ [3]	ResNet-101-C4	34.9	55.7
Faster R-CNN w FPN [6]	ResNet-101-FPN	36.2	59.1
Faster R-CNN by G-RMI [4]	Inception-ResNet-v2 [19]	34.7	55.5
Faster R-CNN w TDM [18]	Inception-ResNet-v2-TDM	36.8	57.7
One-stage methods			
YOLOv2 [13]	DarkNet-19 [13]	21.6	44.0
SSD513 [9, 2]	ResNet-101-SSD	31.2	50.4
DSSD513 [2]	ResNet-101-DSSD	33.2	53.3
RetinaNet [7]	ResNet-101-FPN	39.1	59.1



Method	mAP	time
[B] SSD321	28.0	61
[C] DSSD321	28.0	85
[D] R-FCN	29.9	85
[E] SSD513	31.2	125
[F] DSSD513	33.2	156
[G] FPN FRCN	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	**37.8**	198
YOLOv3-320	28.2	**22**
YOLOv3-416	31.0	29
YOLOv3-608	33.0	51

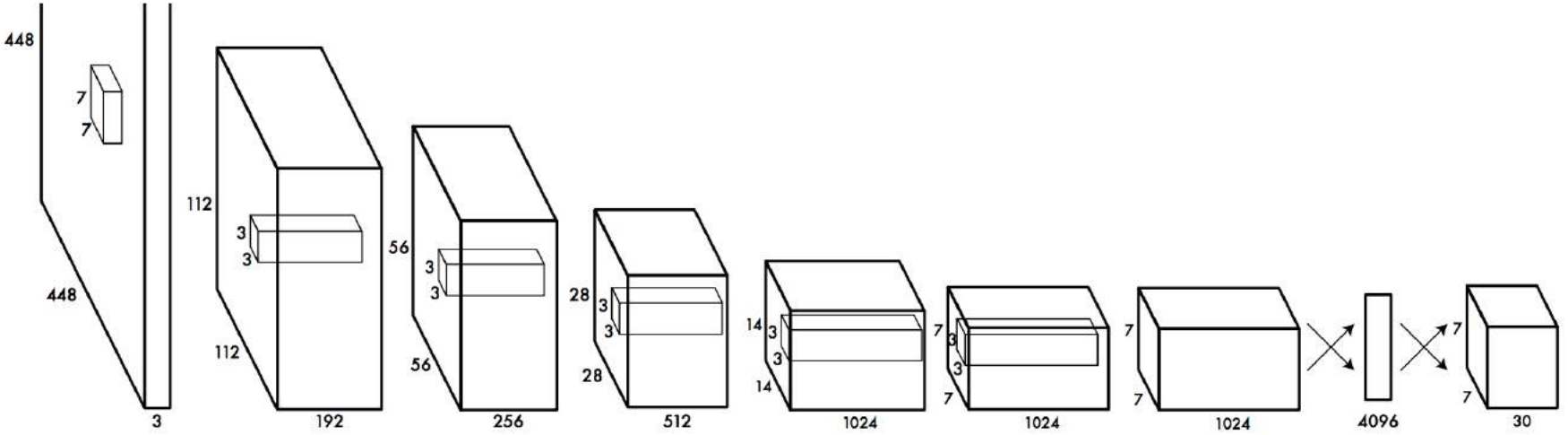
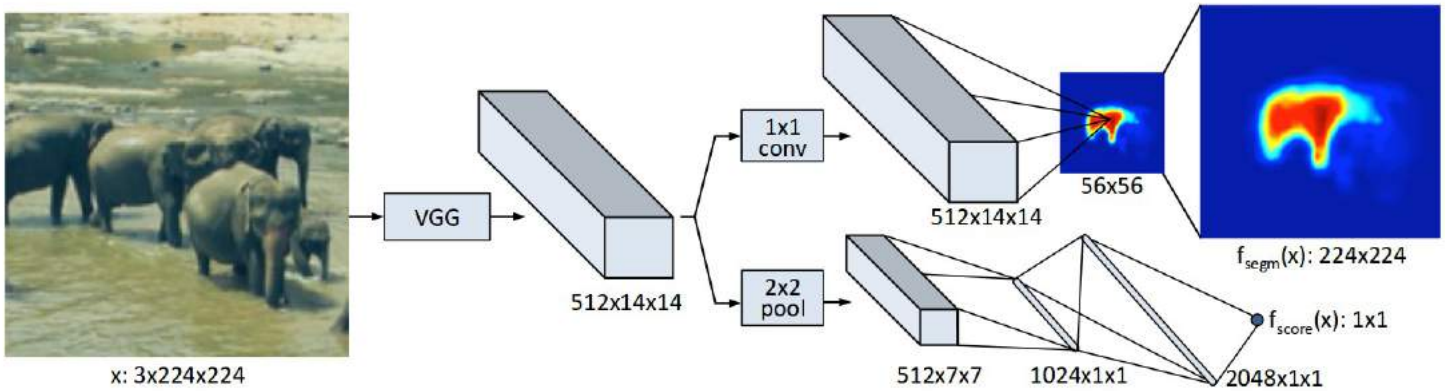


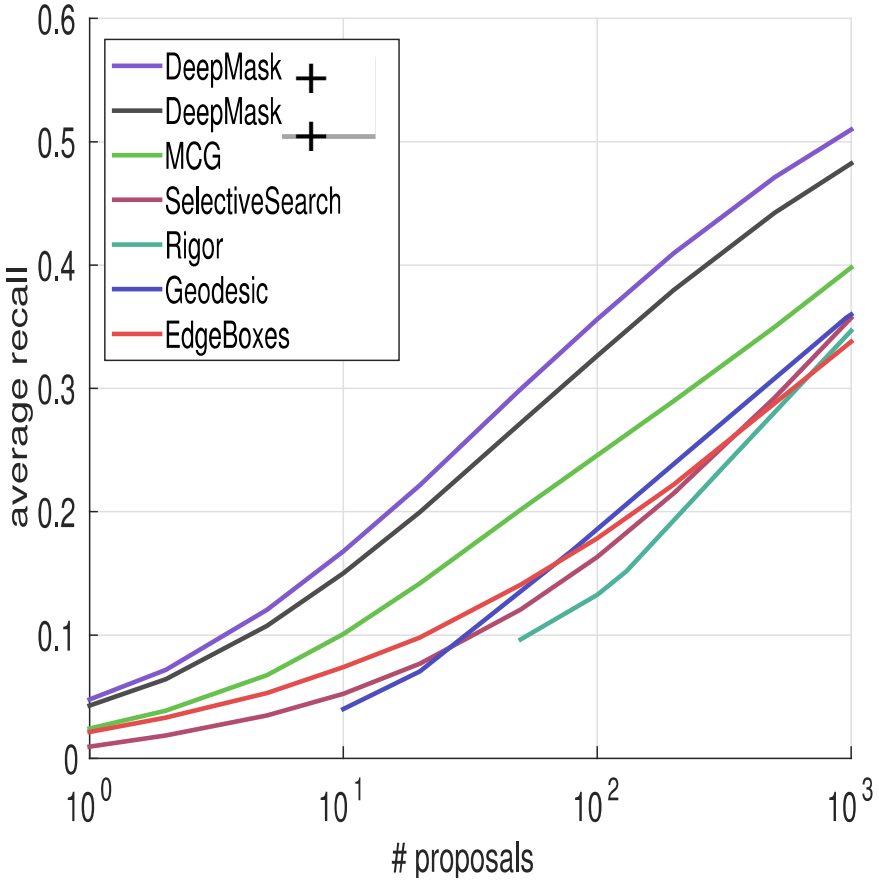
Figure credit: Jonathan Hui

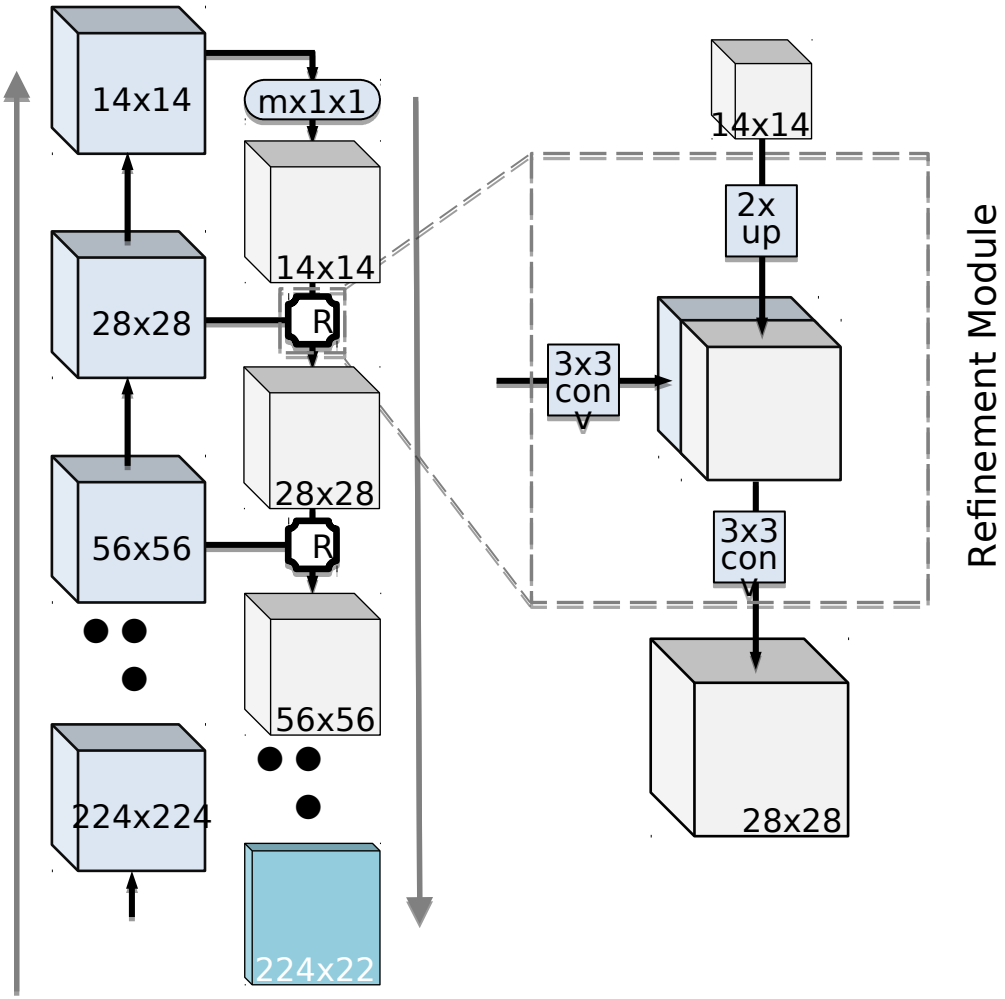
DeepMask: Segmenting and Localizing Objects

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



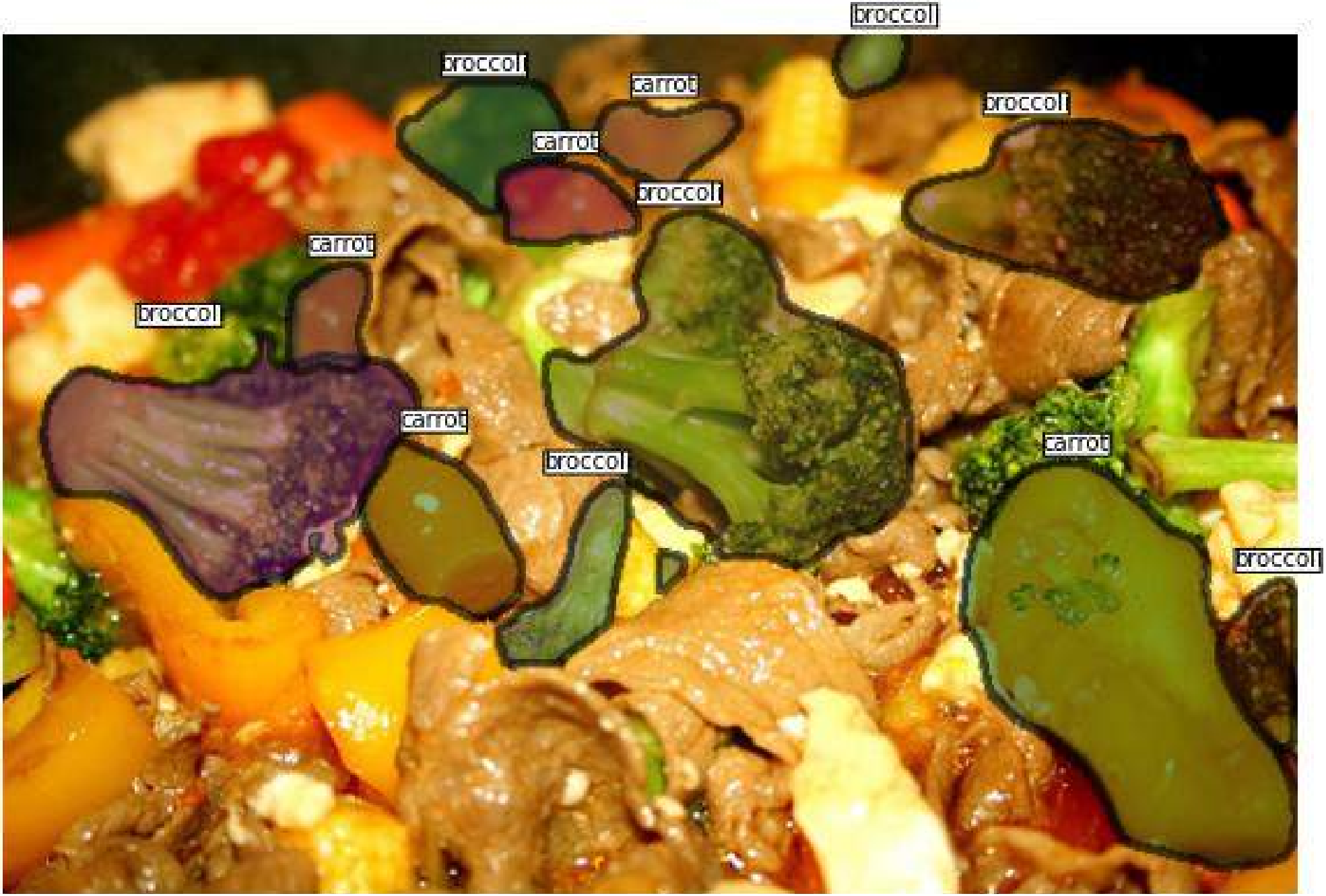
Iterative Refinement: DeepMask++





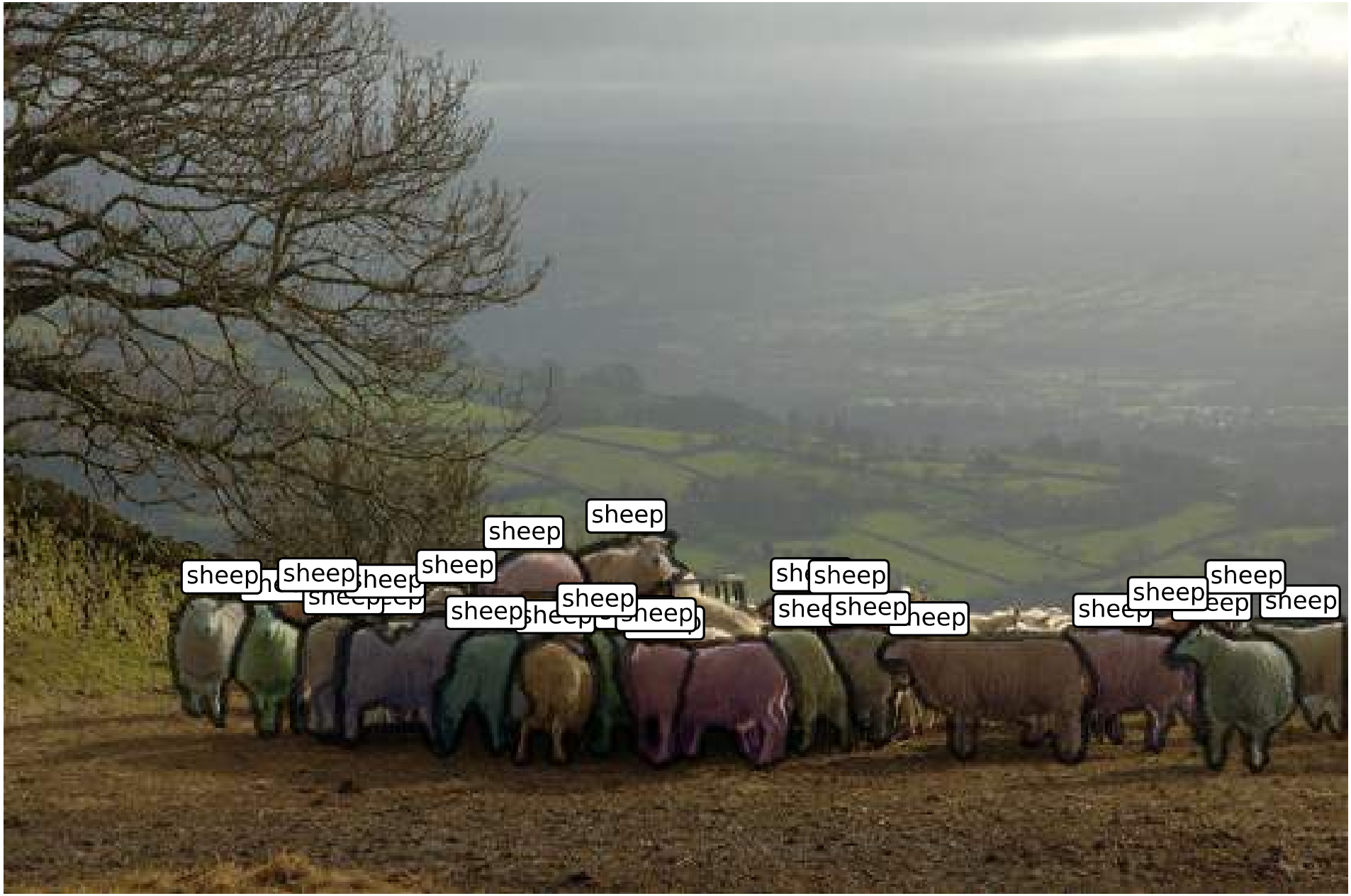
Results

Results


broccoli
broccoli
carrot
broccoli
carrot
broccoli
carrot
broccoli
carrot
broccoli
carrot
broccoli

