What are deep vision models learning?

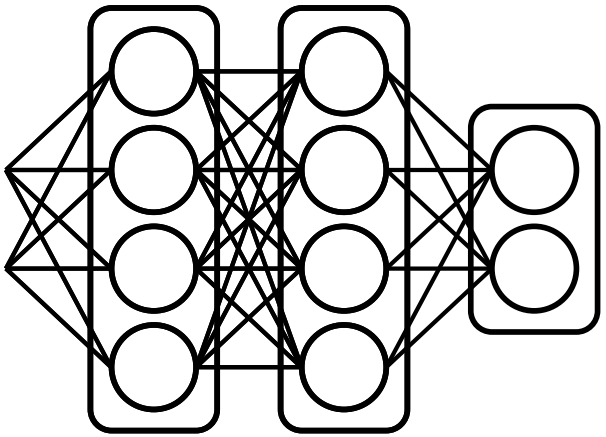
David Crandall

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch
Joint work with Alexei A. Efros & Abhinav Gupta

ICCV 2015

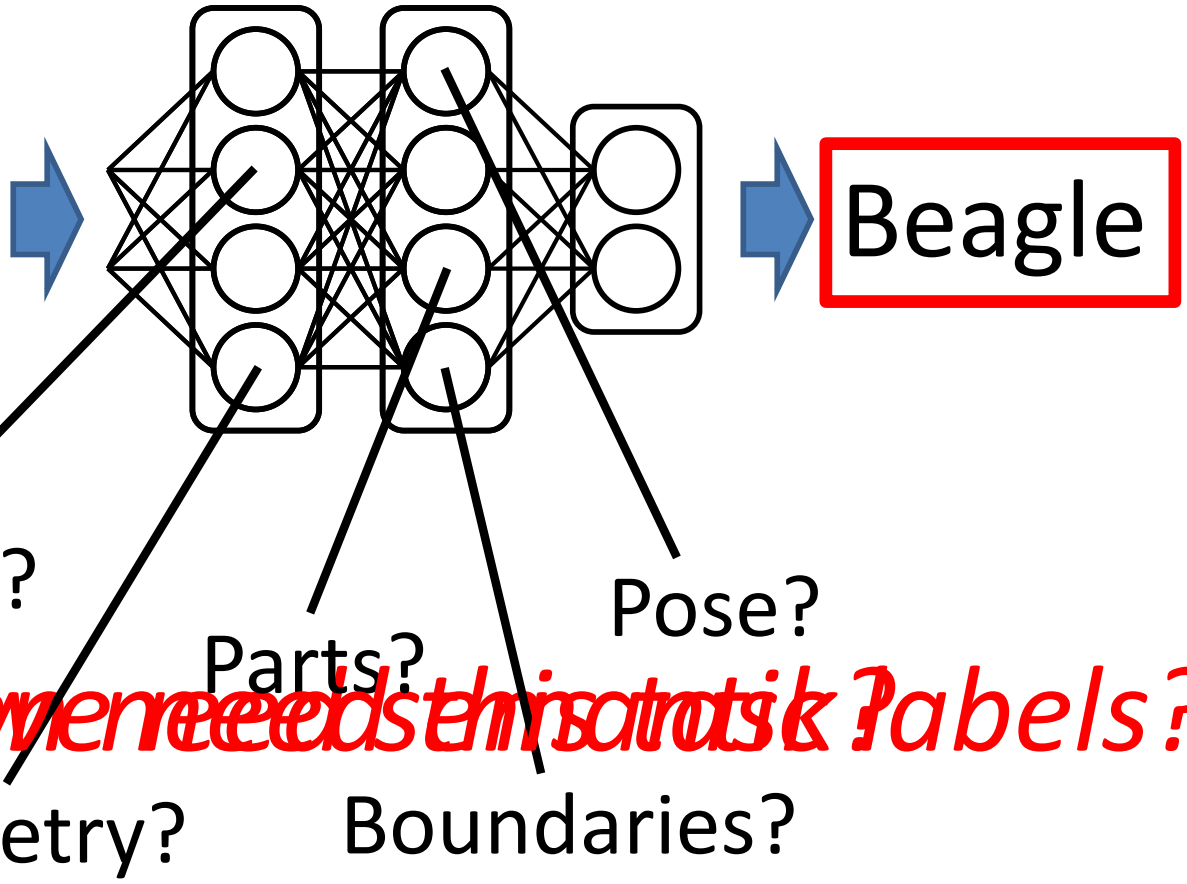
ImageNet + Deep Learning



Beagle

- Image Retrieval
- Detection (RCNN)
- Segmentation (FCN)
- Depth Estimation
- …

ImageNet + Deep Learning



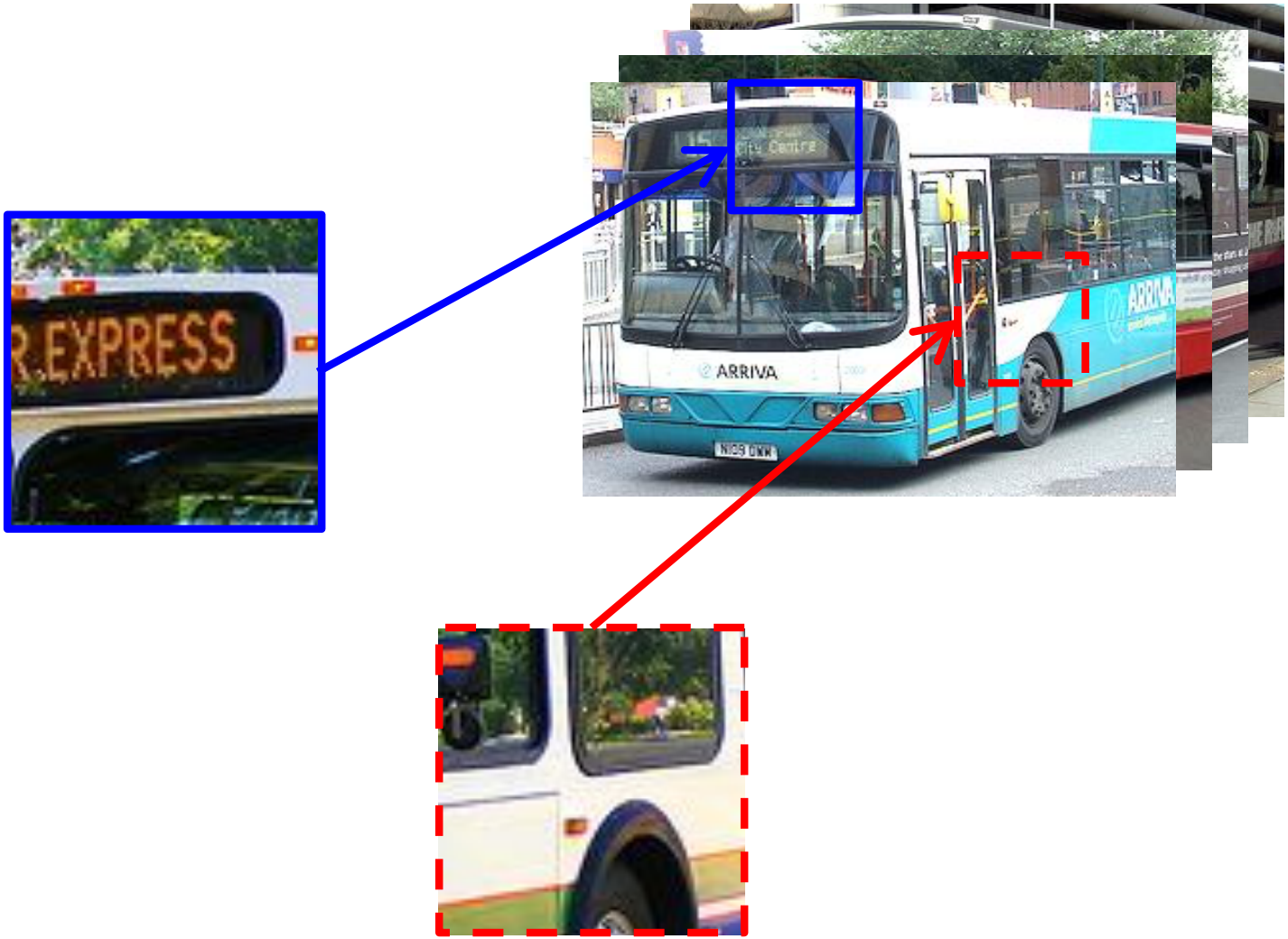
Do we even need this task?
Do we even need labels?