Results

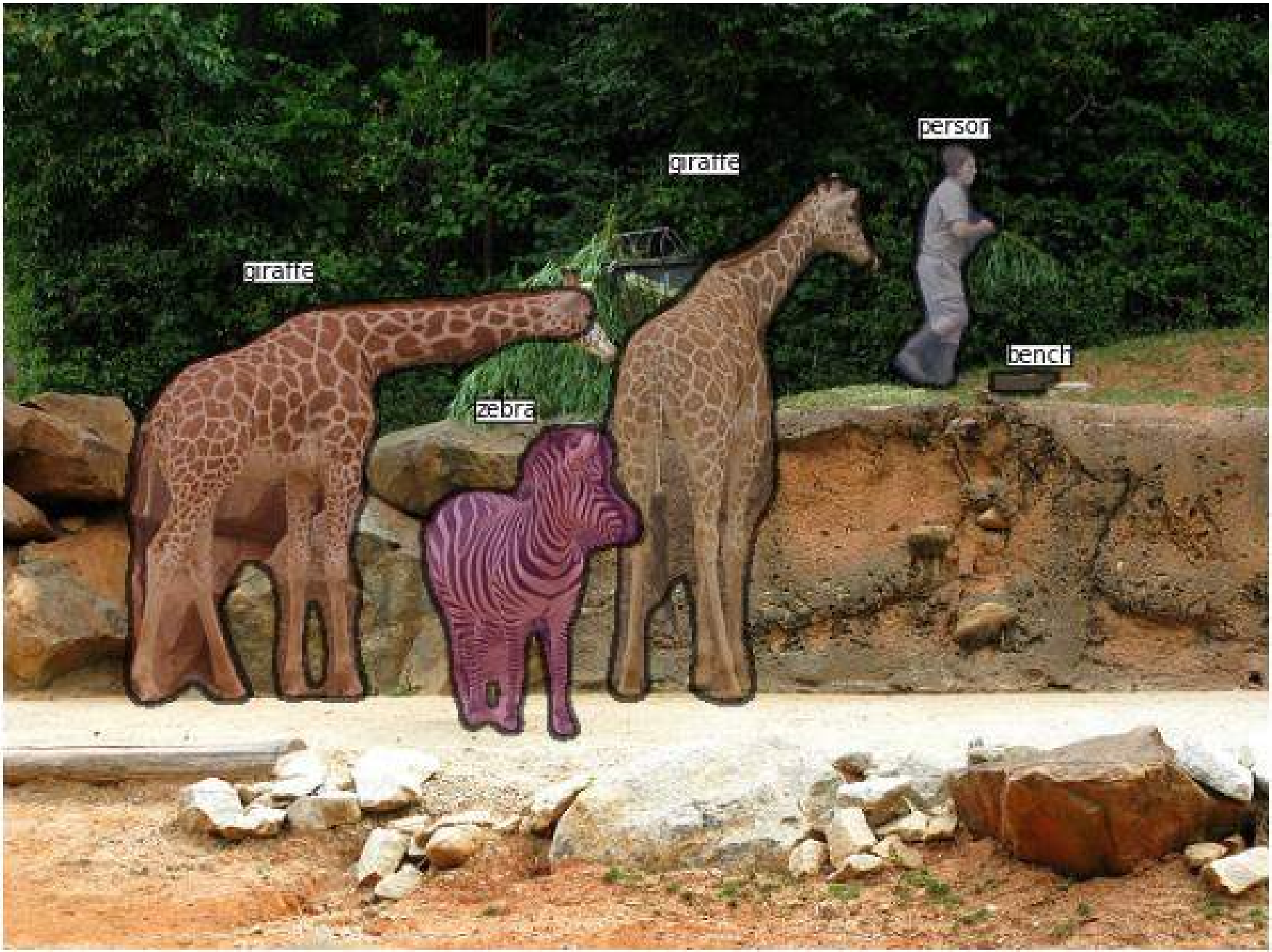
Results


sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep

Results


donut
sandwich
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut

Results

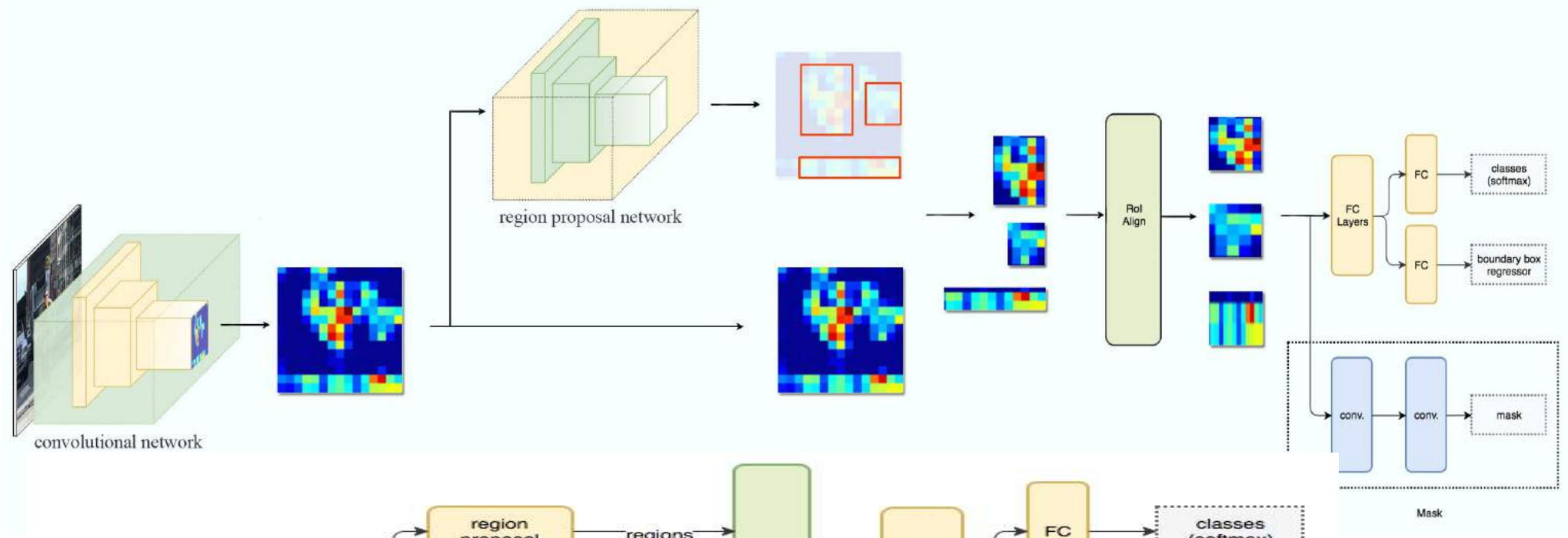


Progress in Computer Vision

- [He 2017]

Mask R-CNN

[He, Gkioxari, Dollar, Girshick arXiv:1703.06870]



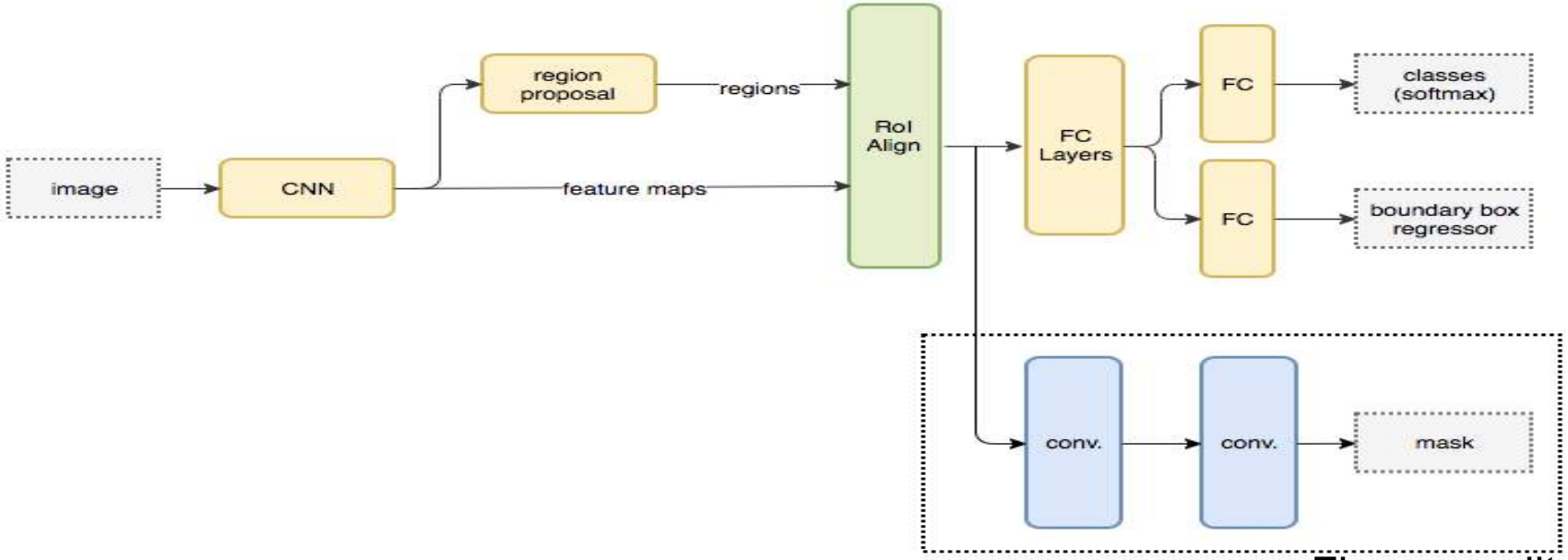
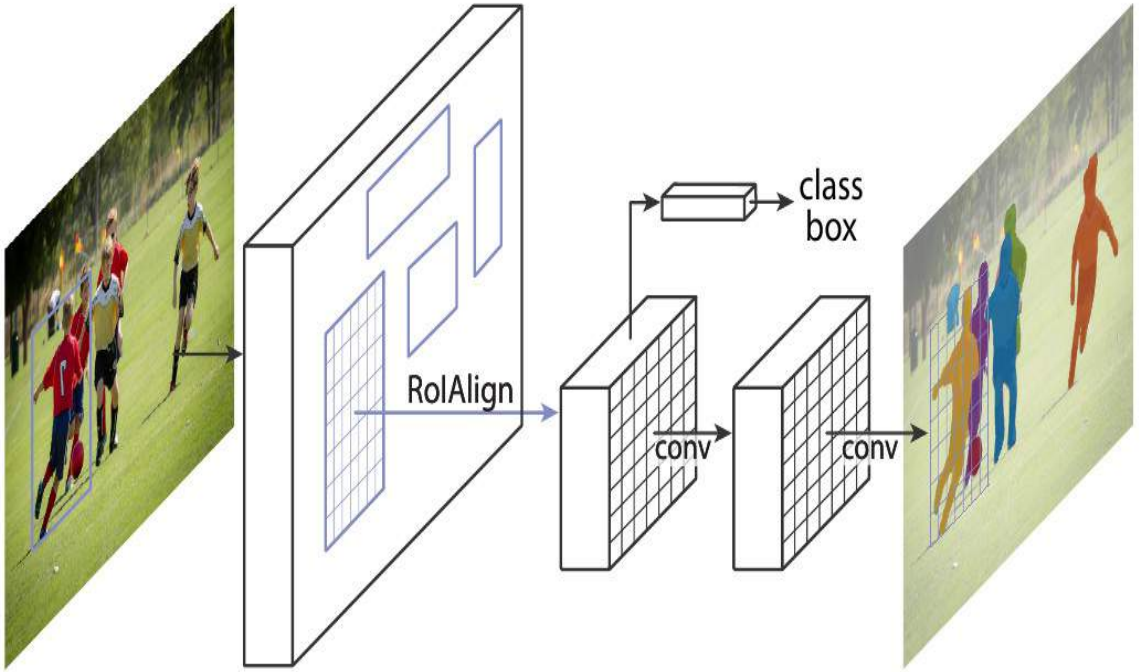


Figure credit: Jonathan Hui

Mask R-CNN: instance segmentation

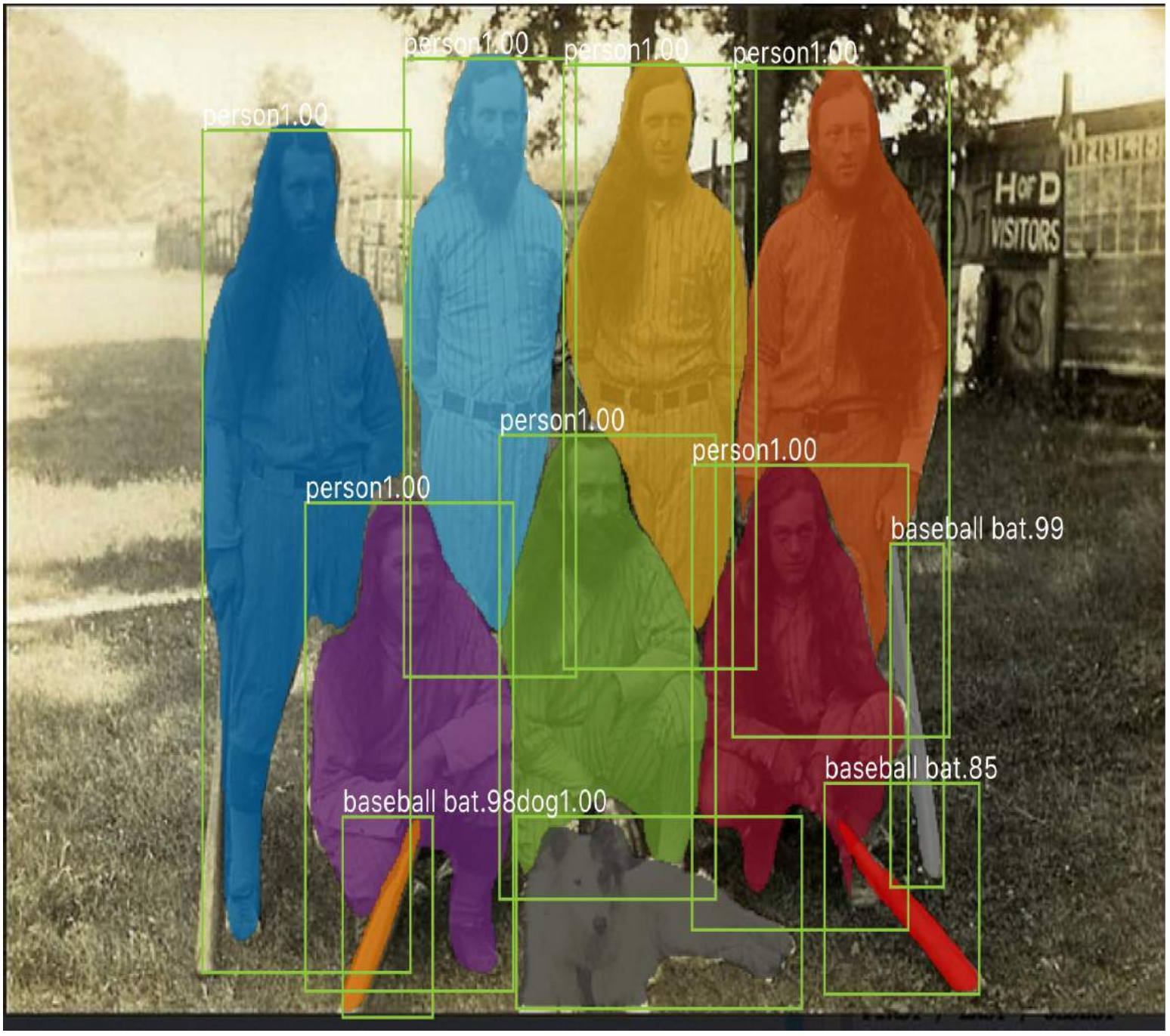
- [He, Gkioxari, Dollar, Girshick arXiv:1703.06870]
- ConvNet produces an object mask for each region of interest
- Combined ventral and dorsal pathways