Context Prediction for Images

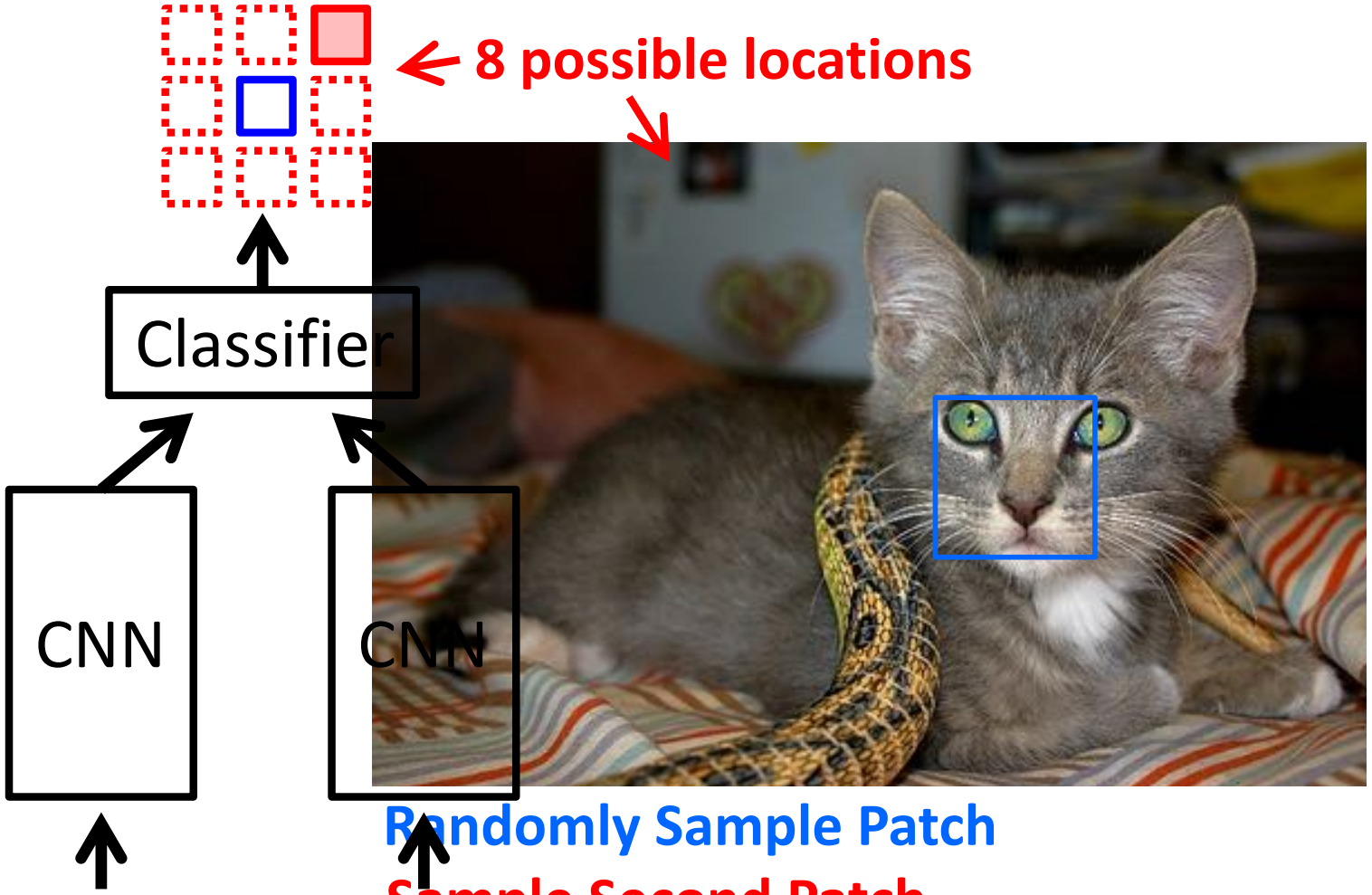
A

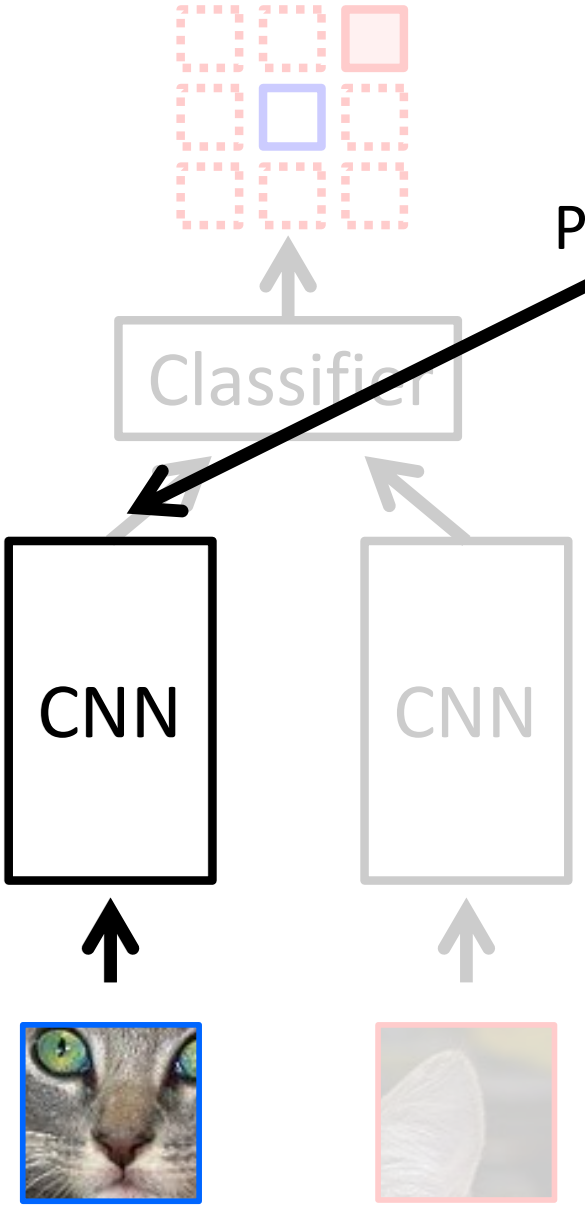
B

Semantics from a non-semantic task



Relative Position Task



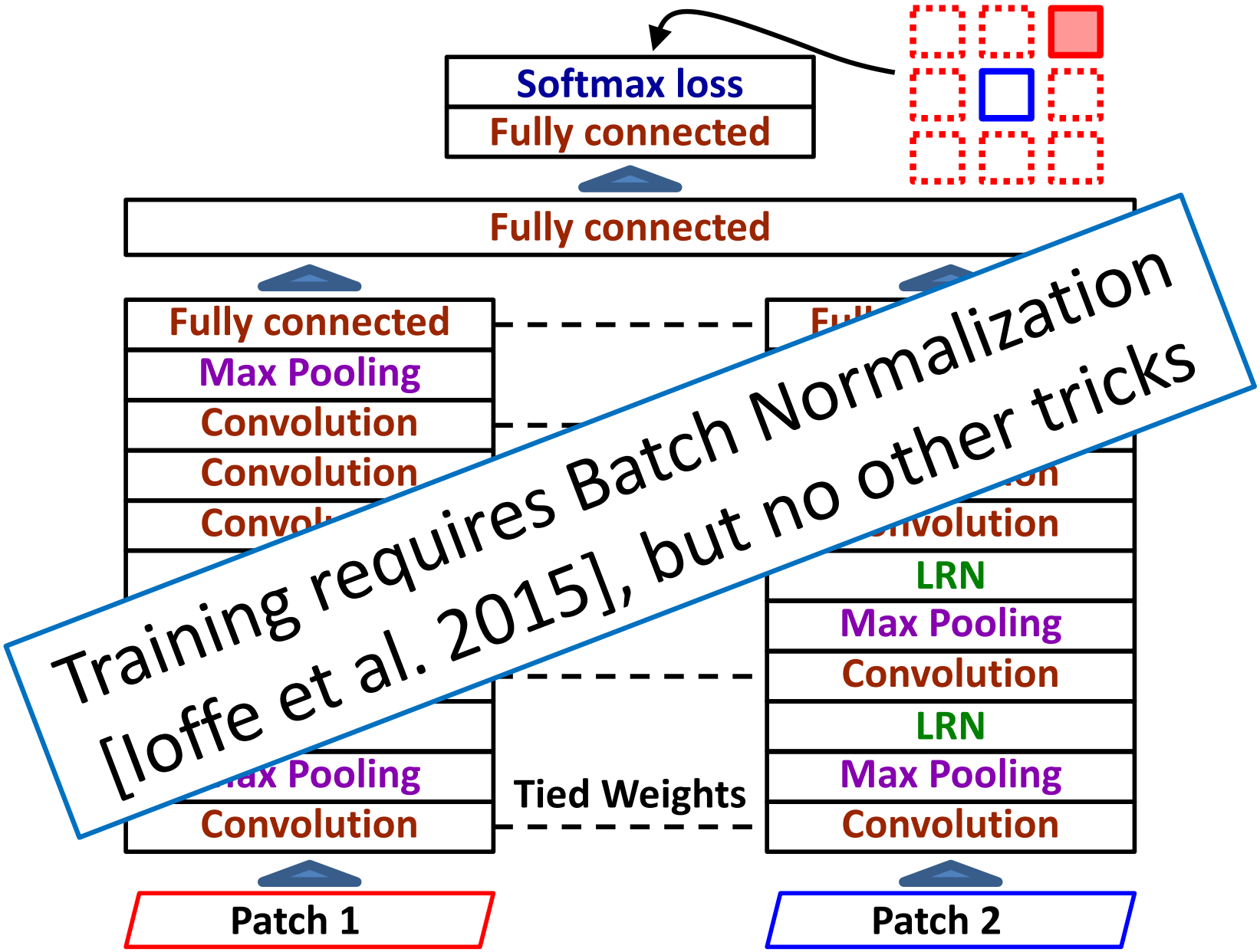


Input

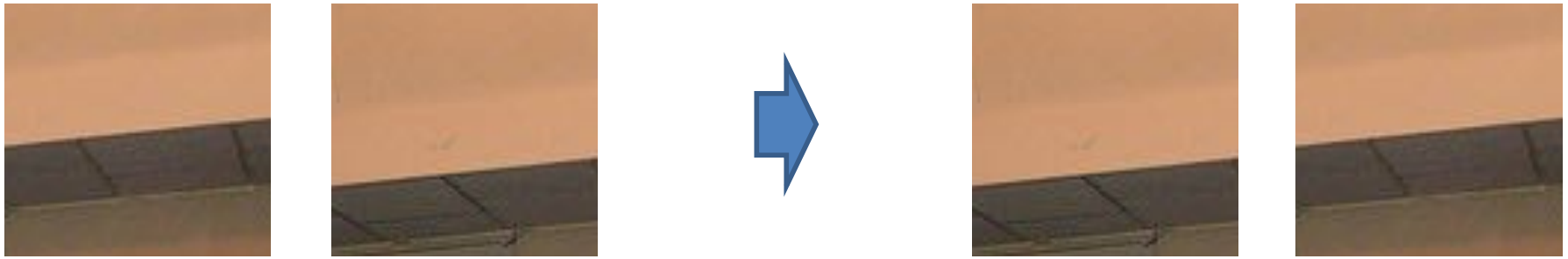
Nearest Neighbors

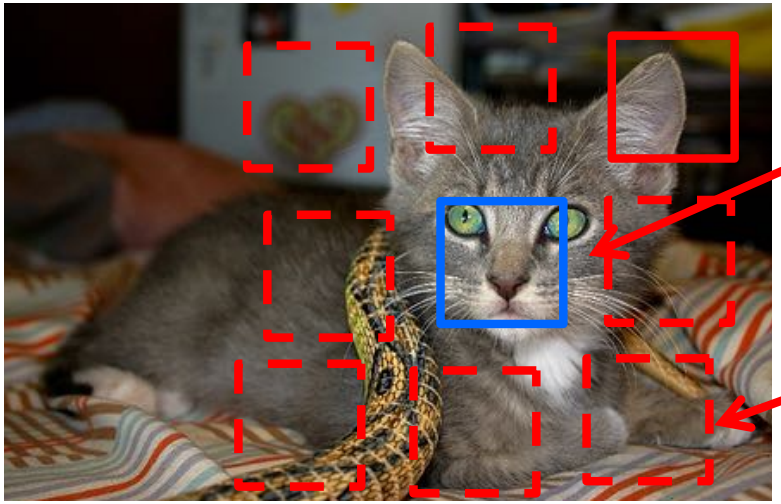
Note: connects ***across*** instances!