	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [20] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

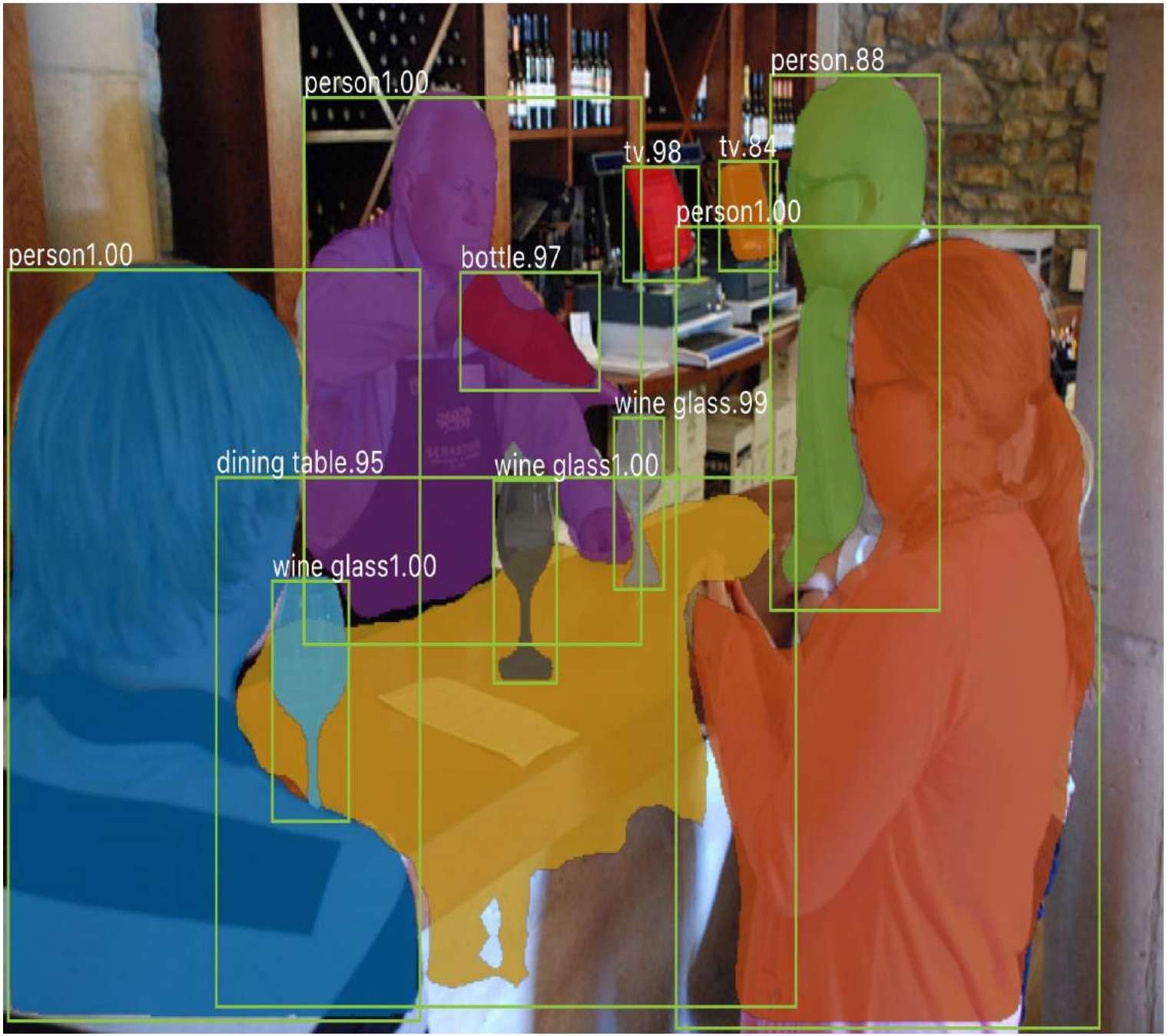
Mask-RCNN Results on COCO dataset

- **Individual objects are segmented.**



Mask-RCNN Results on COCO dataset

- **Individual objects are segmented.**



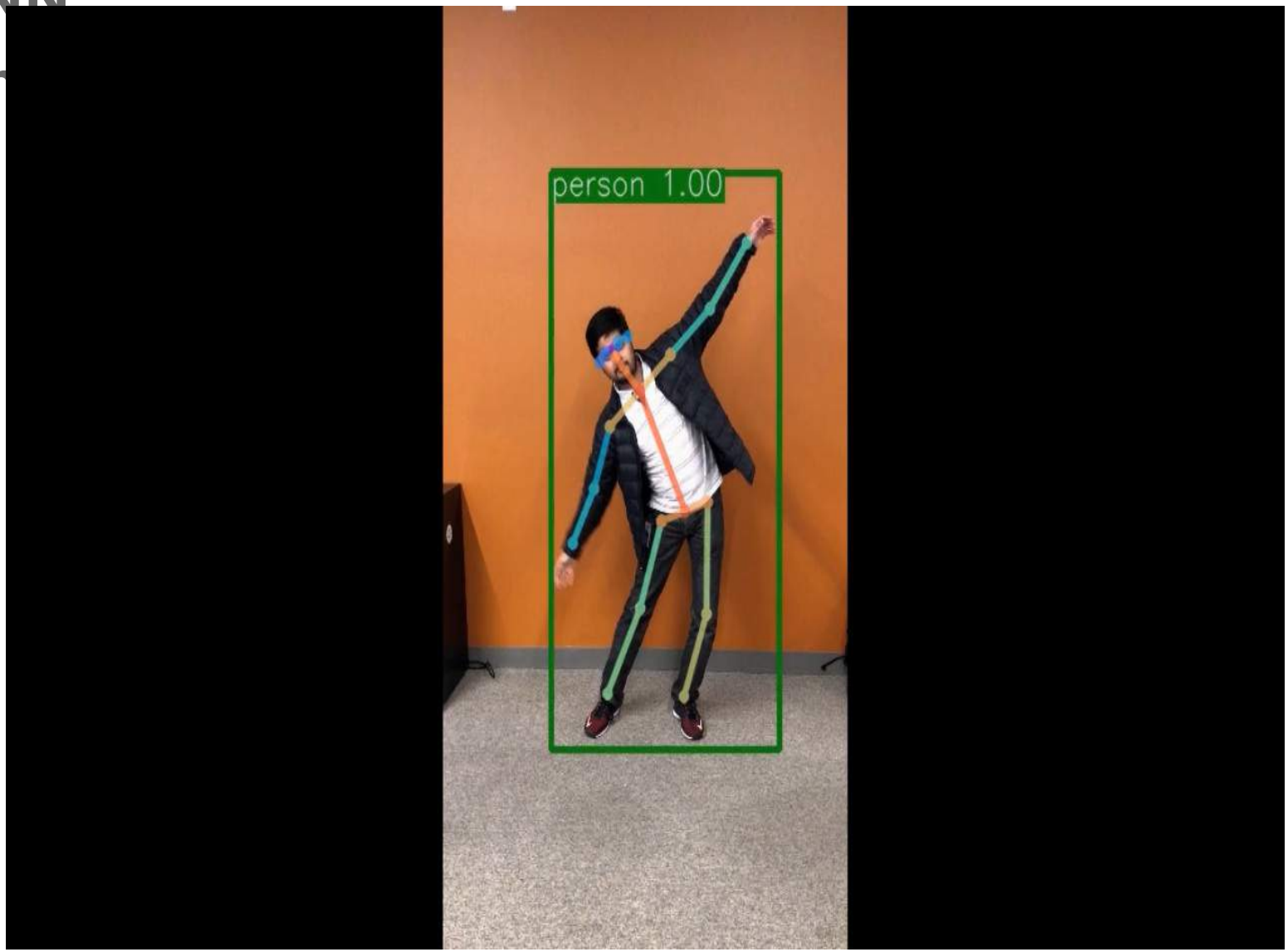
Mask R-CNN Results on COCO test set

Mask R-CNN Results on COCO test set

Figure 4. More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

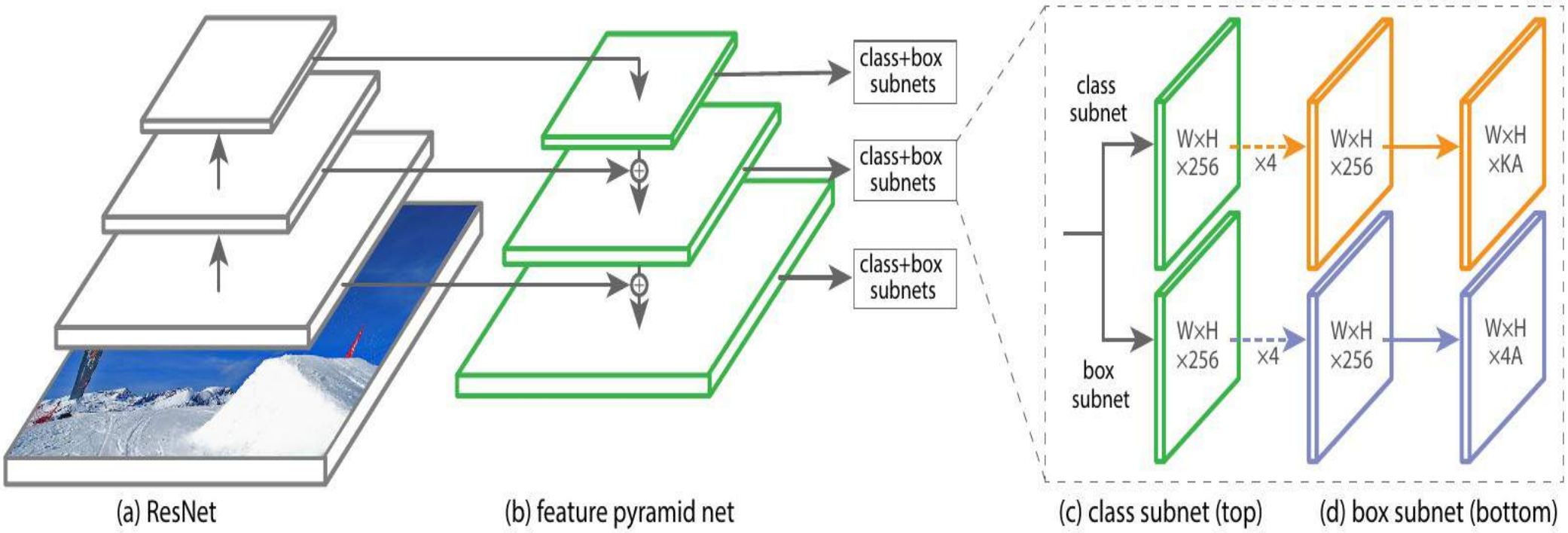
Real-Time Pose Estimation on Mobile Devices

- **Maks R-CNN running on Caffe2Go**



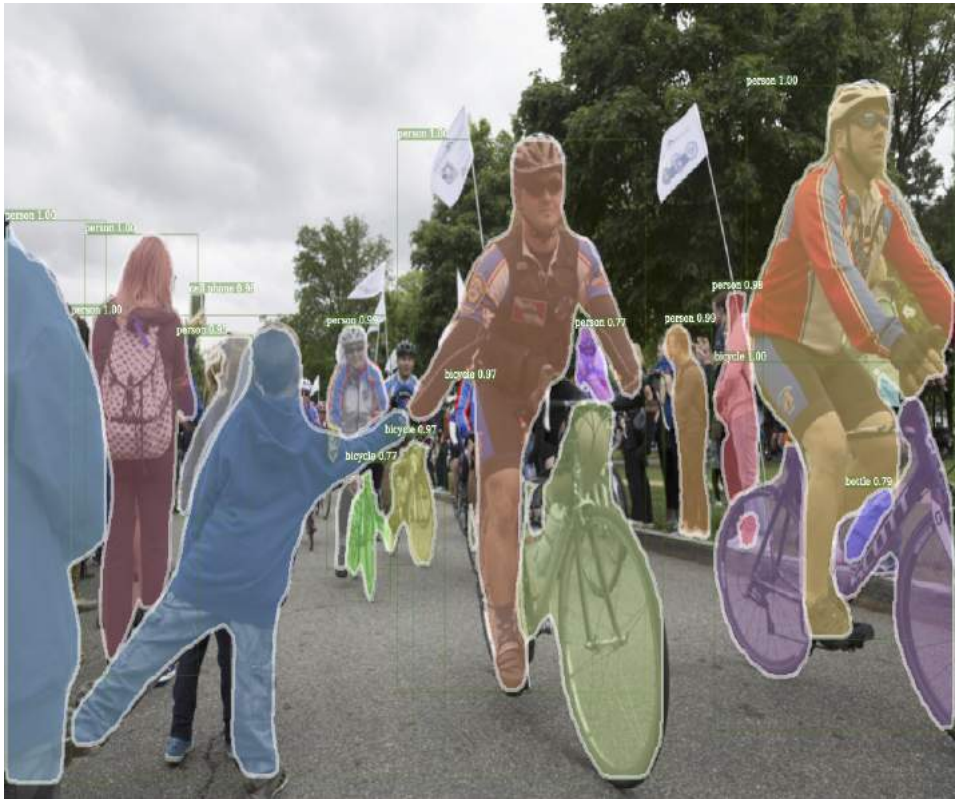
RetinaNet, feature pyramid network

- One-pass object detection
- [Lin et al. ArXiv:1708.02002]



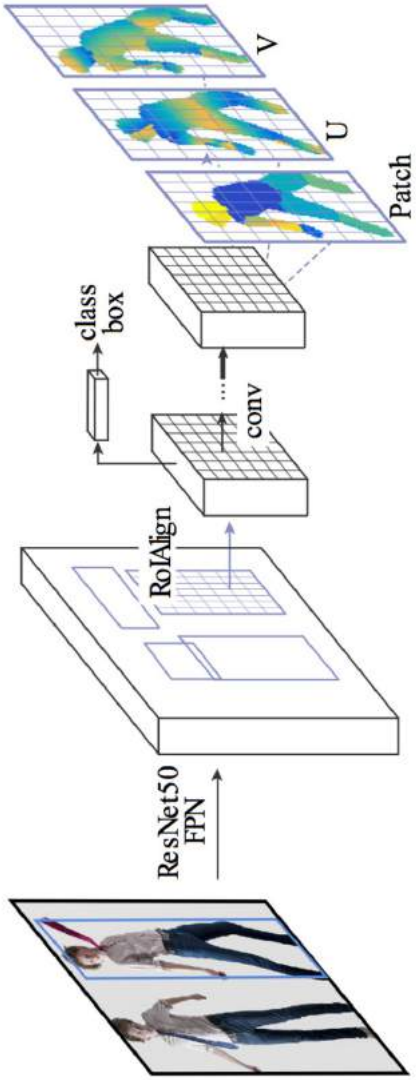
Detectron: open source vision in PyTorch

https://github.com/facebookresearch/maskrcnn-benchmark



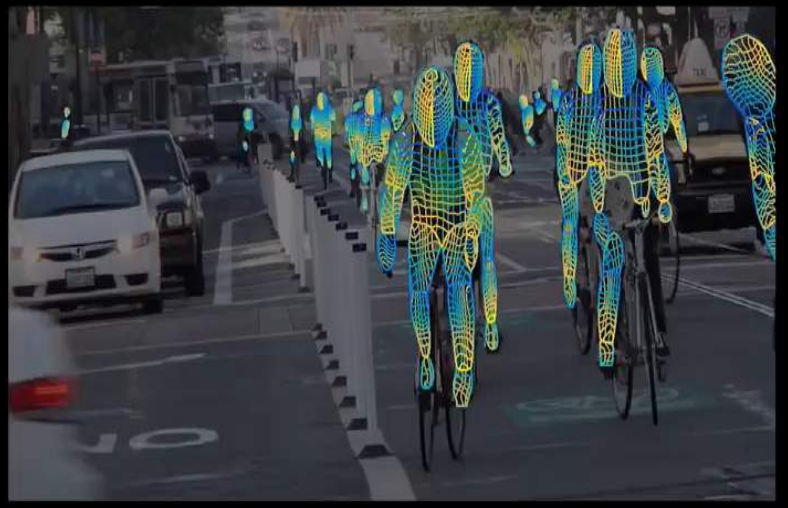
DensePose: real-time body pose estimation

- **[Guler, Neverova, Kokkinos CVPR 2018]**
  http://densepose.org 20 fps on a single GPU



DensePose:

Dense Human Pose Estimation In The Wild



Rıza Alp Güler * Natalia Neverova Iasonas Kokkinos
INRIA, CentraleSupélec *Facebook AI Research* *Facebook AI Research*

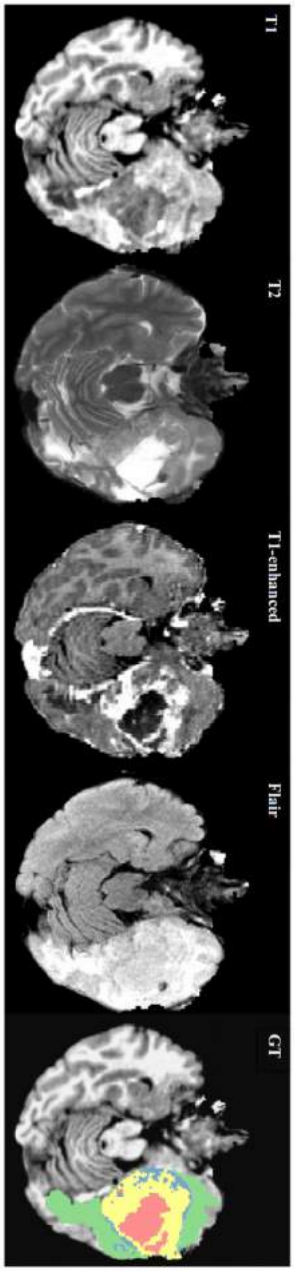
* Rıza Alp Güler was with Facebook AI Research during this work.