Architecture

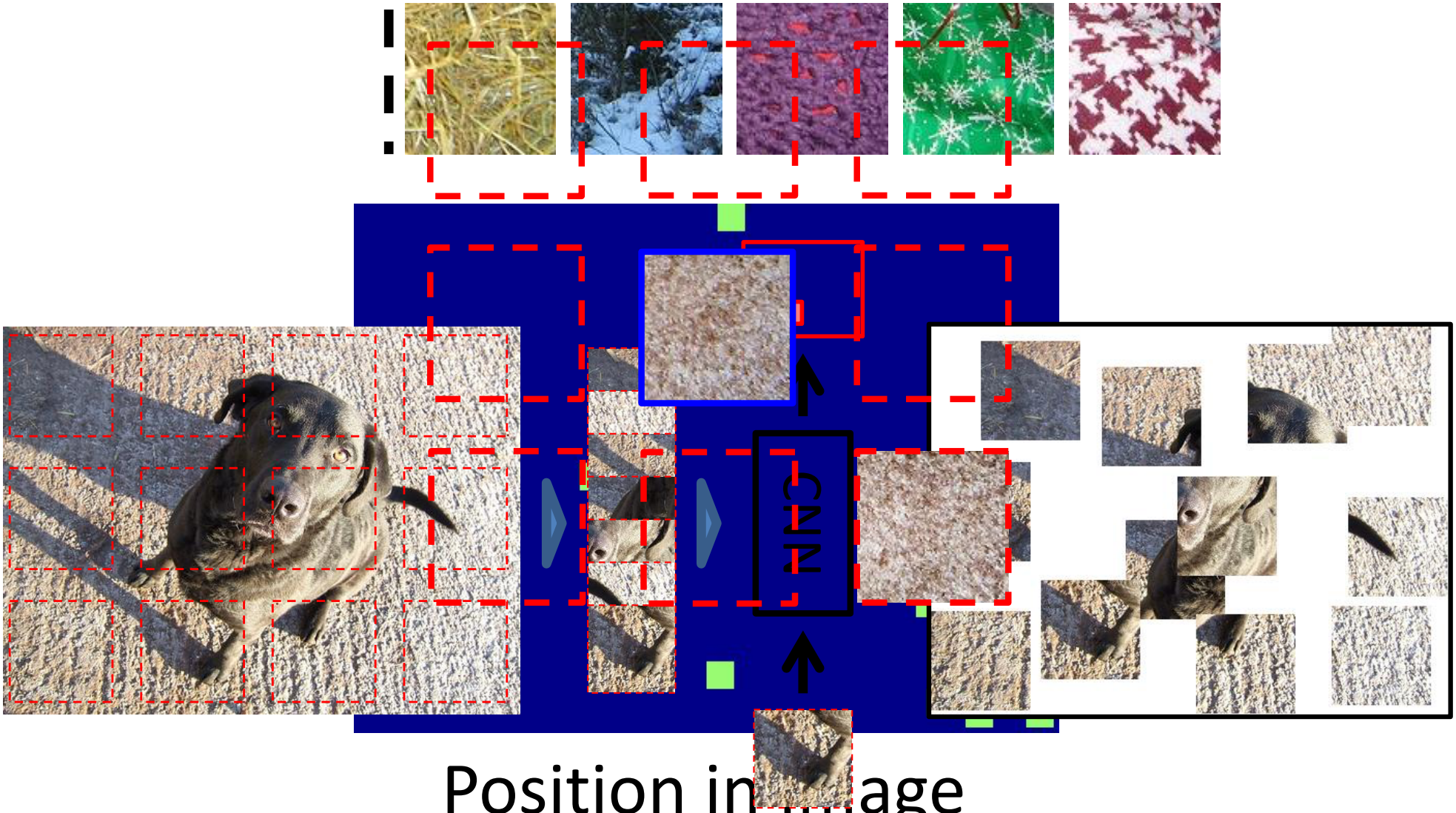


Avoiding Trivial Shortcuts

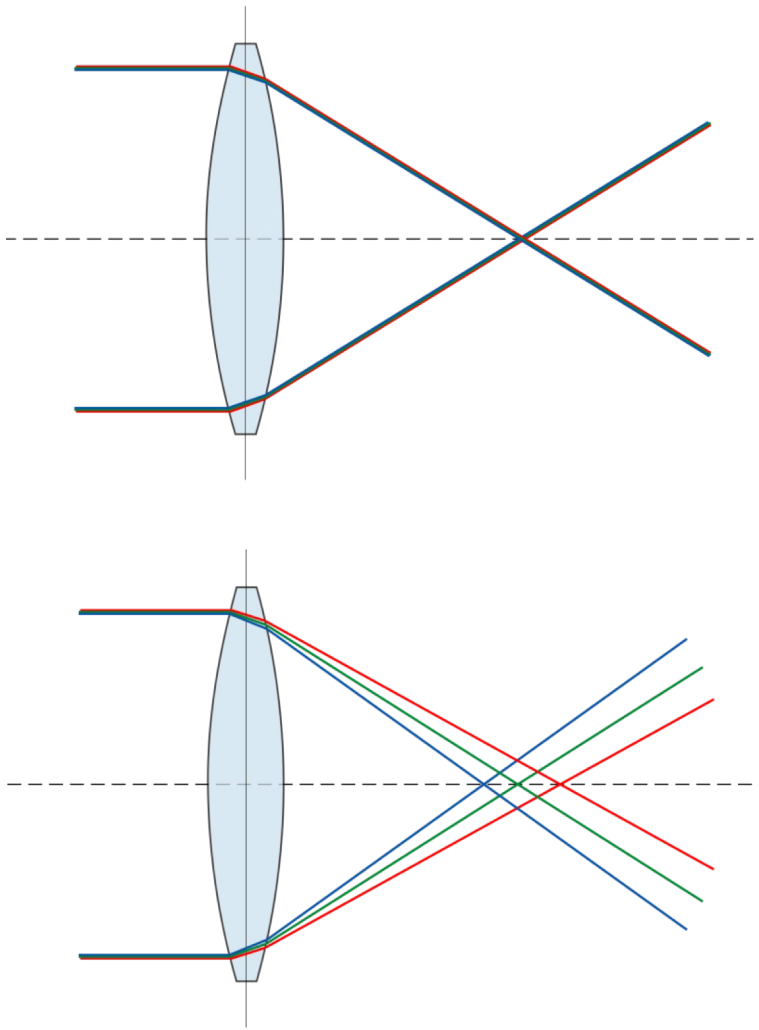


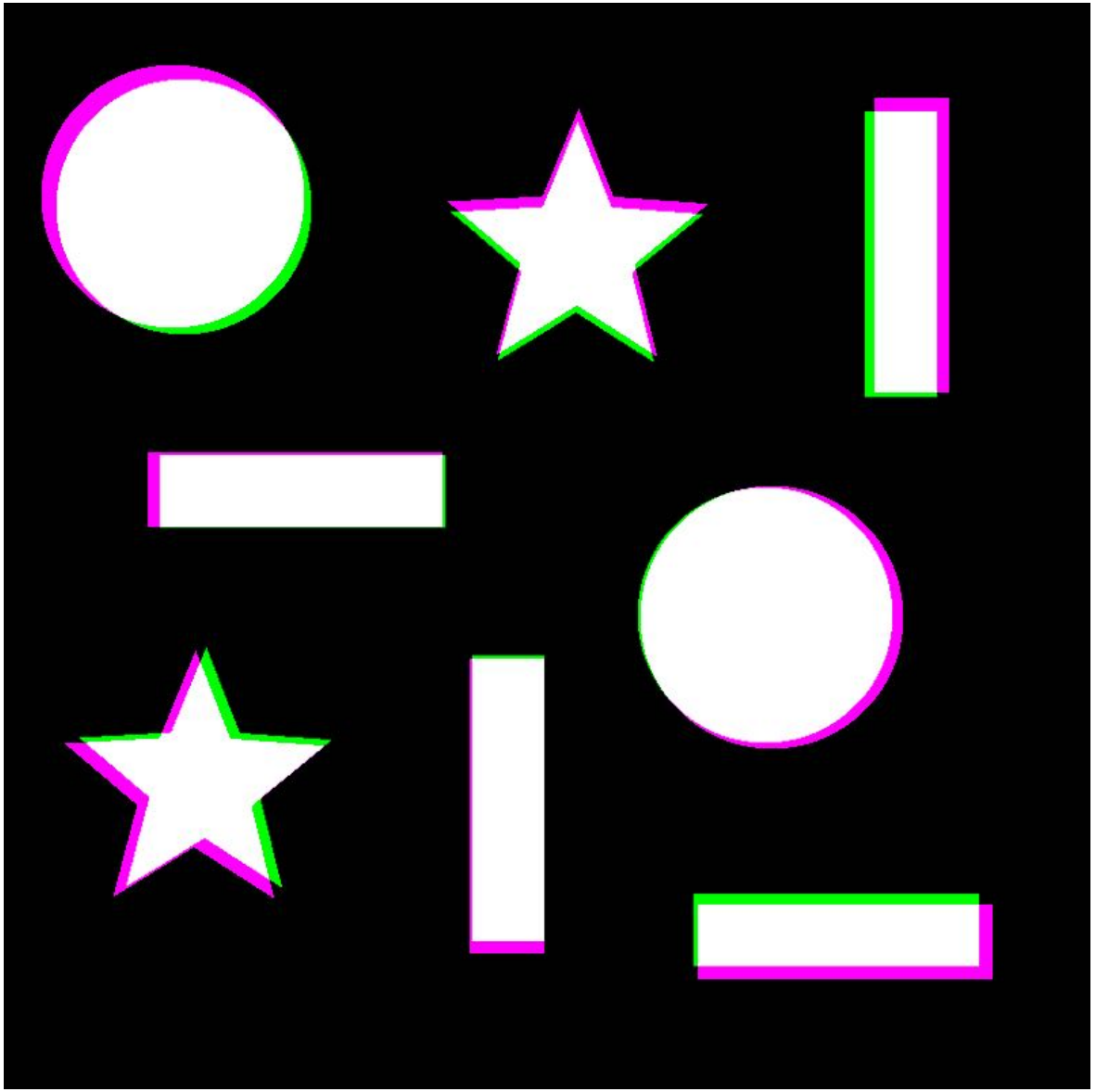


A Not-So “Trivial” Shortcut