ConvNet in Connectomics [Jain, Turaga, Seung 2007]

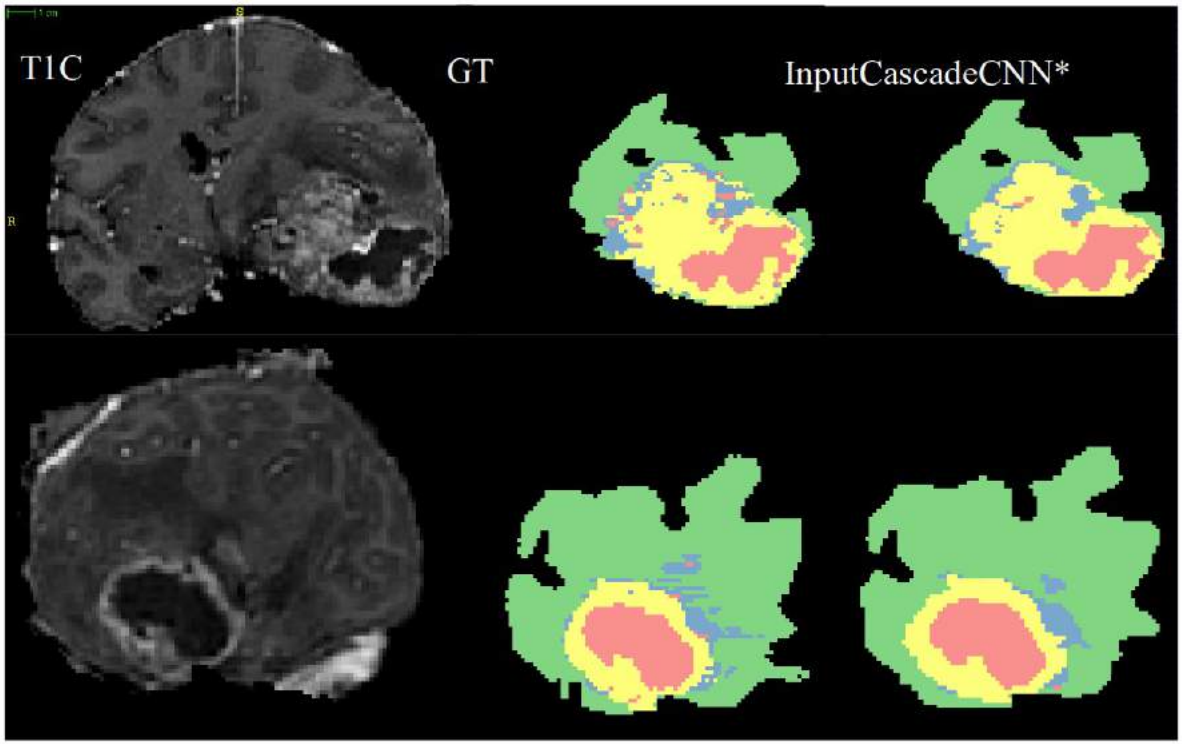
- 3D ConvNet
  - Volumetric
  - Images
- Each voxel labeled as "membrane" or "non-membrane using a 7x7x7 voxel neighborhood
- Has become a standard method in connectomics

VIDEO

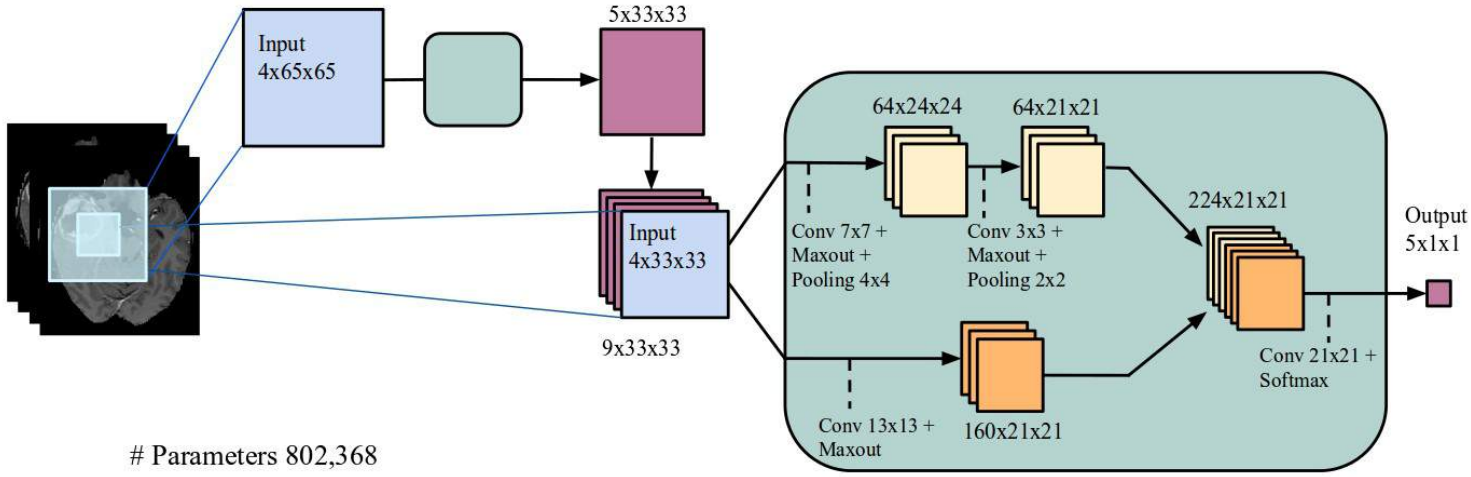
Brain Tumor Detection



- **[Havaei et al. 2015]**
  - Arxiv:1505.03540
- **InputCascadeCNN architecture**
  - 802,368 parameters
- **Trained on 30 patients.**
- **State of the art results on BR**



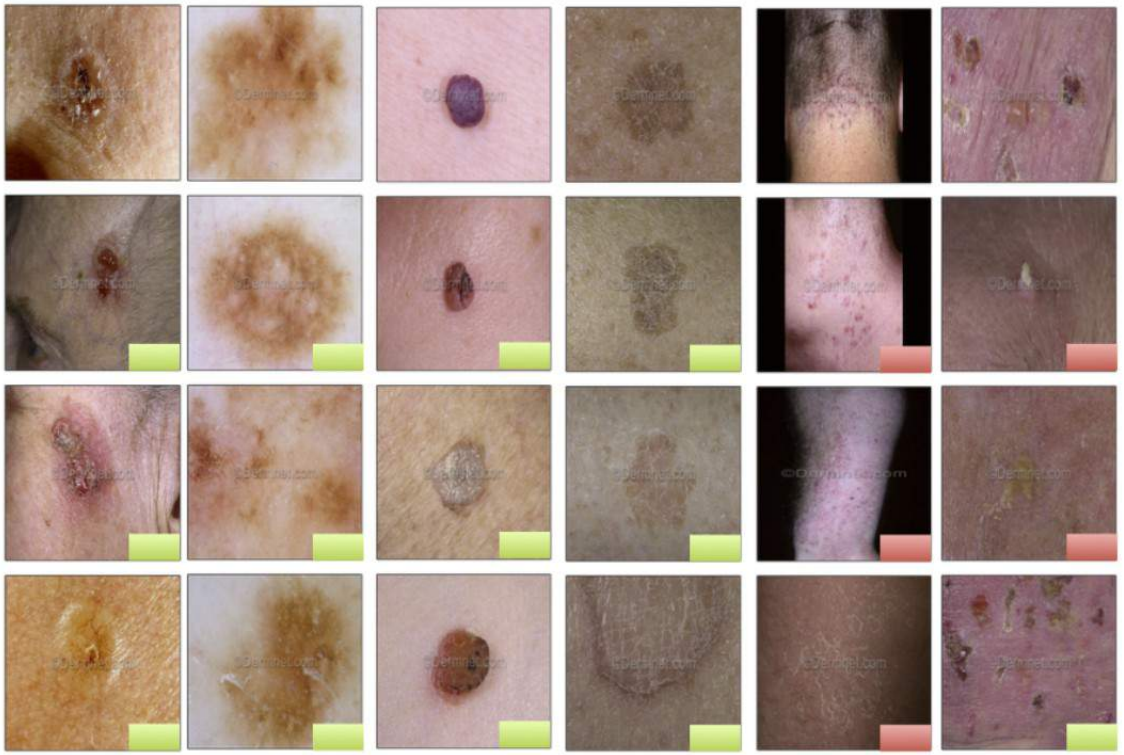
edema, enhanced tumor, necrosis, non-enhanced tumor.

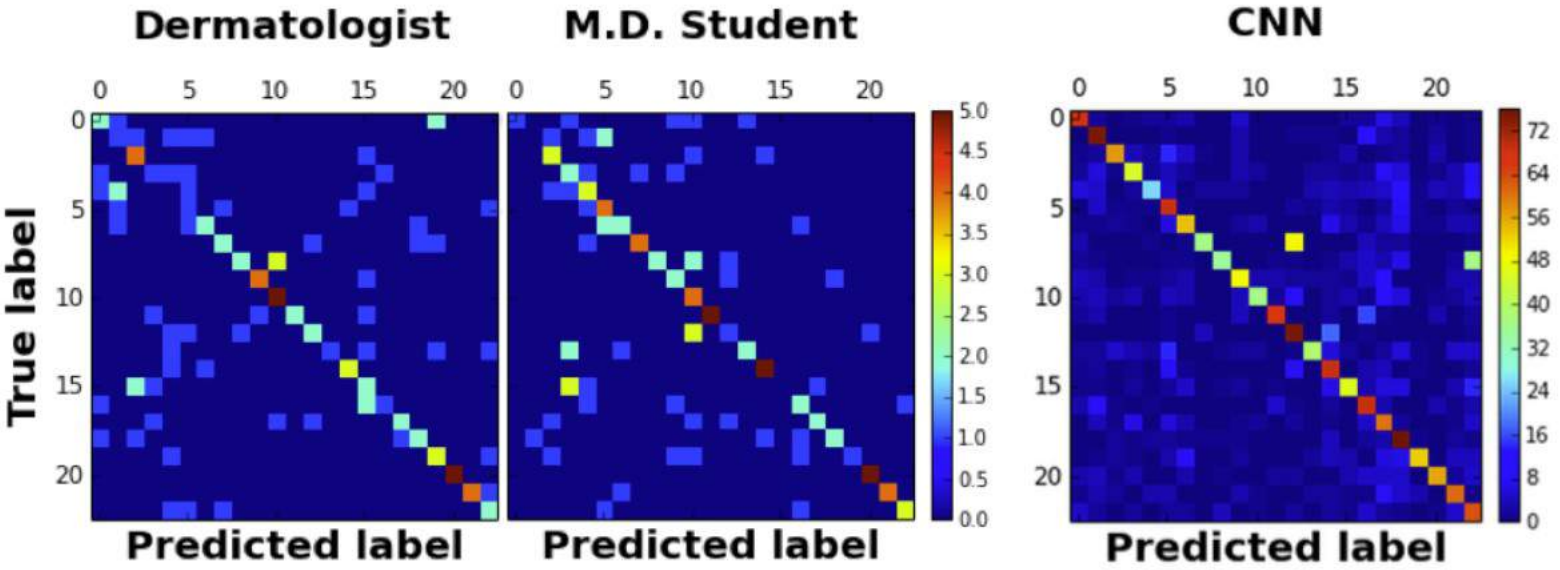


Parameters 802,368

Skin Cancer Detection

- **[Esteva, Kuprel, Thrun 2015]**
- **Trained on 23,000 images**
  - 23 categories
- **Accuracy (disease):**
  - Dermatologist: 46.5%
  - Dermato Resident: 52.7%
  - ConvNet Ensemble: 60.0%
- **Accuracy (cancer detection)**

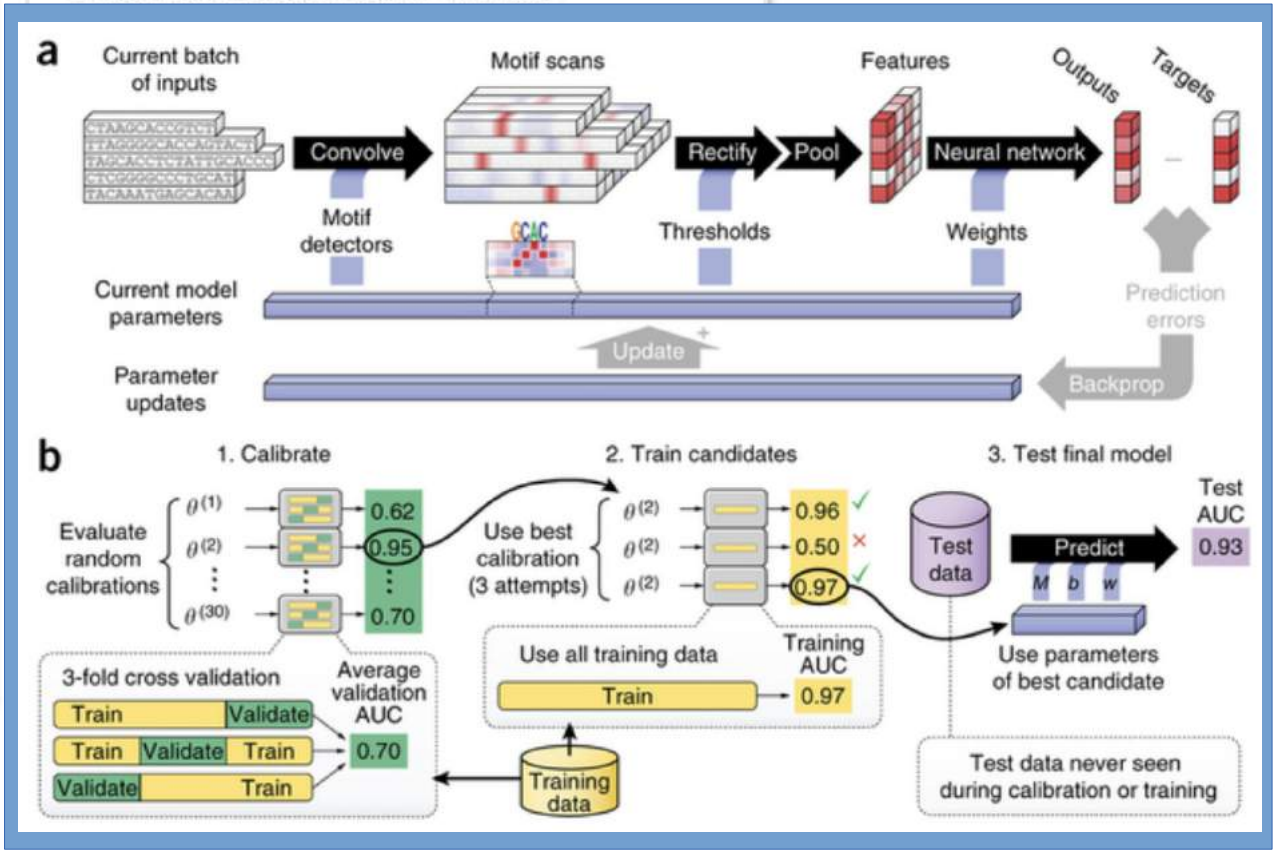




Predicting DNA/RNA – Protein Binding with ConvNets

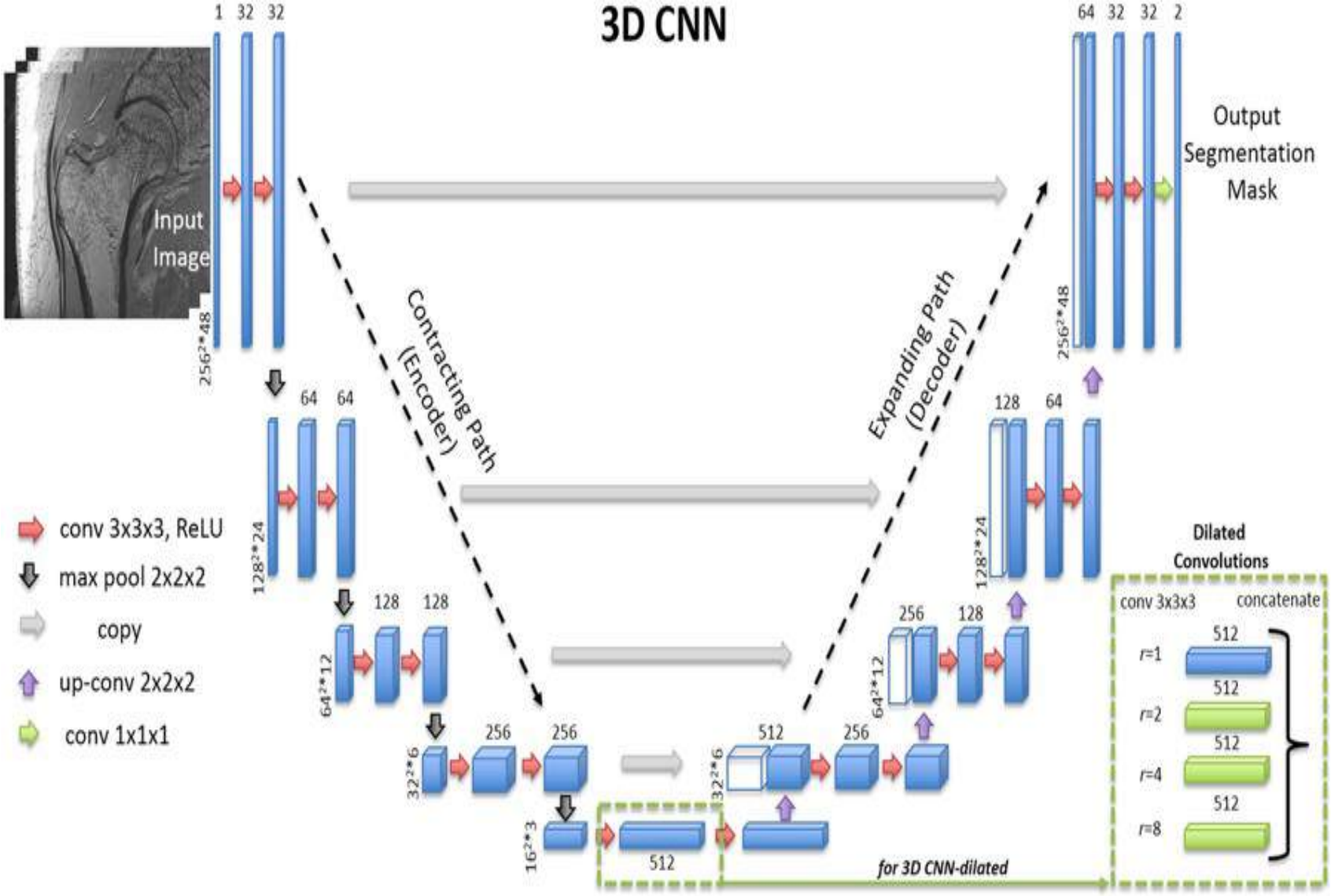
Y LeCun

"Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning" by B Alipanahi, A Delong, M Weirauch, B Frey,

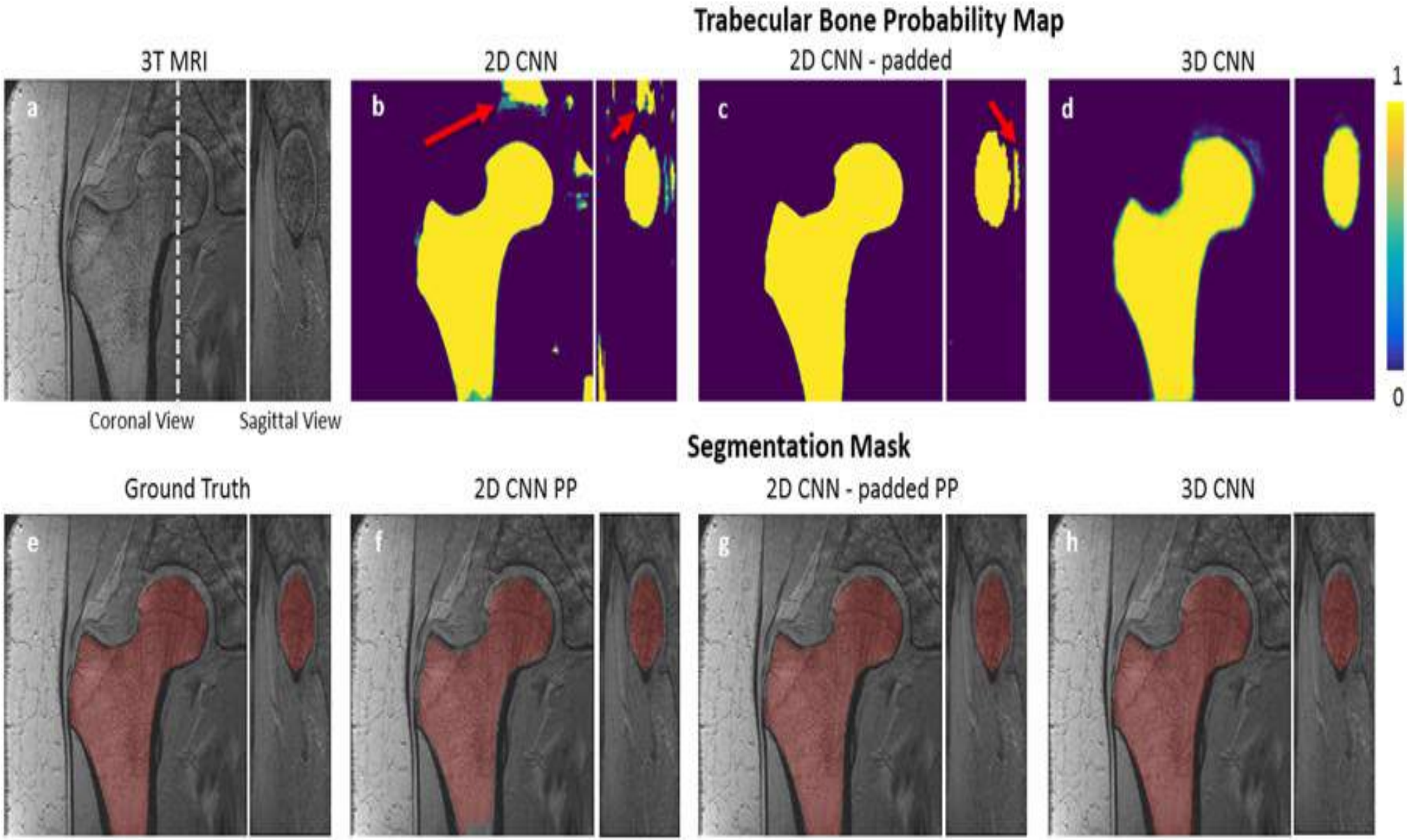


3D ConvNet for Medical Image Analysis

- Segmentation Femur from MR Images
- [Deniz et al. Nature 2018]



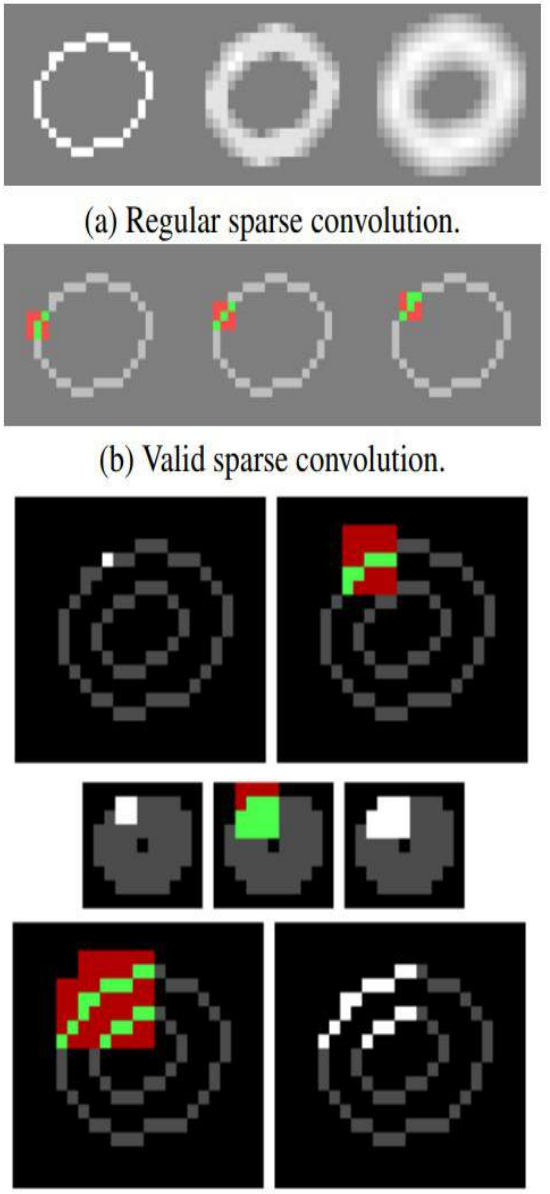
3D ConvNet for Medical Image Analysis



3D Semantic Segmentation with Sparse ConvNets

- **ShapeNet competition results ArXiv:1710.06104]**
- **Winner: Submanifold Sparse ConvNet**
  - [Graham & van der M[illegible]

method	mean
SSCN	**86.00**
PdNet	85.49
DCPN	84.32
PCNN	82.29
PtAdLoss	77.96
KDTNet	65.80
DeepPool	42.79
NN	77.57
[19]	84.74



(a) Regular sparse convolution.

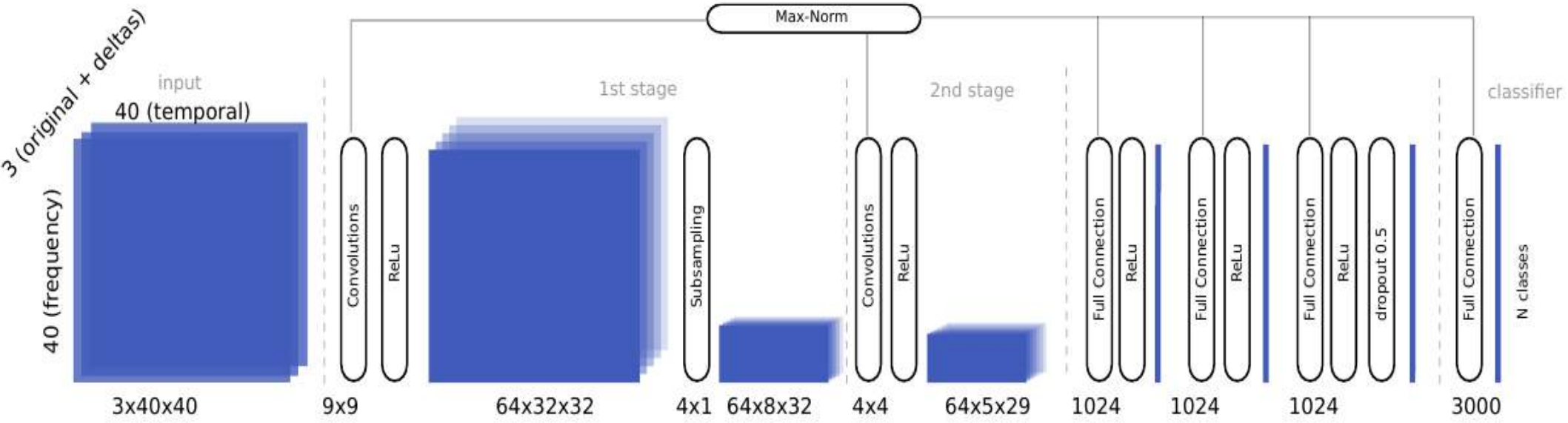
(b) Valid sparse convolution.

) Block with a strided, a valid, and a de-convolution.


Speech Recognition
With
ConvNets

Speech Recognition with Convolutional Nets (NYU/IBM)

Y LeCun

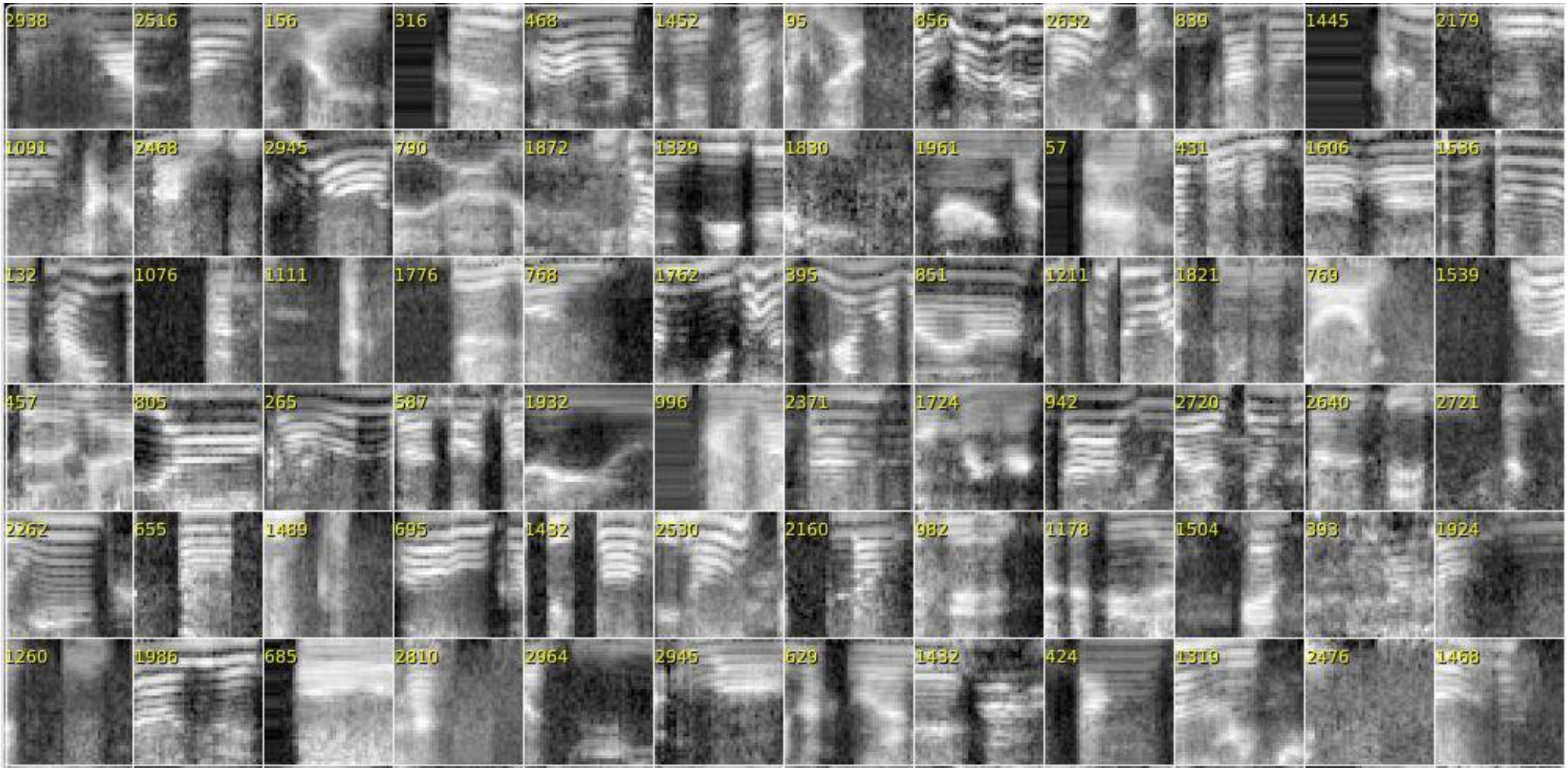


- **Acoustic Model: ConvNet with 7 layers. 54.4 million parameters.**
- **Classifies acoustic signal into 3000 context-dependent subphones categories**
- **ReLU units + dropout for last layers**
- **Trained on GPU. 4 days of training**

Speech Recognition with Convolutional Nets (NYU/IBM)

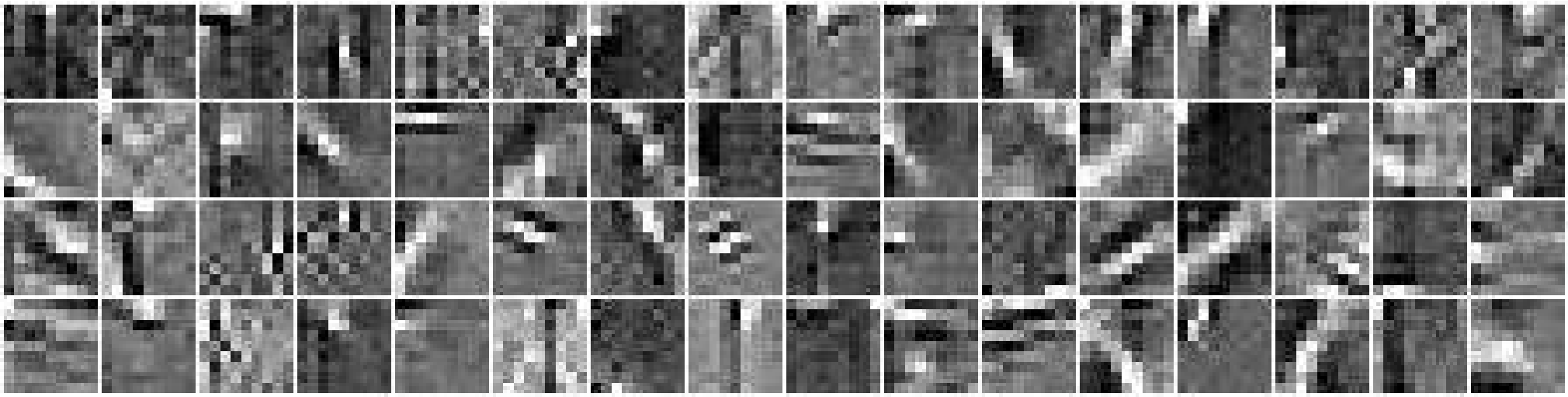
Training samples.