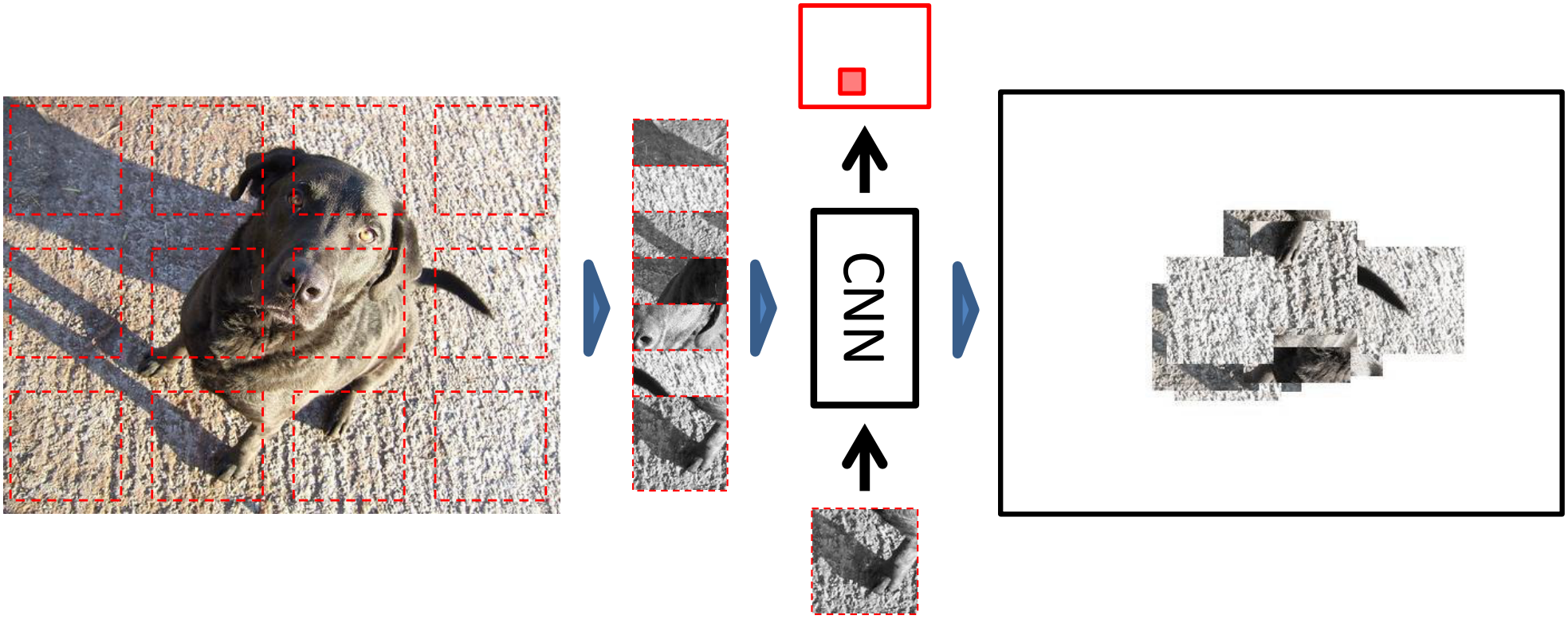


Chromatic Aberration

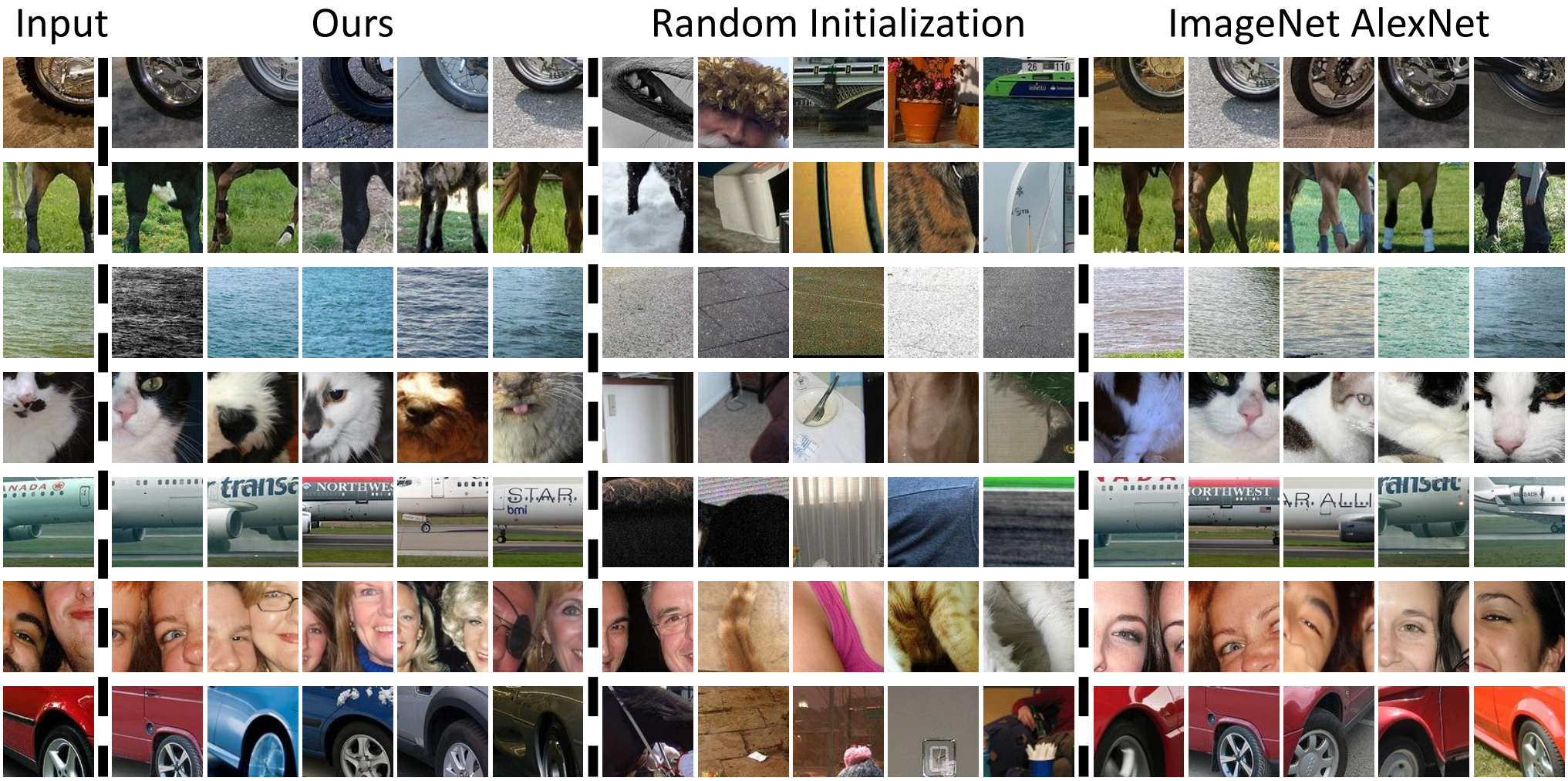




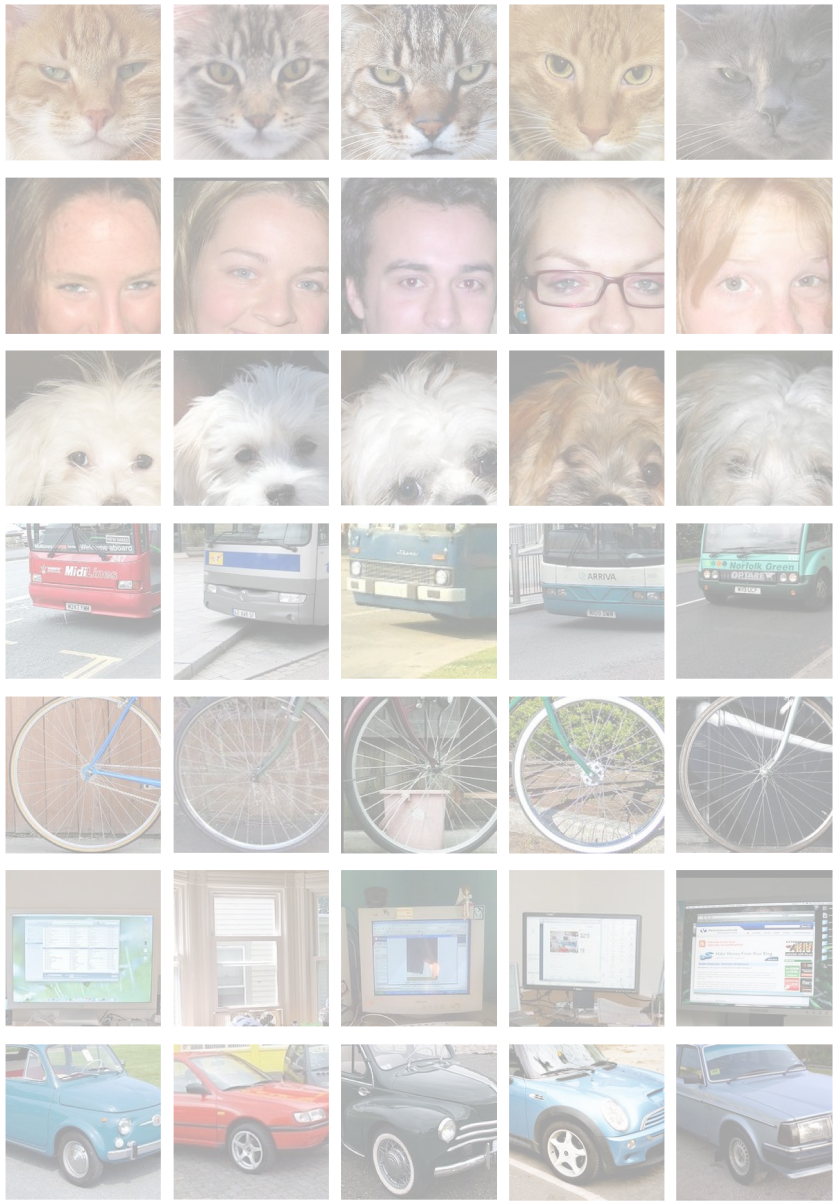
Chromatic Aberration

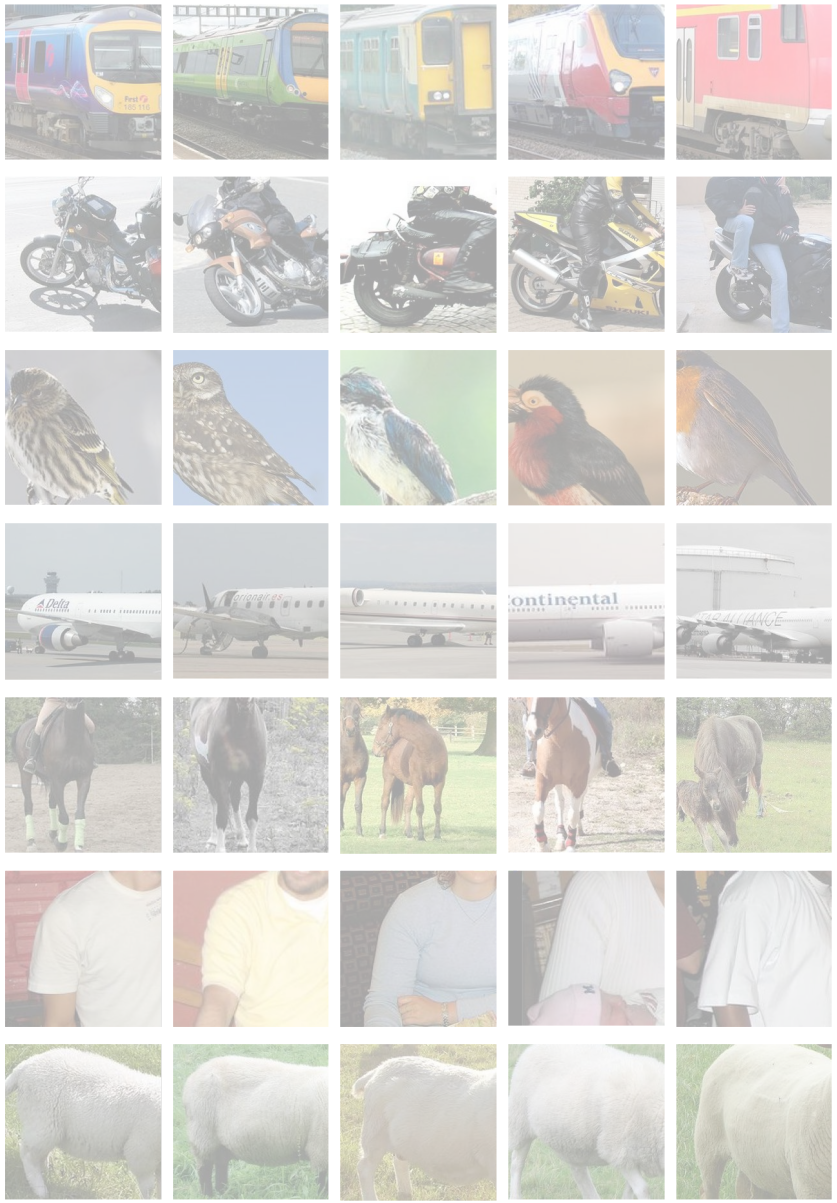


What is learned?