- 40 MEL-frequency Cepstral Coefficients
- Window: 40 frames, 10ms each



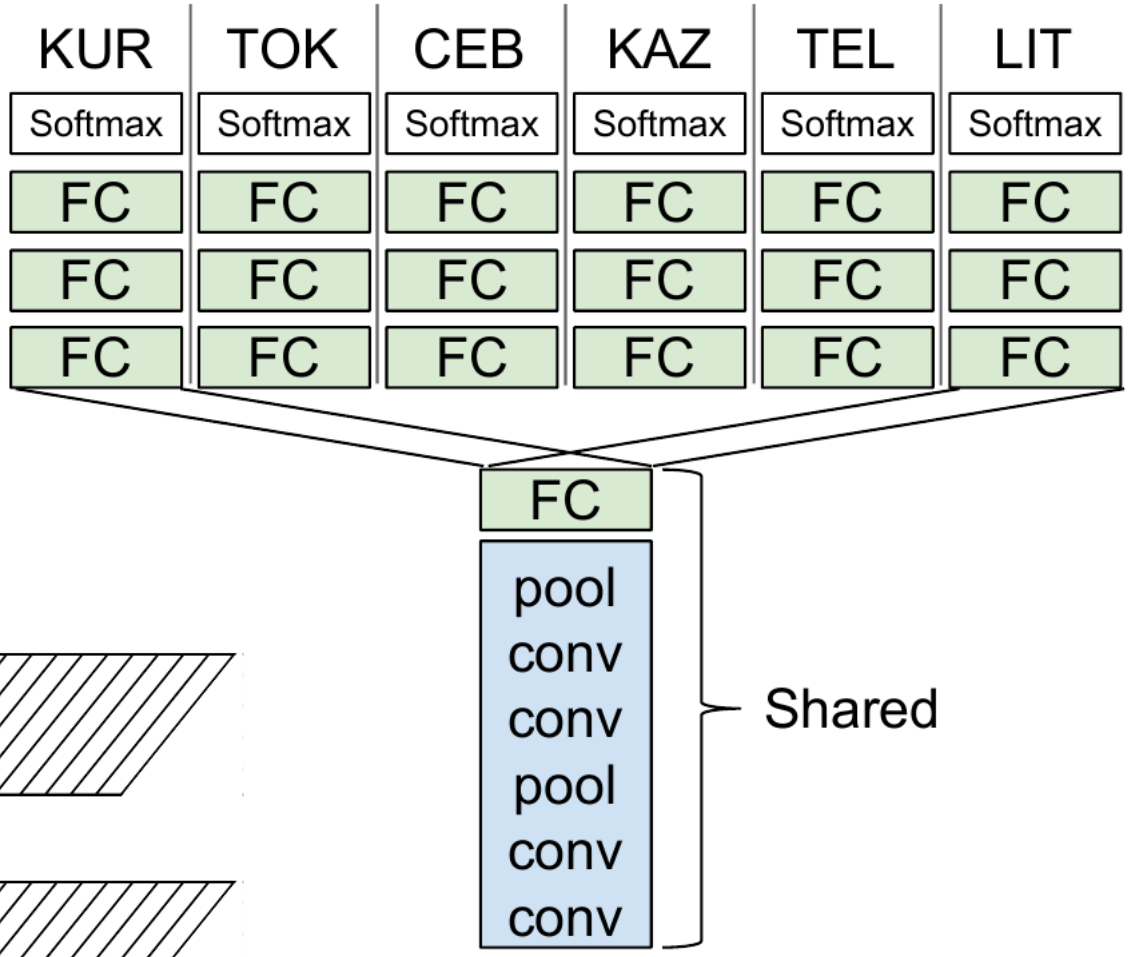
Speech Recognition with Convolutional Nets (NYU/IBM)

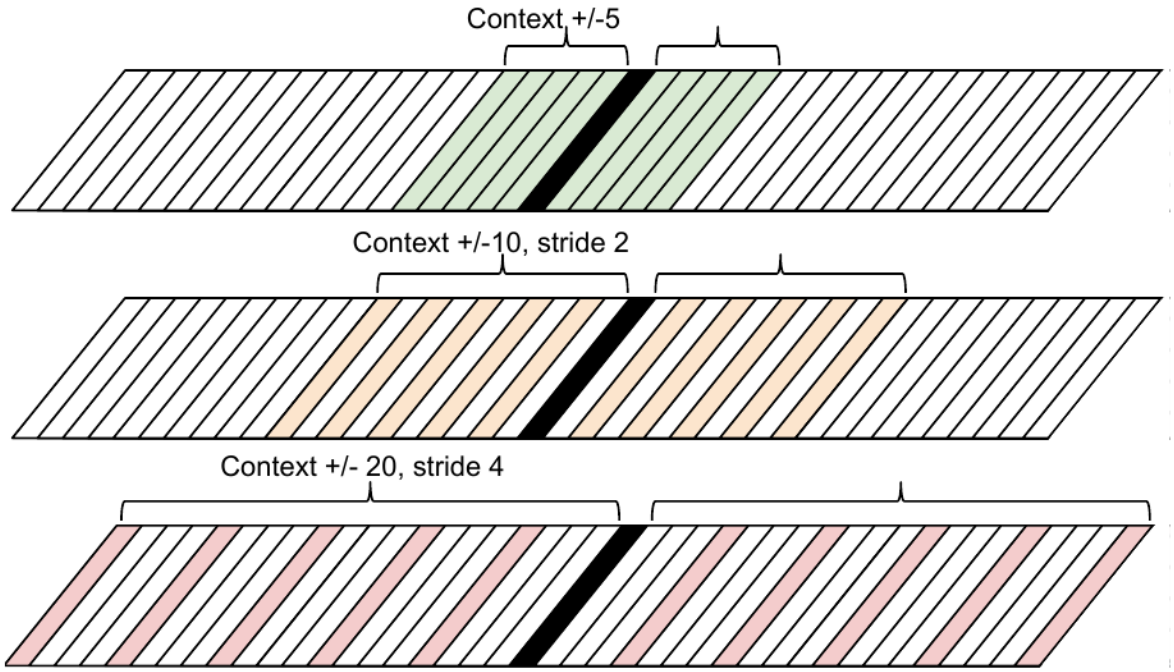
- **Convolution Kernels at Layer 1:**
  - 64 kernels of size 9x9



Speech Recognition with Convolutional Nets (NYU/IBM)

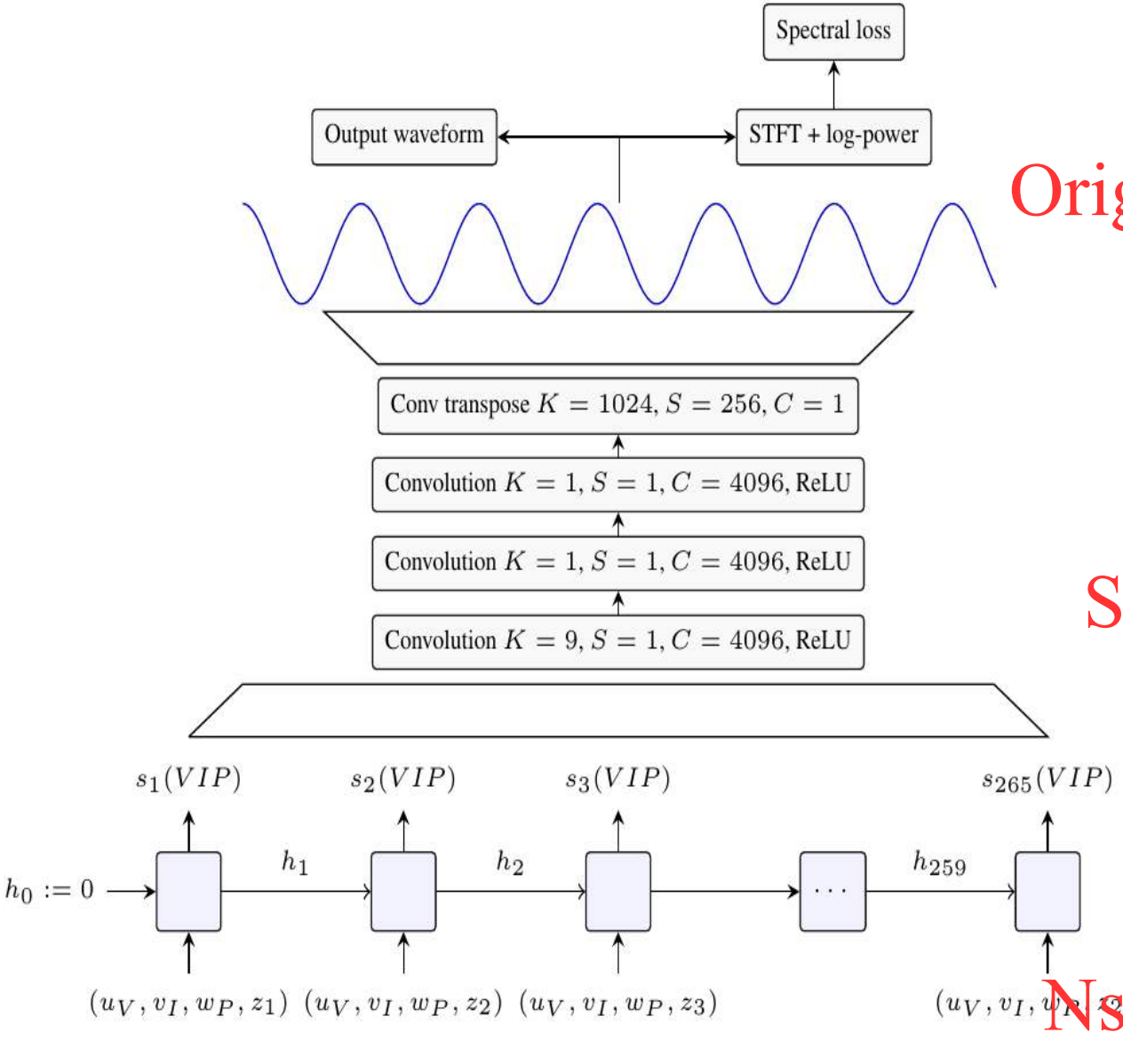
- **Multilingual recognizer**
- **Multiscale input**
  - Large context window



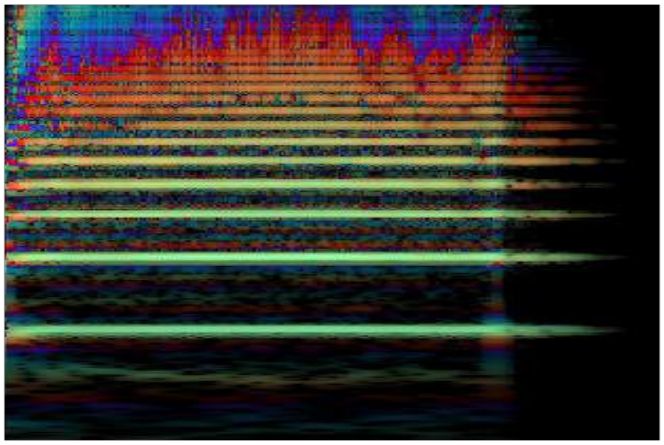


SING: Symbol to Instrument Neural Generator

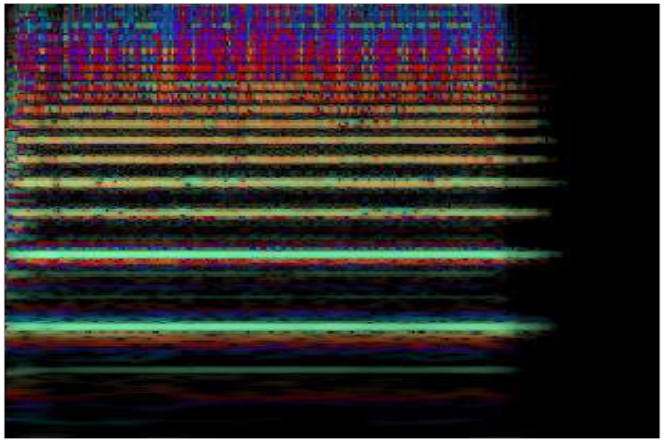
- **[Defossez NeurIPS 18]**



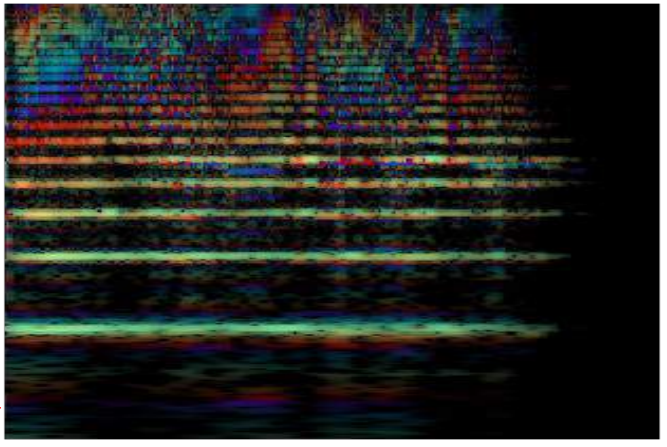
Original



SING



Nsynth



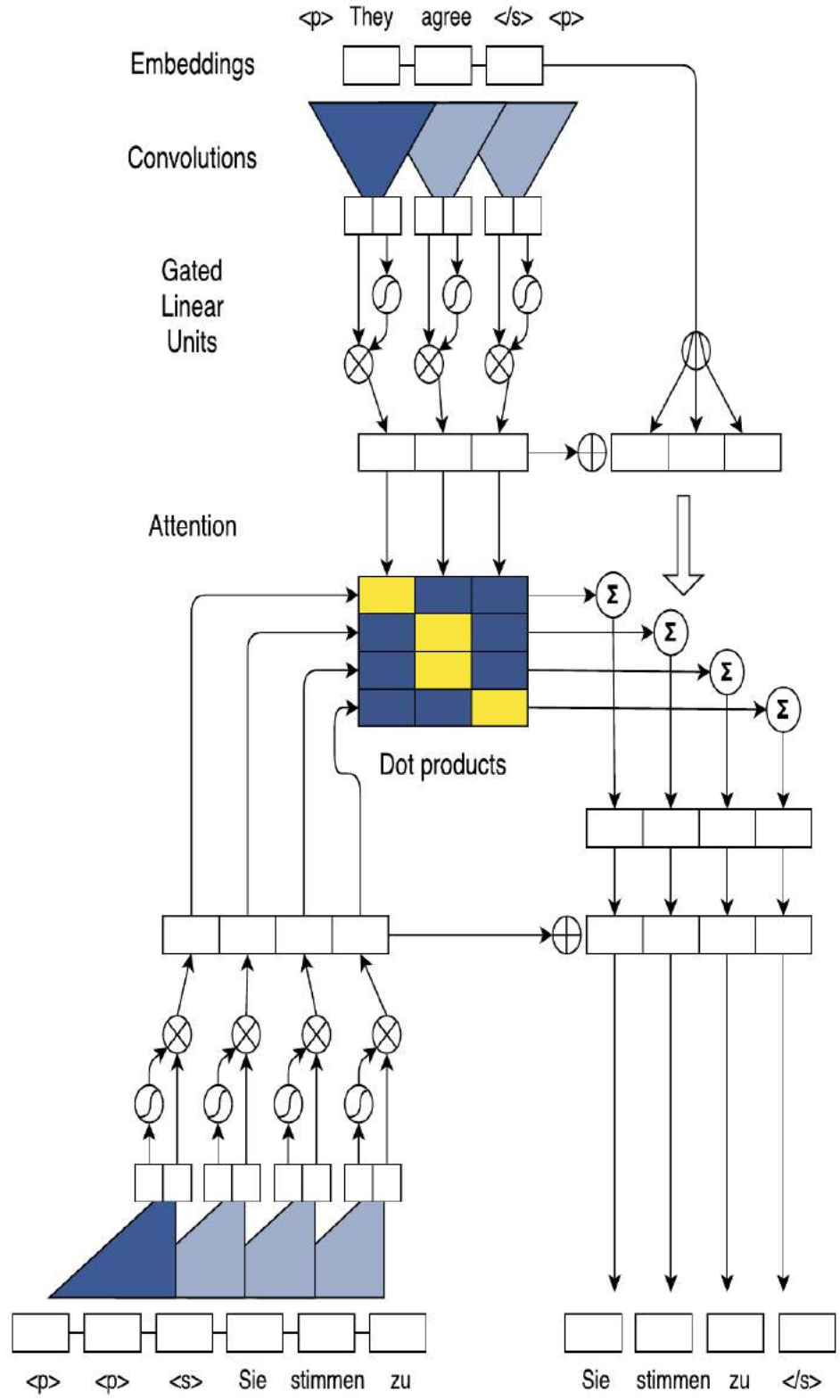
FairSeq for Translation

- [Gehring et al. ArXiv:1705.03122]

WMT'16 English-Romanian	BLEU
Sennrich et al. (2016b) GRU (BPE 90K)	28.1
ConvS2S (Word 80K)	29.45
ConvS2S (BPE 40K)	29.88

WMT'14 English-German	BLEU
Luong et al. (2015) LSTM (Word 50K)	20.9
Kalchbrenner et al. (2016) ByteNet (Char)	23.75
Wu et al. (2016) GNMT (Word 80K)	23.12
Wu et al. (2016) GNMT (Word pieces)	24.61
ConvS2S (BPE 40K)	25.16

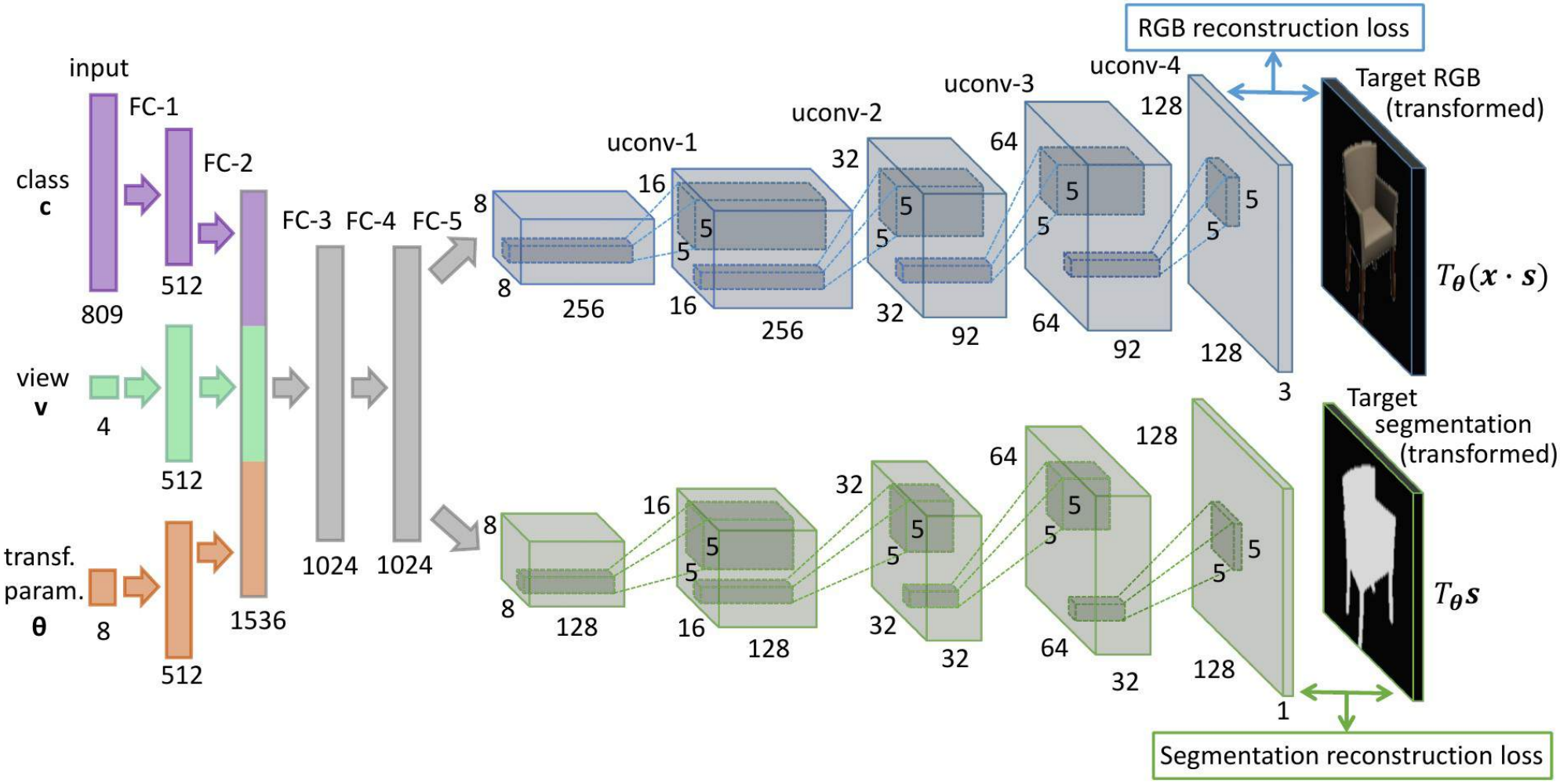
WMT'14 English-French	BLEU
Wu et al. (2016) GNMT (Word 80K)	37.90
Wu et al. (2016) GNMT (Word pieces)	38.95
Wu et al. (2016) GNMT (Word pieces) + RL	39.92
ConvS2S (BPE 40K)	40.46



Supervised ConvNets that Draw Pictures

Y LeCun

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

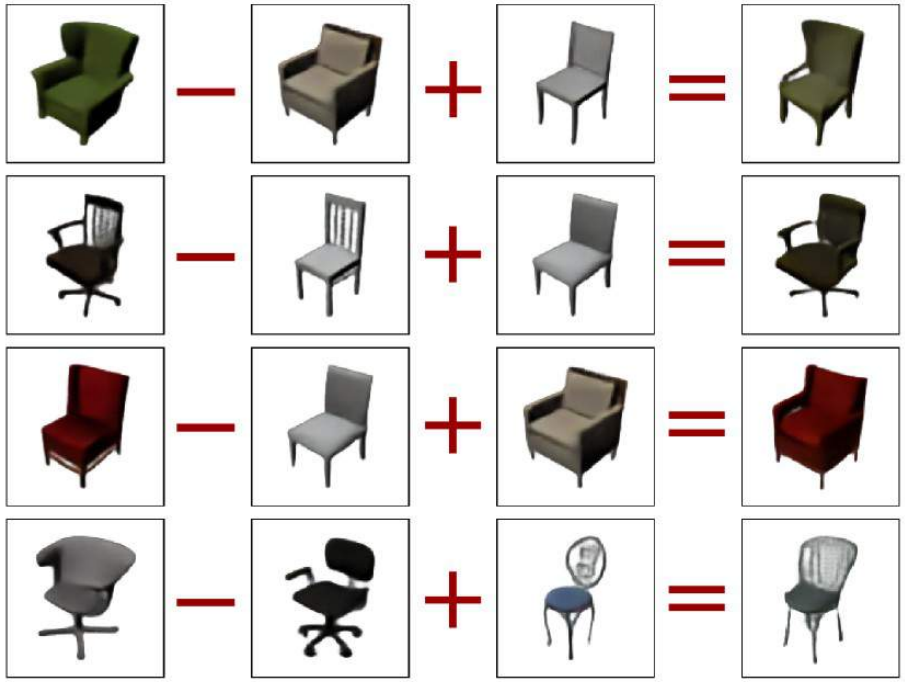


Supervised ConvNets that Draw Pictures

Y LeCun

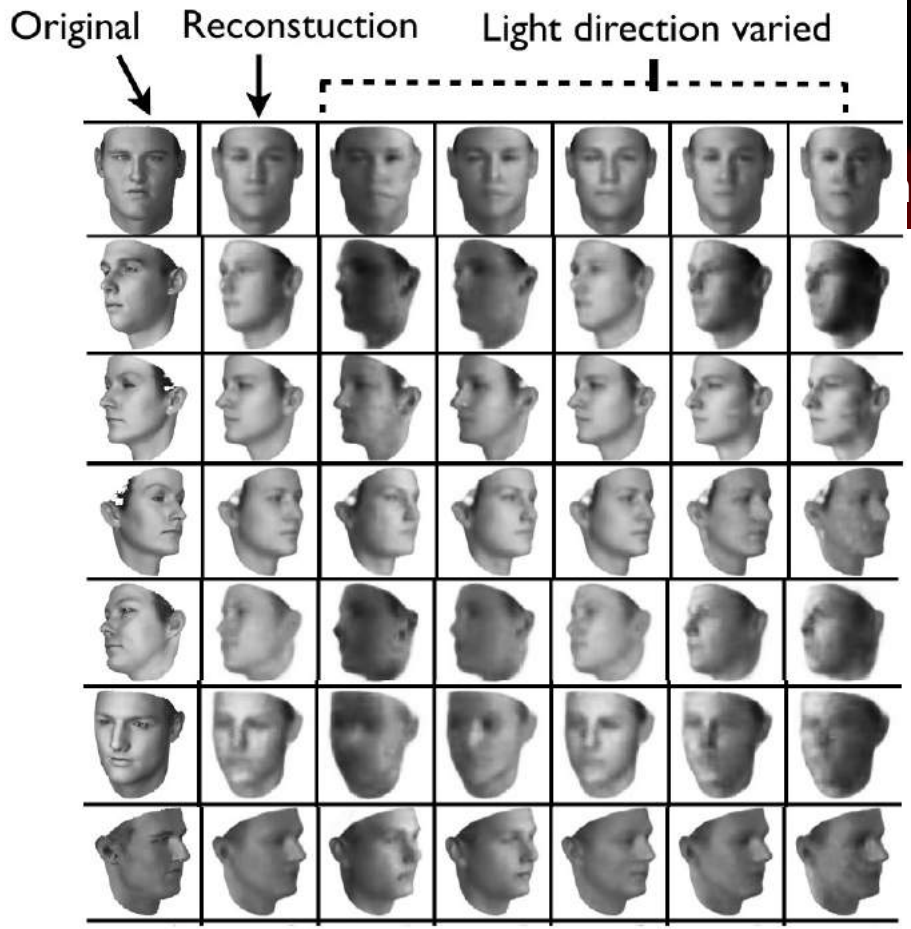
Generating Chairs

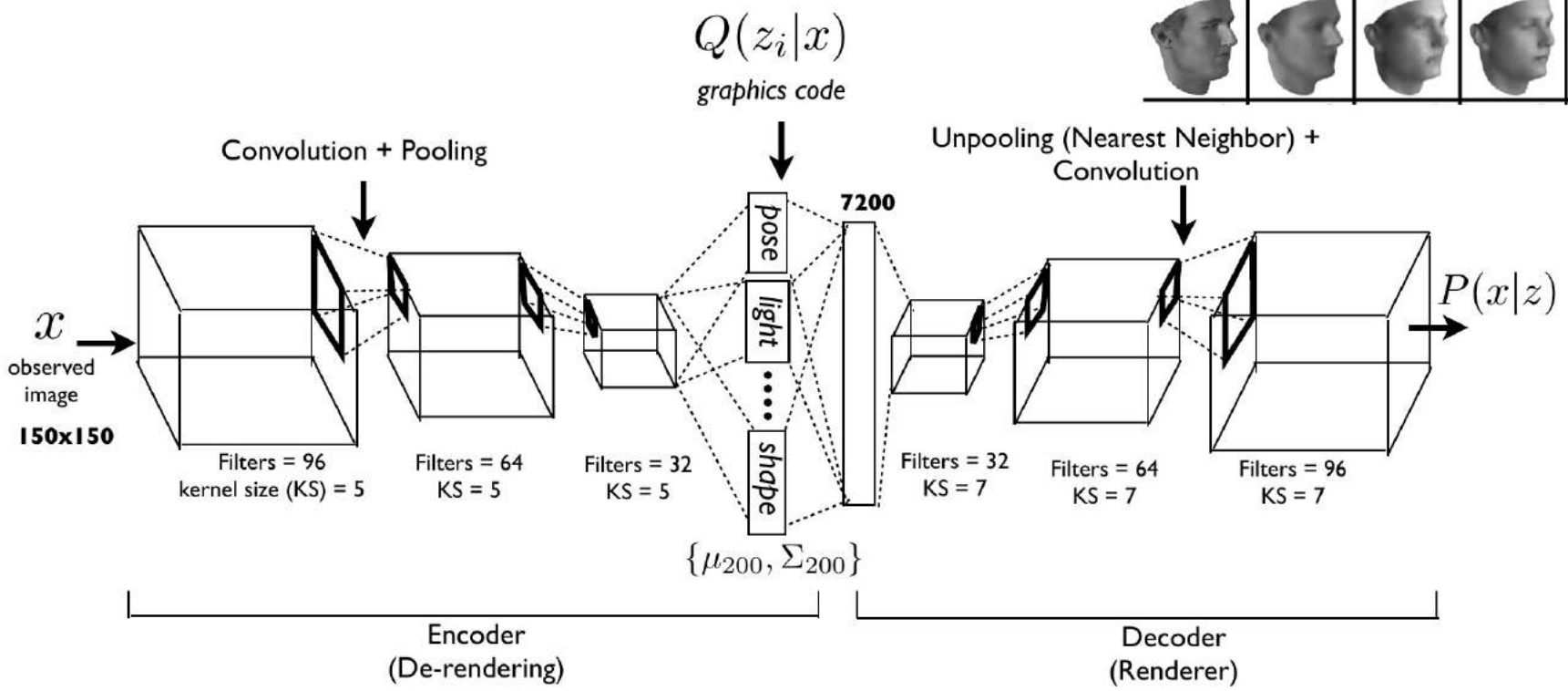
Chair Arithmetic in Feature Spa



Convolutional Encoder-Decoder

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





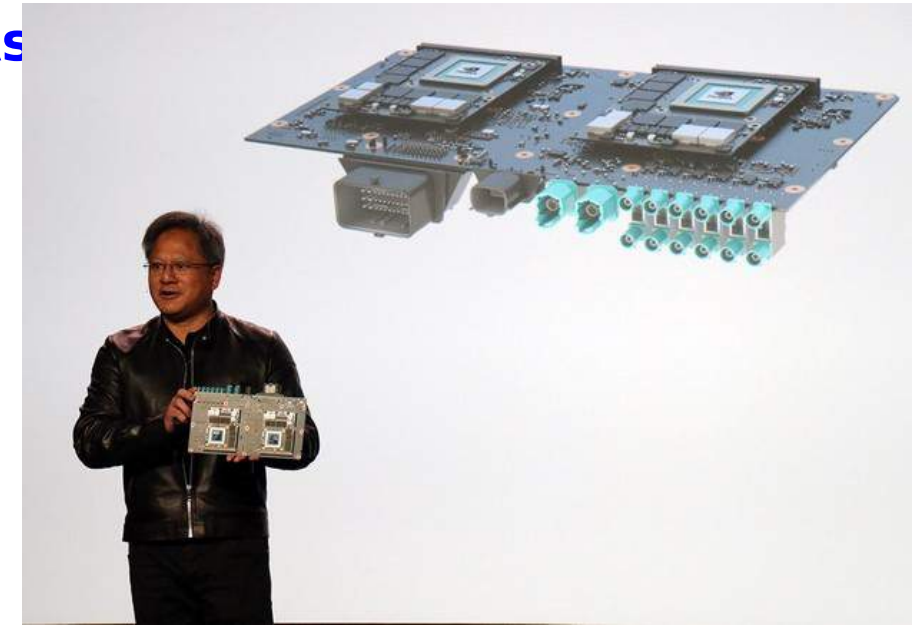

ConvNets are Everywhere
(or soon will be)

NVIDIA: ConvNet-Based Driver Assistance

Drive-PX2: Open Platform for Driver As

Embedded Super-Computer: 42 TOPS

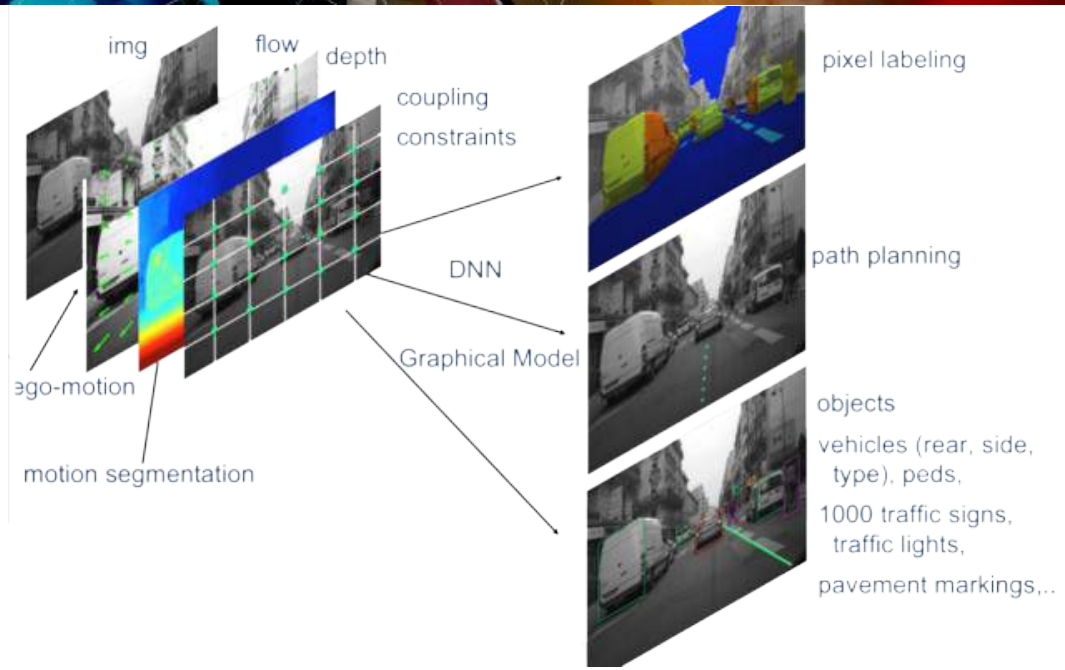
- (=150 Macbook Pros)