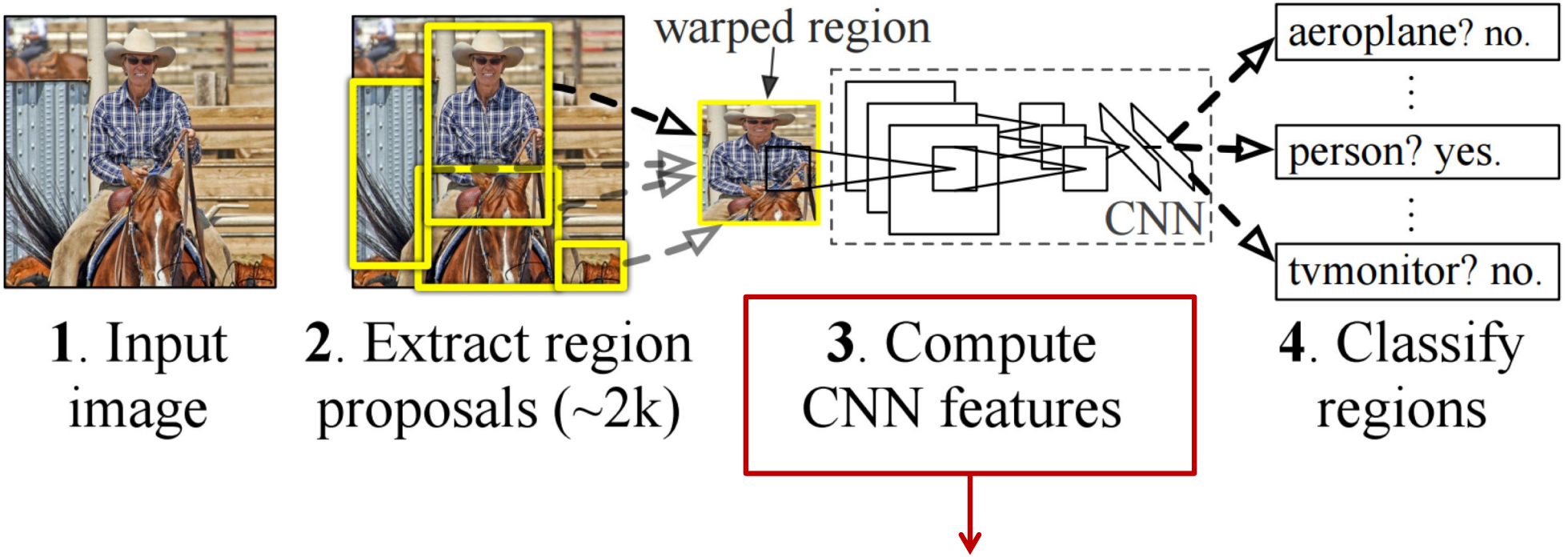


Mined from Pascal VOC2011



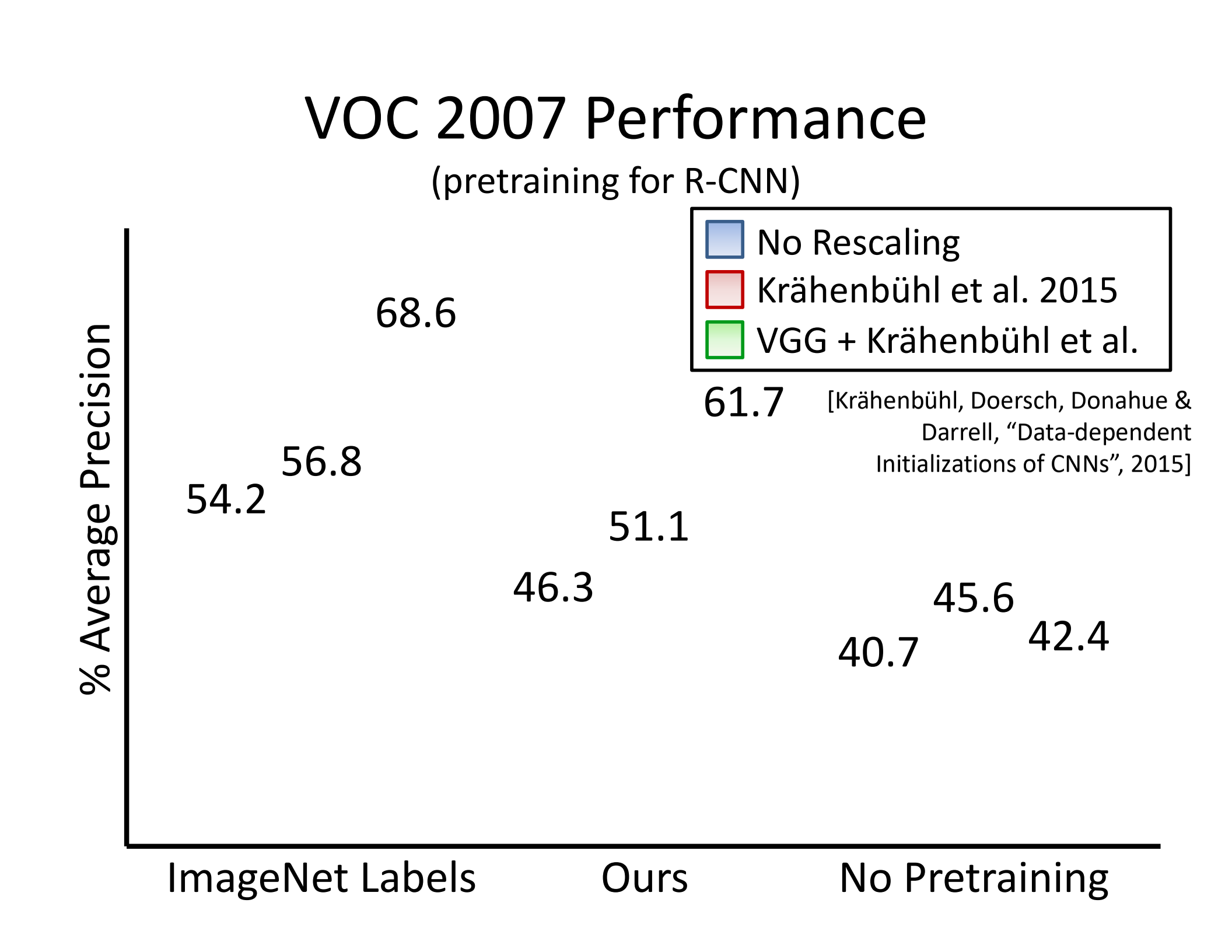


Pre-Training for R-CNN

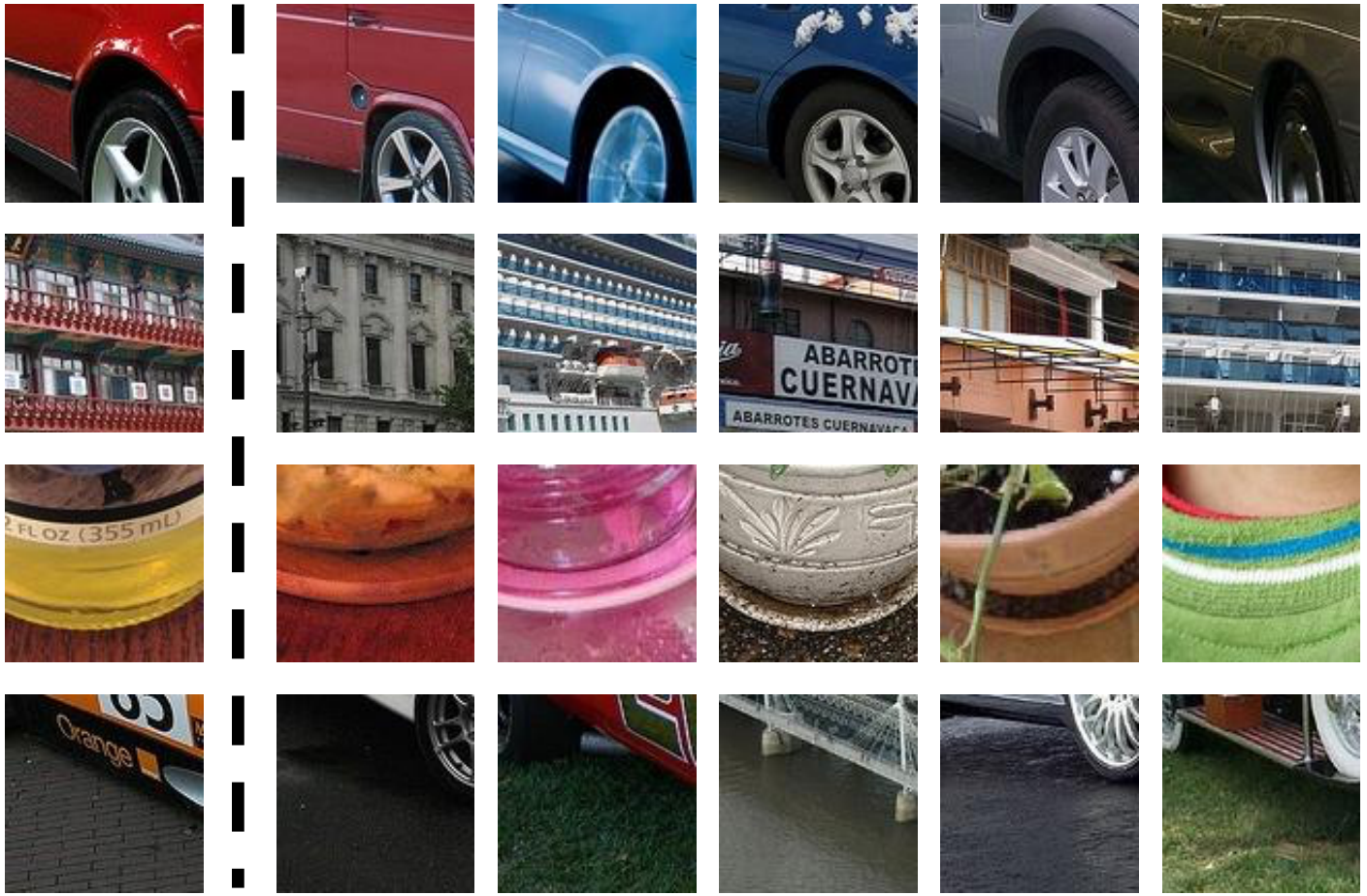


Pre-train on relative-position task, w/o labels

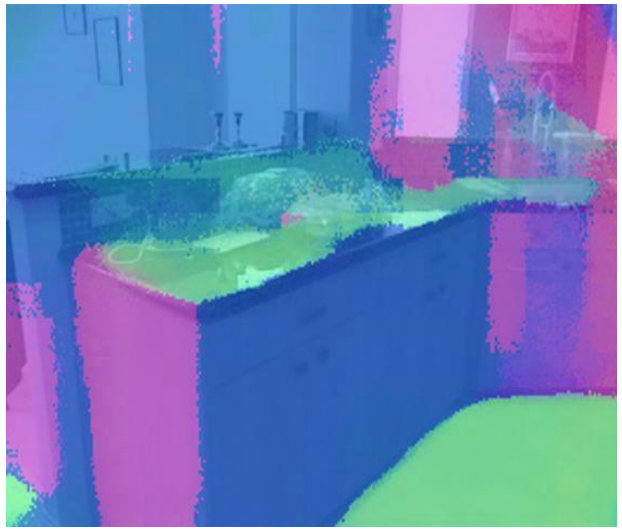
[Girshick et al. 2014]


VOC 2007 Performance
(pretraining for R-CNN)
No Rescaling
Krähenbühl et al. 2015
VGG + Krähenbühl et al.
% Average Precision
68.6
61.7
56.8
54.2
51.1
46.3
45.6
42.4
40.7
ImageNet Labels
Ours
No Pretraining
[Krähenbühl, Doersch, Donahue & Darrell, "Data-dependent Initializations of CNNs", 2015]

Capturing Geometry?



Surface-normal Estimation

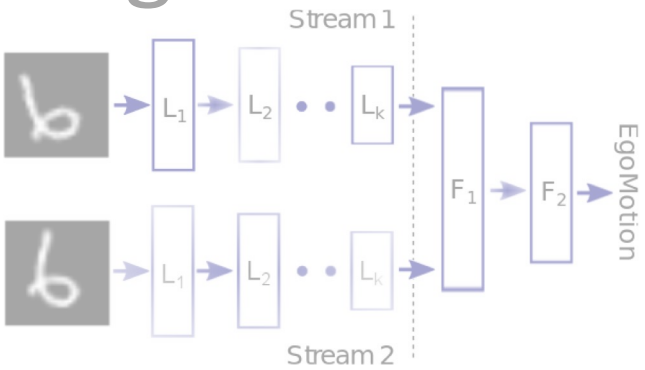


Method	Error (Lower Better)		% Good Pixels (Higher Better)		
	Mean	Median	11.25°	22.5°	30.0°
No Pretraining	38.6	26.5	33.1	46.8	52.5
Ours	**33.2**	21.3	36.0	51.2	57.8
ImageNet Labels	33.3	**20.8**	**36.7**	**51.7**	**58.1**

So, do we need semantic labels?

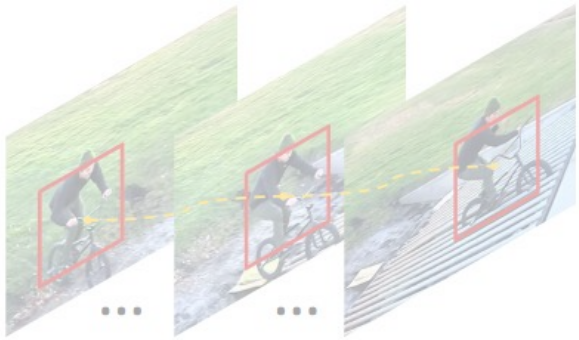
“Self-Supervision” and the Future

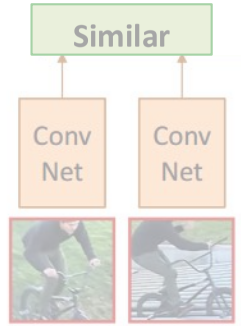
Ego-Motion



[Agrawal et al. 2015; Jayaraman et al. 2015]

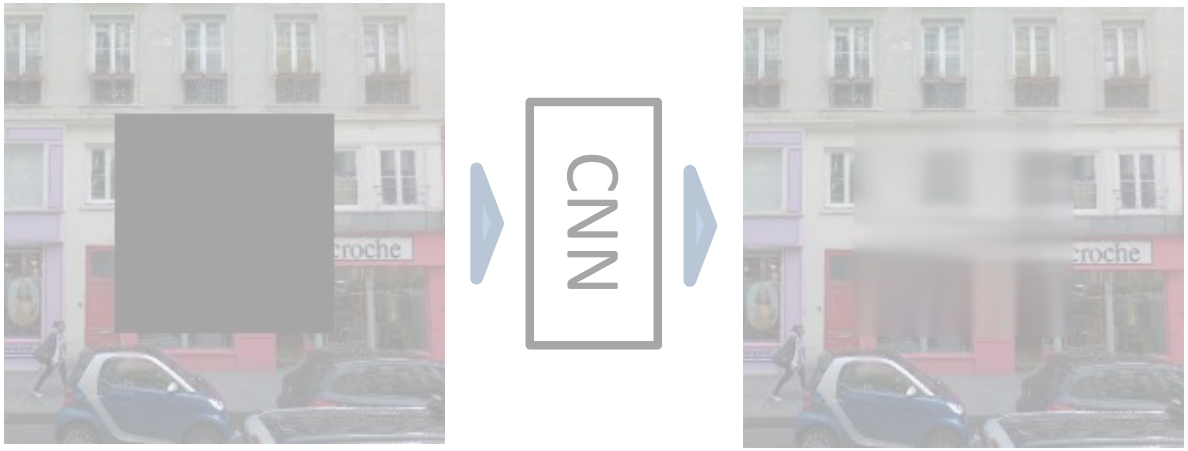
Video





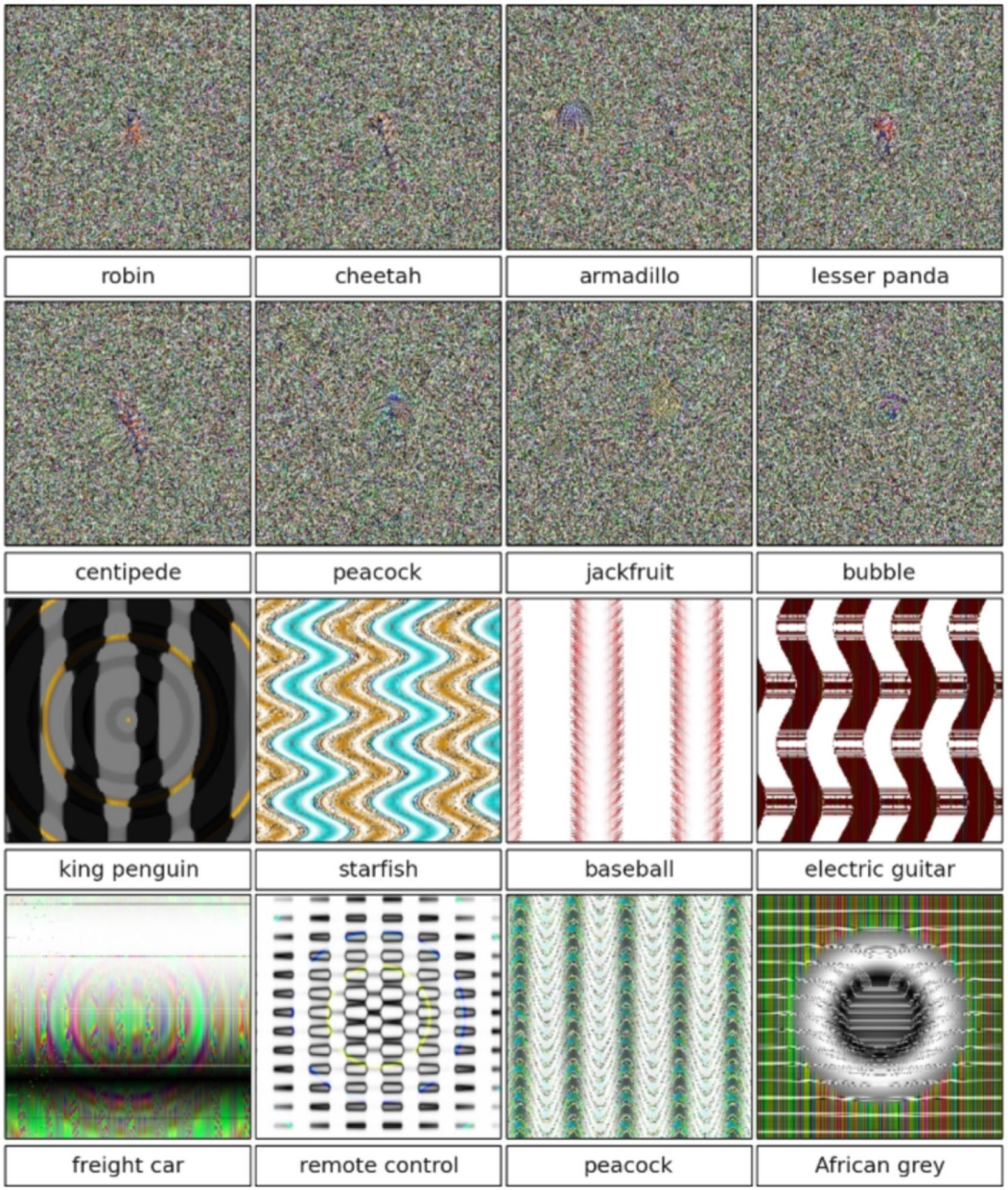
[Wang et al. 2015; Srivastava et al 2015; …]

Context



[Doersch et al. 2014; Pathak et al. 2015; Isola et al. 2015]

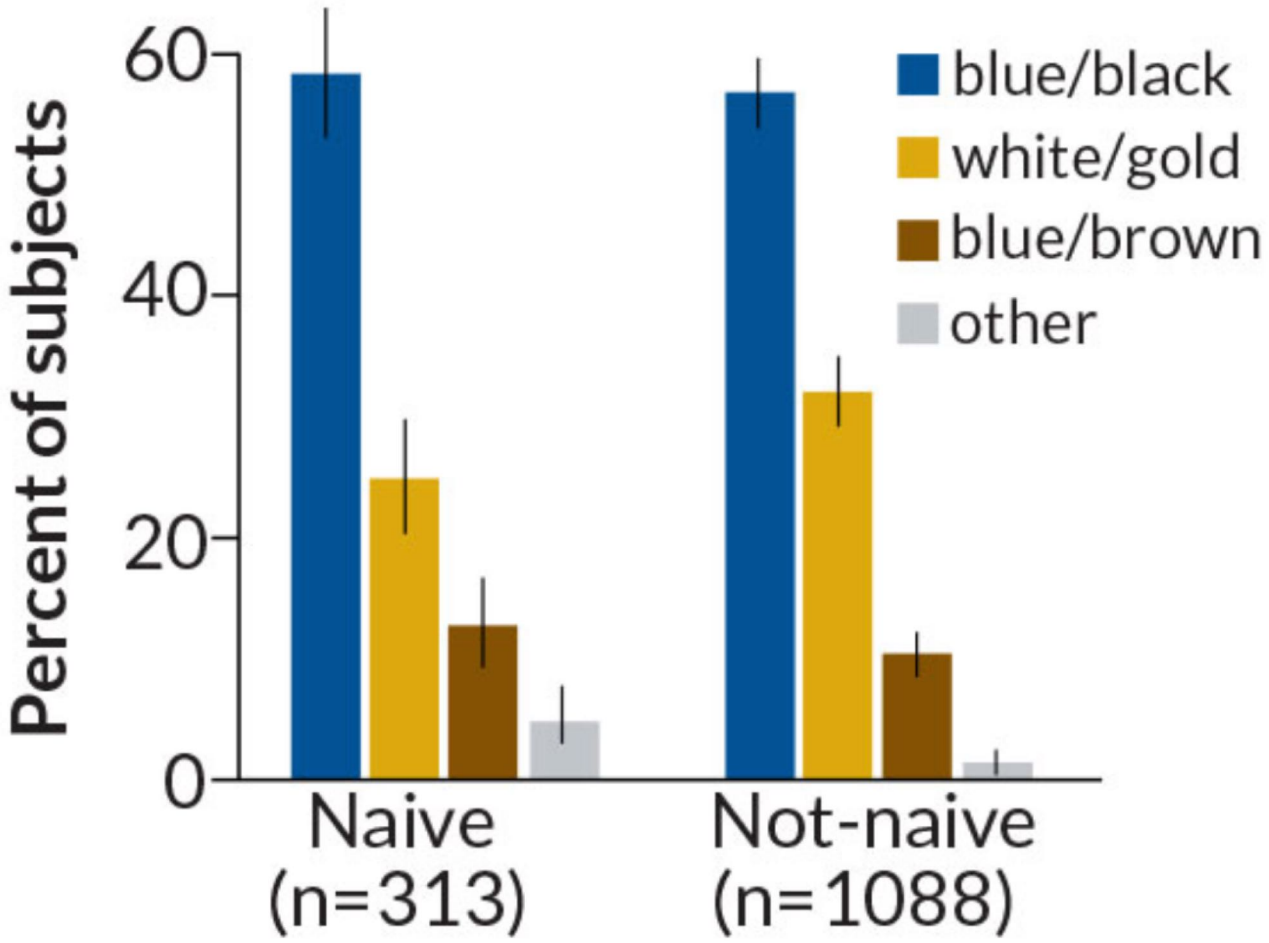
What are CNNs learning?



[Nguyen et al]

The Dress divided the Internet, but it's really about subtraction

BY **RACHEL EHRENBERG** 12:49PM, MAY 14, 2015



CNN mysteries

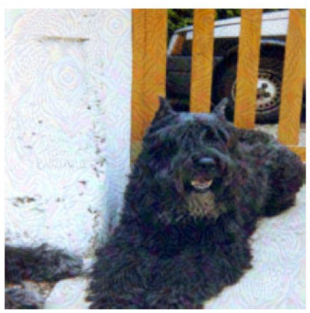
CNNs can sometimes outperform humans. Even after adding extreme noise, all these digits are recognized correctly!

But, only first column of these are recognized correctly. Second column digits are all wrong!

Goodfellow et al

Evidence of overfitting?

Goodfellow et al


wool


Indian elephant


Indian elephant


African grey


tabby


African grey


common newt


carousel

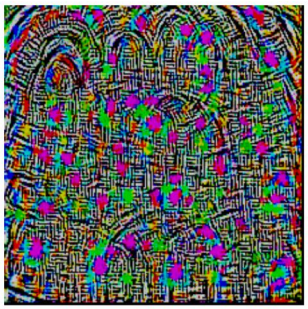

grey fox

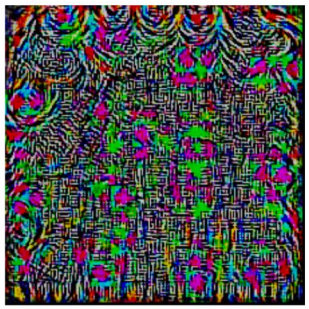

macaw

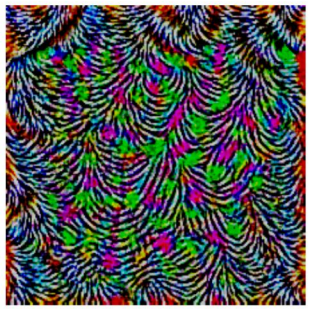

three-toed sloth

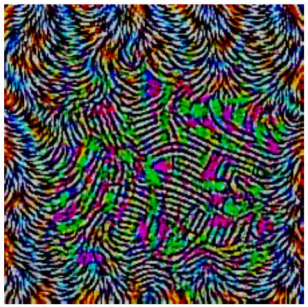

macaw

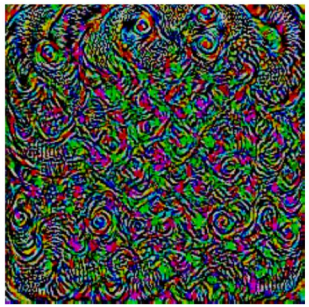
Moosavi-Dezfooli et al, Universal adversarial perturbations, CVPR 2017.