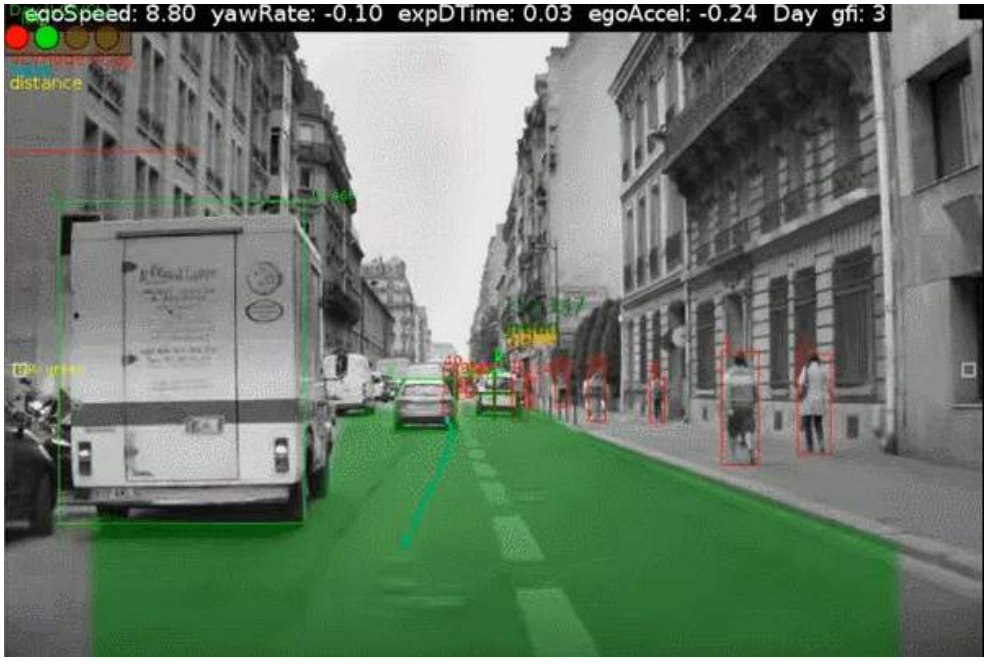


MobilEye: ConvNet-Based Driver Assistance

Deployed in the latest

Tesla Model S and Model X





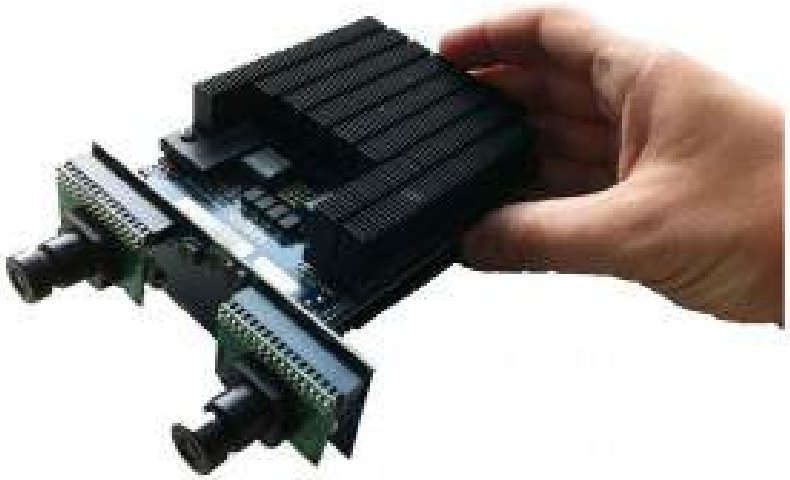
Deep Learning is Everywhere (ConvNets are Everywhere)

- **Lots of applications at Facebook, Google, Microsoft, Baidu, Twitter, IBM…**
  - Image recognition for photo collection search
  - Image/Video Content filtering: spam, nudity, violence.
  - Search, Newsfeed ranking
- **People upload 800 million photos on Facebook every day**
  - (2 billion photos per day if we count Instagram, Messenger and Whatsapp)
- **Each photo on Facebook goes through two ConvNets within 2 seconds**
  - One for image recognition/tagging
  - One for face recognition (not activated in Europe).
- **Soon ConvNets will really be everywhere:**
  - self-driving cars, medical imaging, augemnted reality, mobile devices, smart cameras, robots, toys…..


ConvNet
Hardware

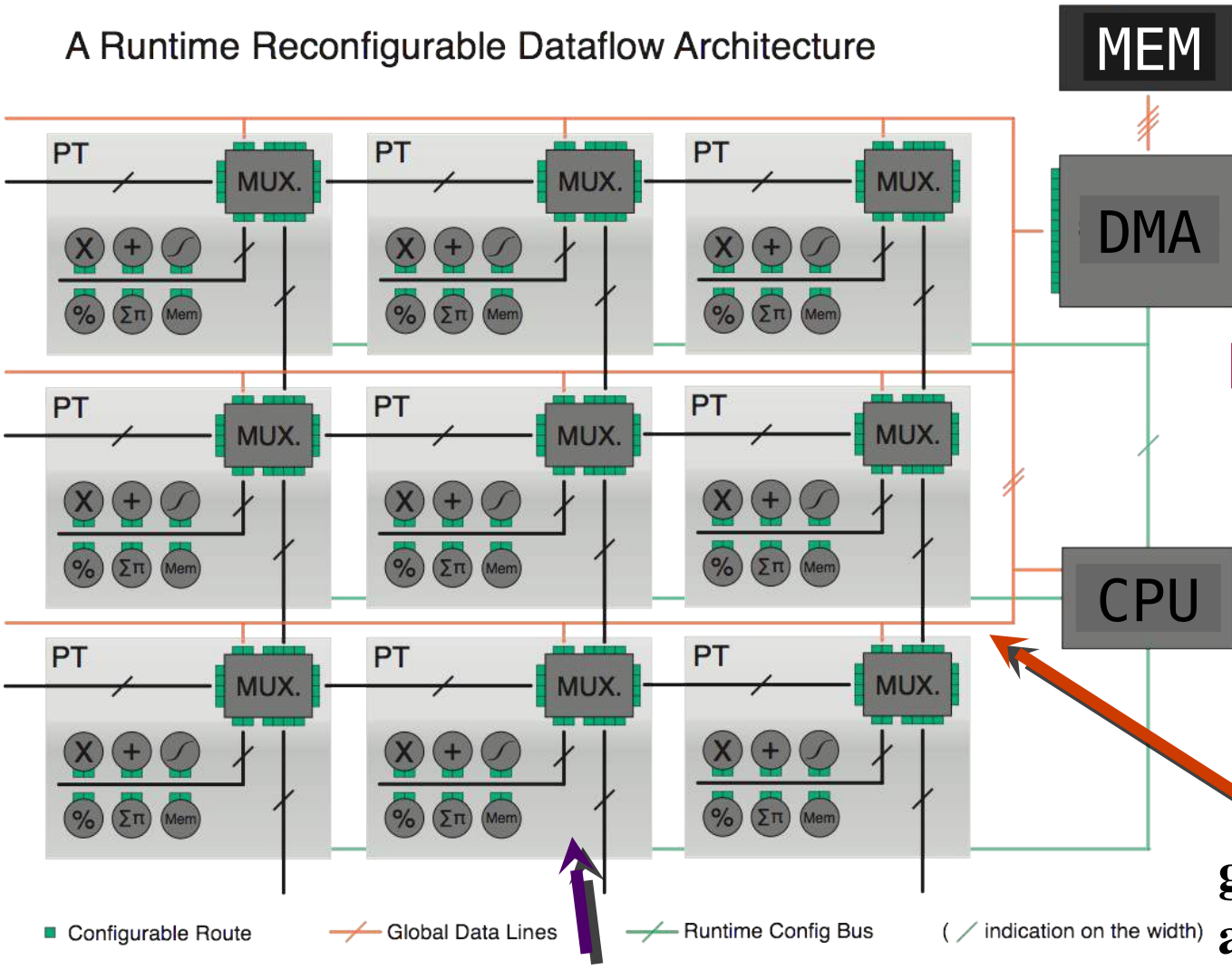
NeuFlow architecture (NYU + Purdue)

- **Collaboration NYU-Purdue: Eugenio Culurciello's e-Lab.**
- **Running on Picocomputing 8x10cm high-performance FPGA board**
  - Virtex 6 LX240T: 680 MAC units, 20 neuflow tiles
- **Full scene labeling at 20 frames/sec (50ms/frame) at 320x240**



board with Virtex-6

NewFlow: Architecture



MEM

DMA

Multi-port memory controller (DMA)

[x12 on a V6 LX240T]

CPU

RISC CPU, to reconfigure tiles and data paths, at runtime

global network-on-chip to allow fast reconfiguration

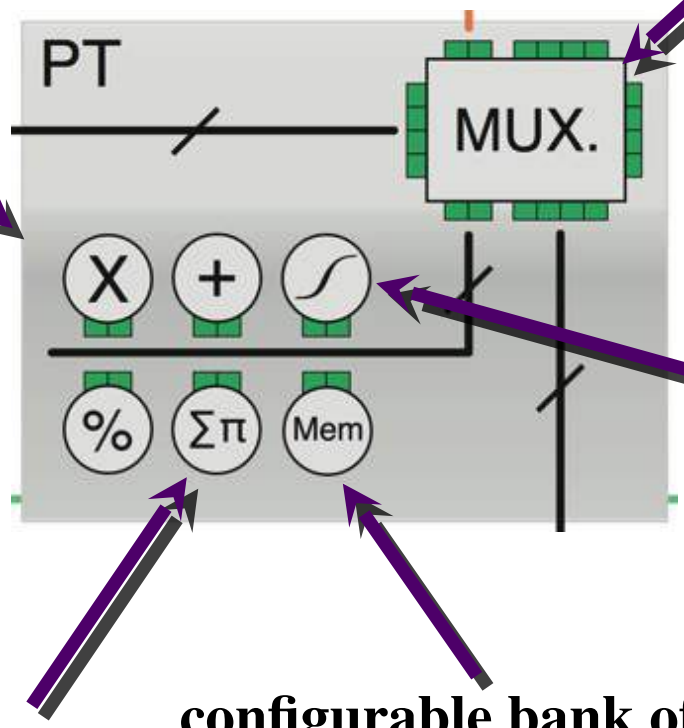
grid of passive processing tiles (PTs)

[x20 on a Virtex6 LX240T]

NewFlow: Processing Tile Architecture

Term-by:-term streaming operators (MUL,DIV,ADD,SUB,MAX)

[x8,2 per tile]



configurable router, to stream data in and out of the tile, to neighbors or DMA ports

[x20]

configurable piece-wise linear or quadratic mapper

[x4]

full 1/2D parallel convolver with 100 MAC units

[x4]

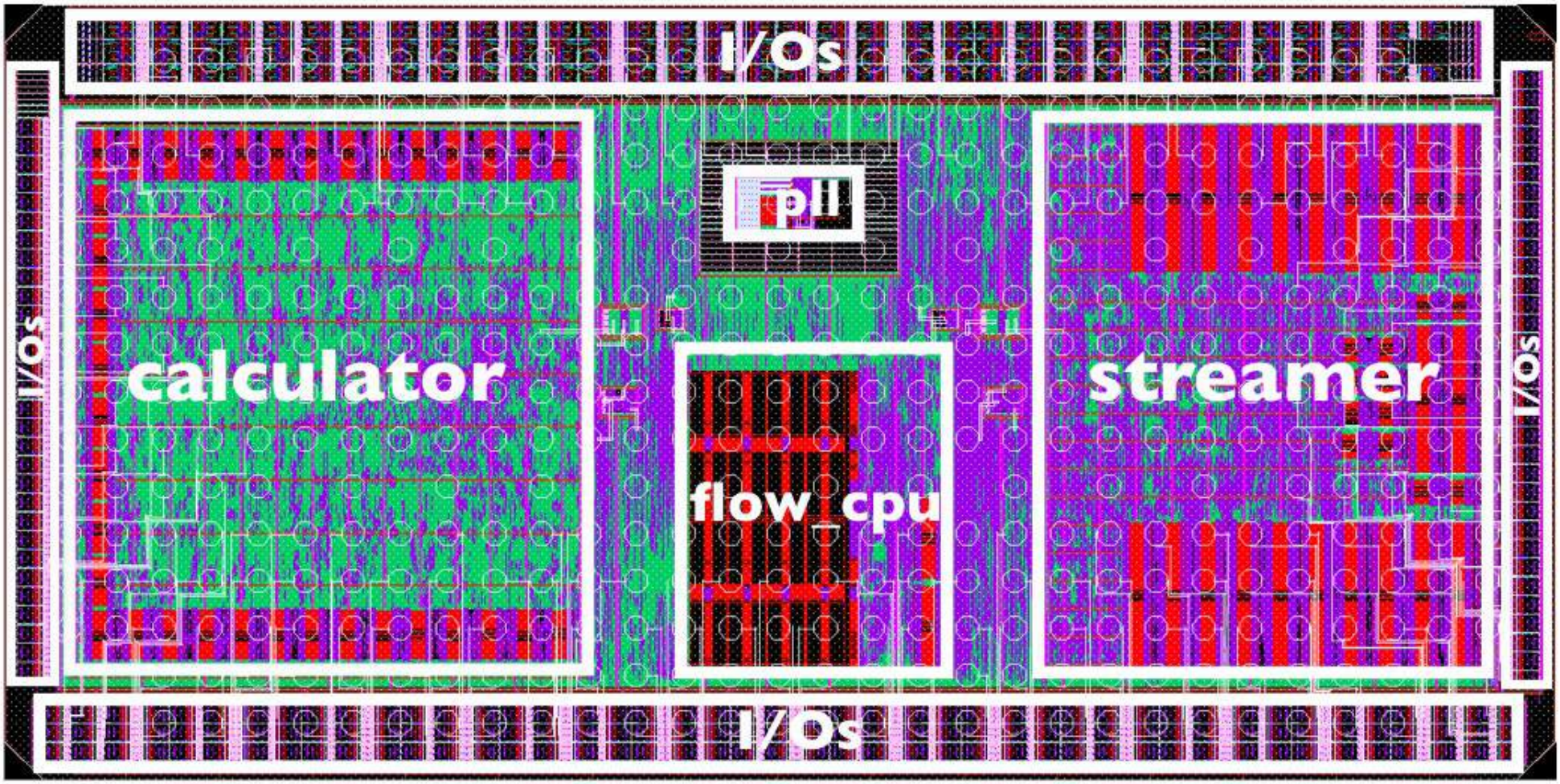
configurable bank of FIFOs , for stream buffering, up to 10kB per PT

[x8]

[Virtex6 LX240T]

NewFlow ASIC: 2.5x5 mm, 45nm, 0.6Watts, >300GOPS

- Collaboration Purdue-NYU: Eugenio Culurciello's e-Lab
- Suitable for vision-enabled embedded and mobile devices
- (but the fabrication was botched...)



[Pham, Jelaca, Farabet, Martini, LeCun, Culurciello 2012]

ConvNet Hardware Today

- **Training is all done on GPUs (on NVIDIA GPUs)**
  - 4 or 8 GPU cards per node. Many nodes per rack. Many racks.
  - O(100 Tflops) per card
  - Training needs performance, programmability, flexibility, accuracy.
  - Power consumption and space is not that important.
- **Many large hardware companies are developing ConvNet accelerators**
  - NVIDIA: evolving from GPU. Embedded applications (automotive)
  - Intel: high-end: Nervana; mid-range: CPU+NN accelerator; automotive: MobilEye; low-end: Movidius low-power parallel DSP.
  - Orcam: low-power ConvNet chip for the visually impaired
  - Qualcomm, Samsung, ARM: ConvNet IP for mobile devices
  - Cadence, Synopsis: ConvNet IP for custom chips
- **Lots of startups are getting into the field**
  - Many in China, Taiwan, Korea, Japan, USA, Europe.
- **Low-power chips today: 10 TOPS/Watt**

Deep Learning Applications Today = ConvNets

- **Practical/deployed applications of deep learning:**
  - Almost exclusively ConvNet trained with Backprop
  - A few recurrent nets used experimentally
  - Applications: image recognition, face recognition, video understanding, speech recognition, natural language understanding.
- **Training**
  - Backprop running on racks of GPUs
- **Production**
  - ConvNets on CPUs, with reduced-precision arithmetics
  - Some are moving to GPUs or FPGA.
- **Scale**
  - People upload several billion images on Facebook every day
  - All of them go through a handful of ConvNets within 2 seconds
    - Generic image tagging, face recognition, caption for the visually impaired, objectionable content detection (nudity, violence), text detection/OCR…..