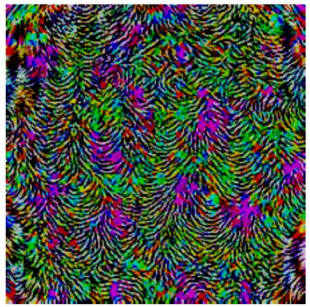

(a) CaffeNet

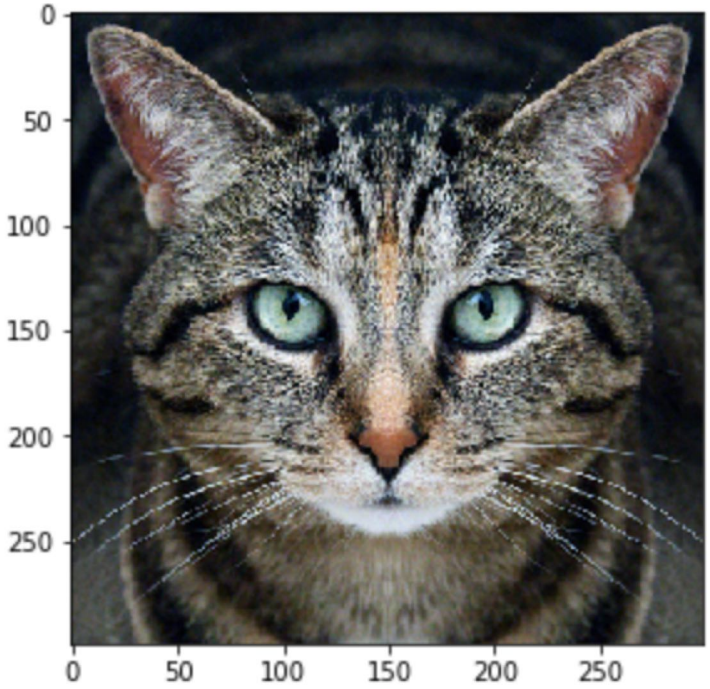

(b) VGG-F


(c) VGG-16

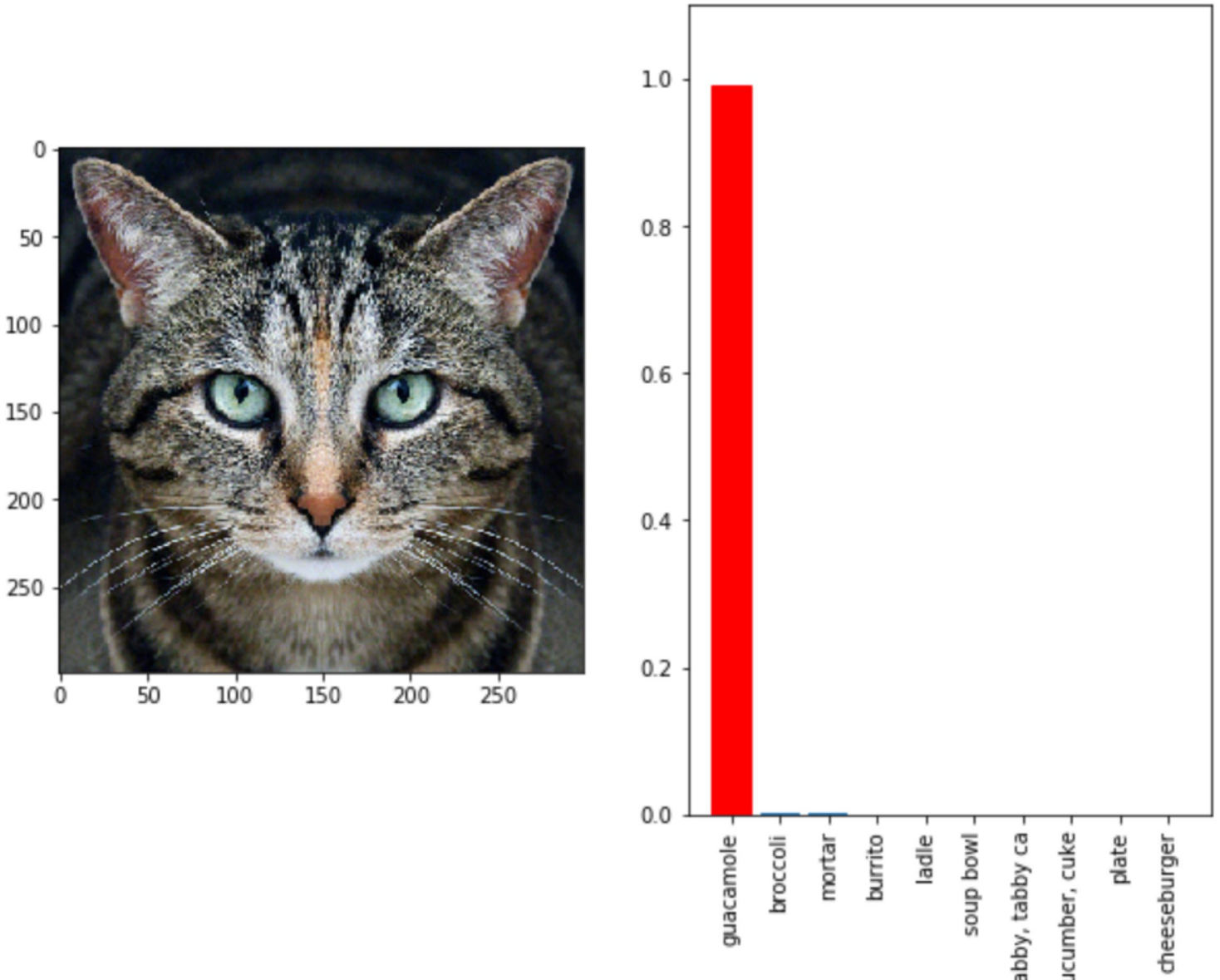

(d) VGG-19


(e) GoogLeNet

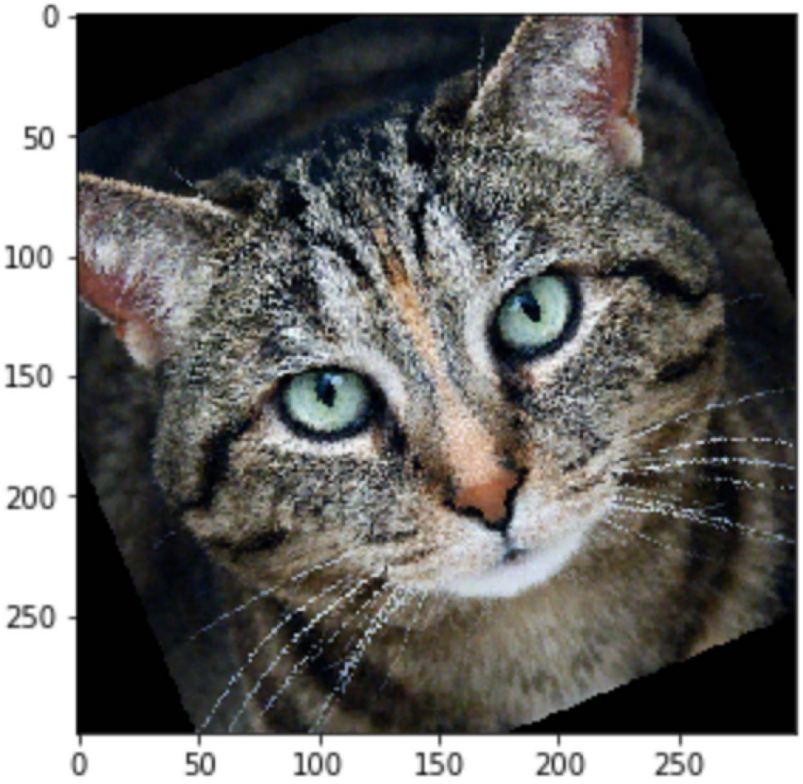

(f) ResNet-152

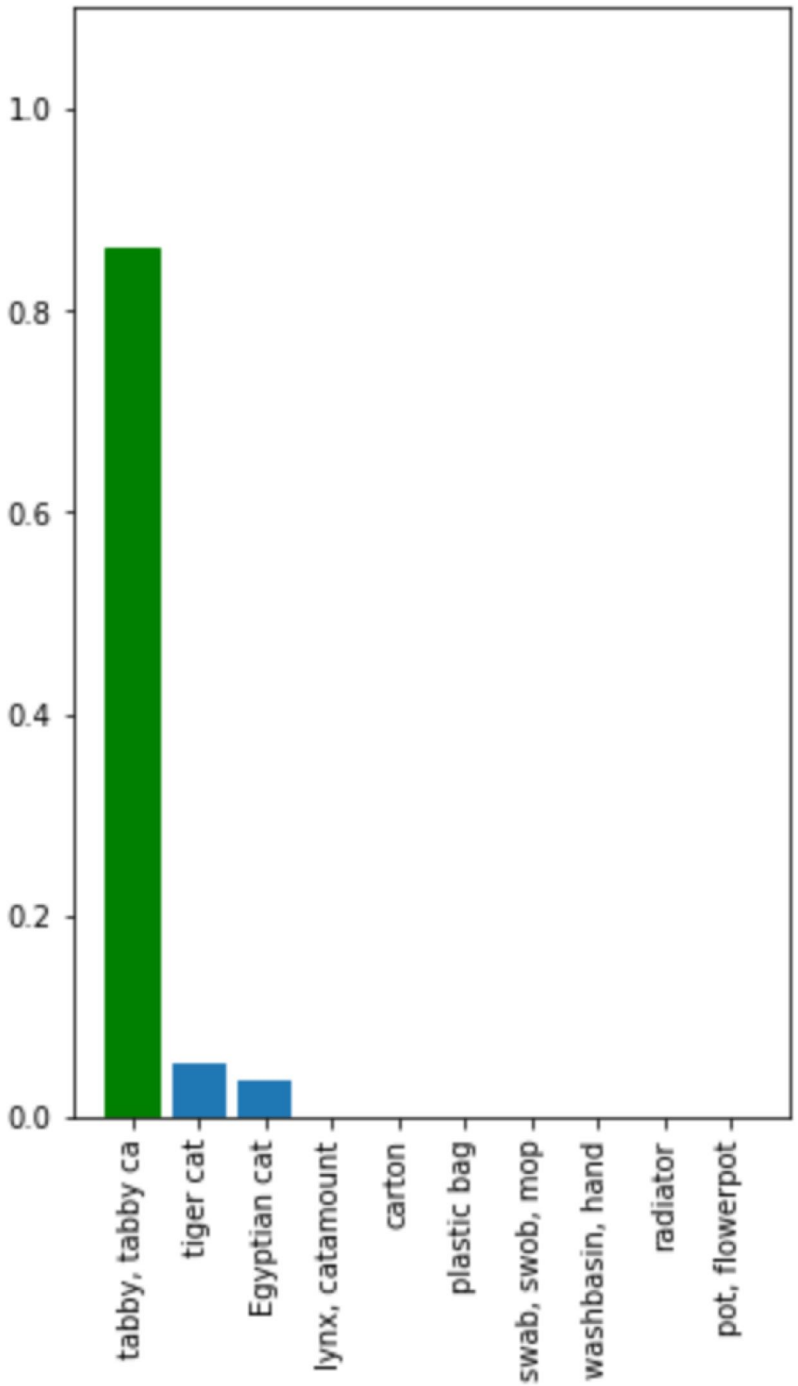


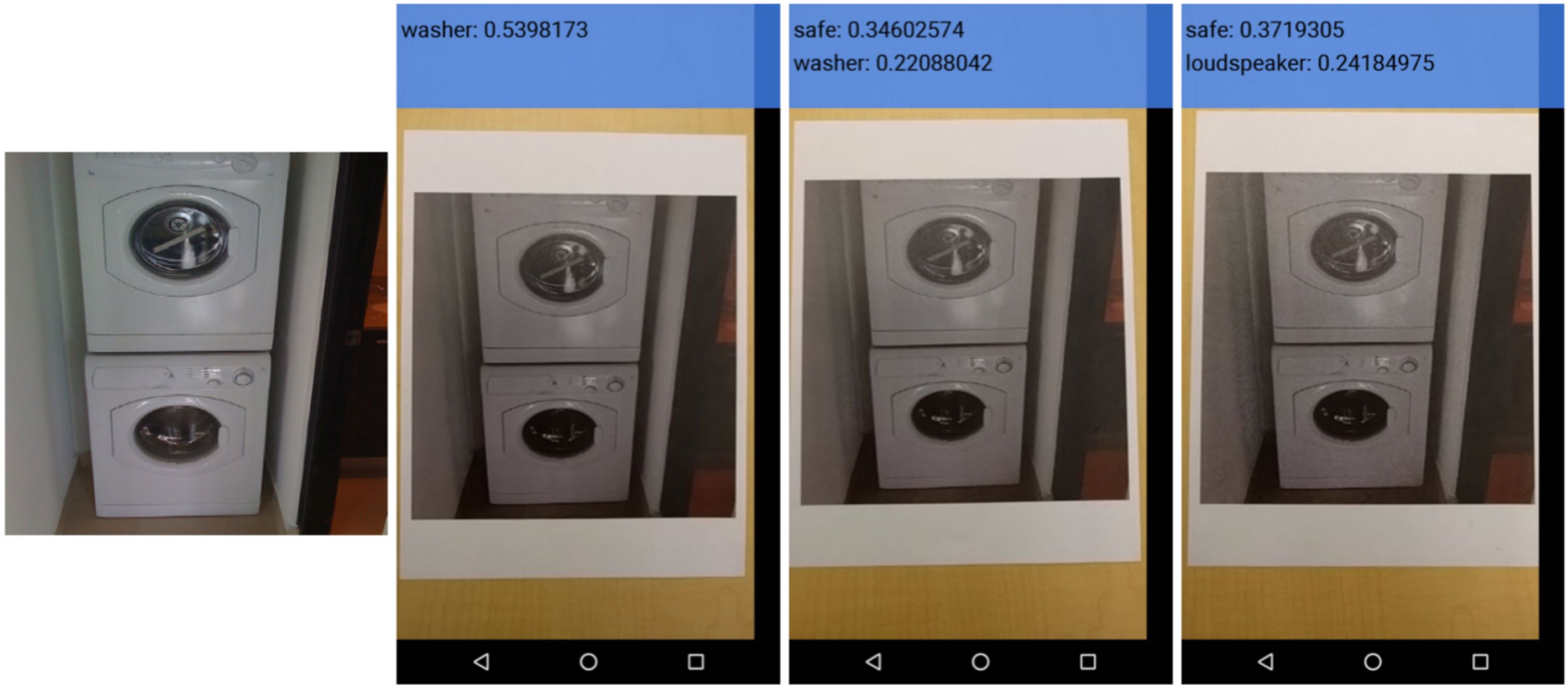
Elsayed 2018



Elsayed 2018

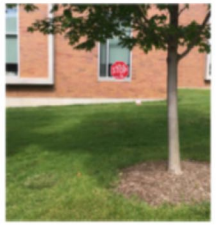
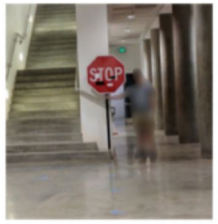

0
50
100
150
200
250
0
50
100
150
200
250


1.0
0.8
0.6
0.4
0.2
0.0
tabby, tabby ca
tiger cat
Egyptian cat
lynx, catamount
carton
plastic bag
swab, swob, mop
washbasin, hand
radiator
pot, flowerpot



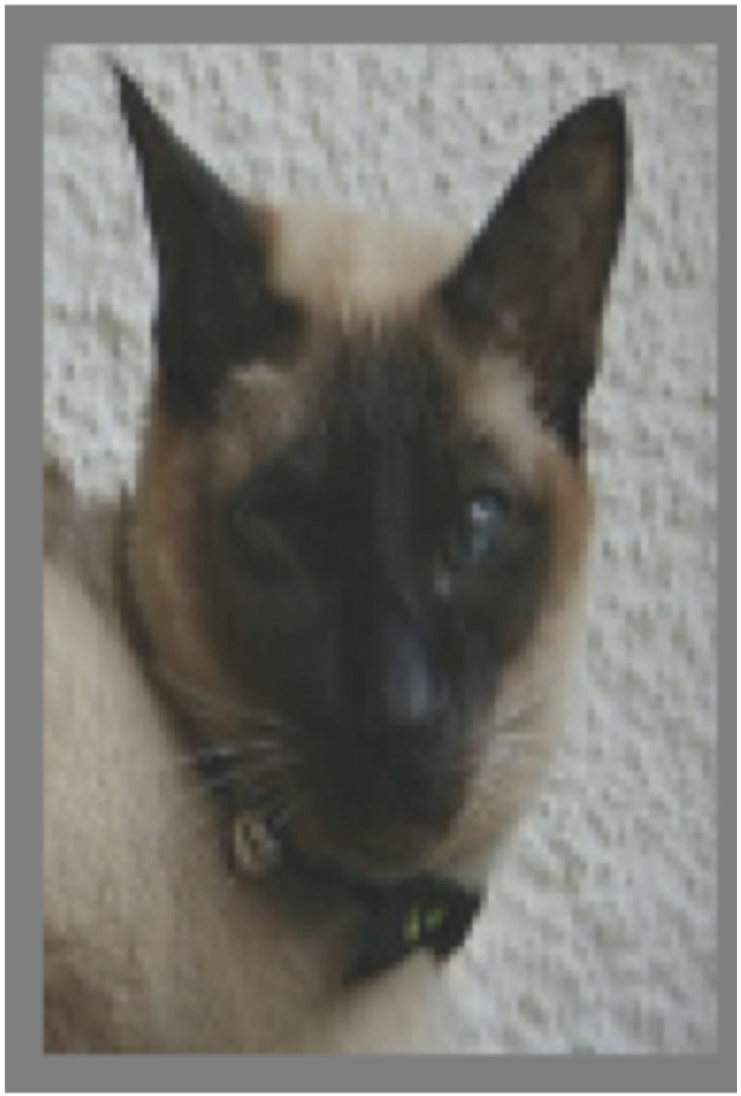
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Kurakin 2016

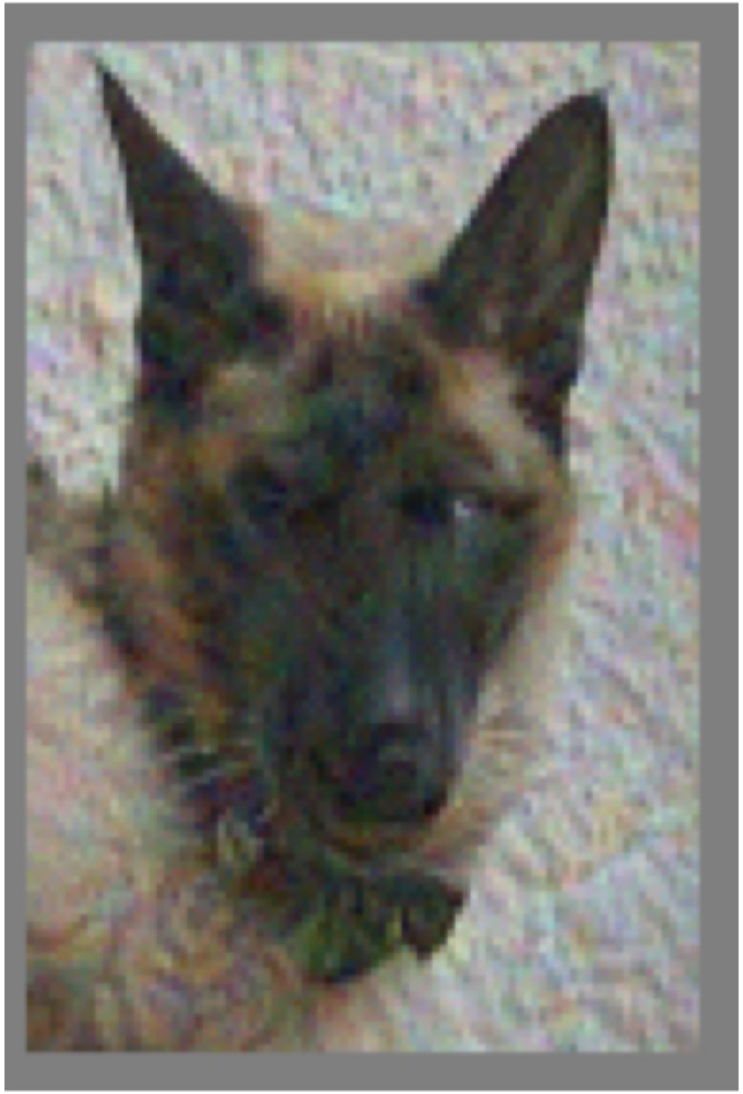
Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholt 2018

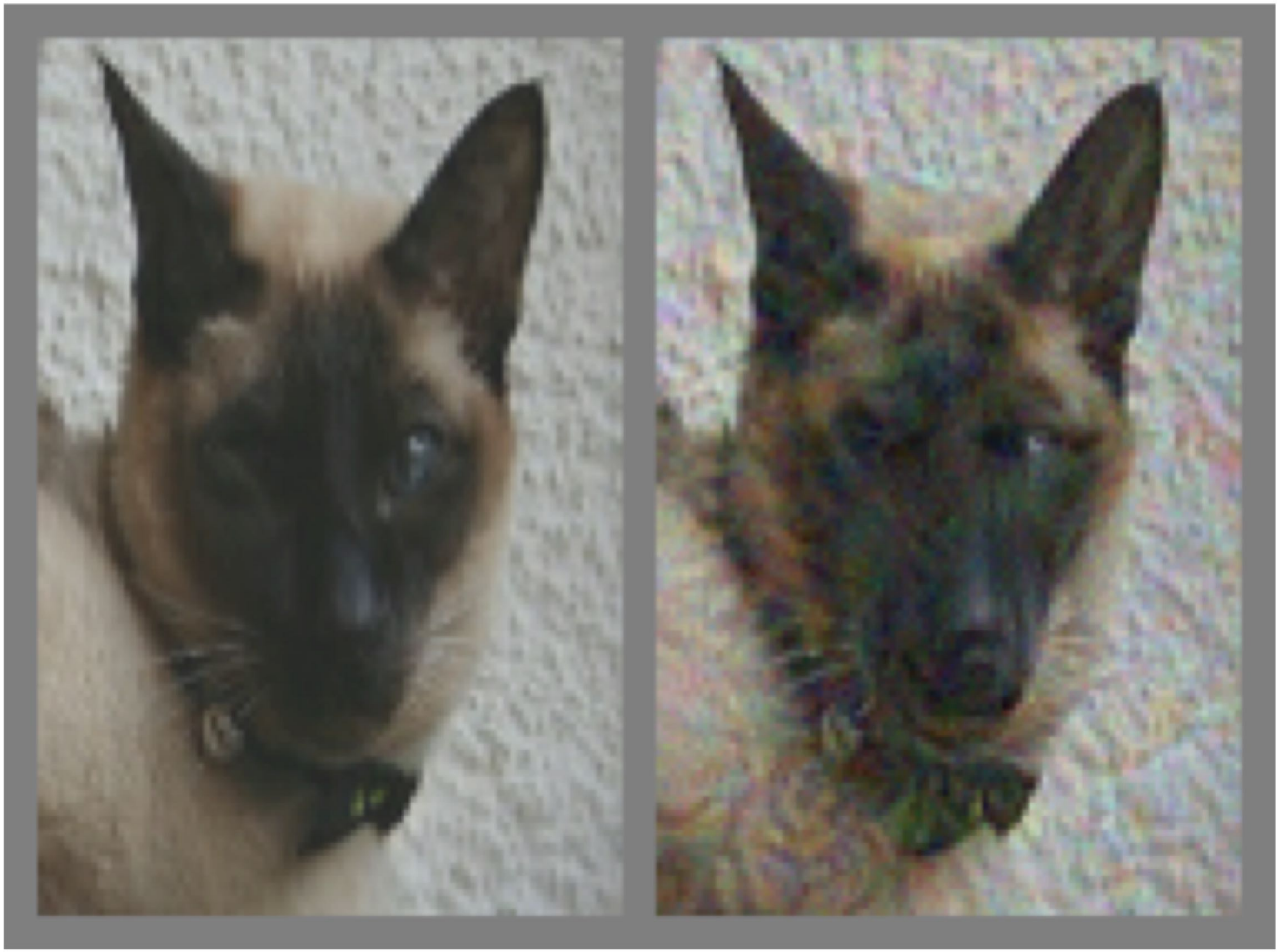

rifle
shield, buck
revolver, si
Fooling Neural Networks
in the Real World
labsix



Elsayed 2018



Elsayed 2018



Elsayed 2018

Visualizing Convnets

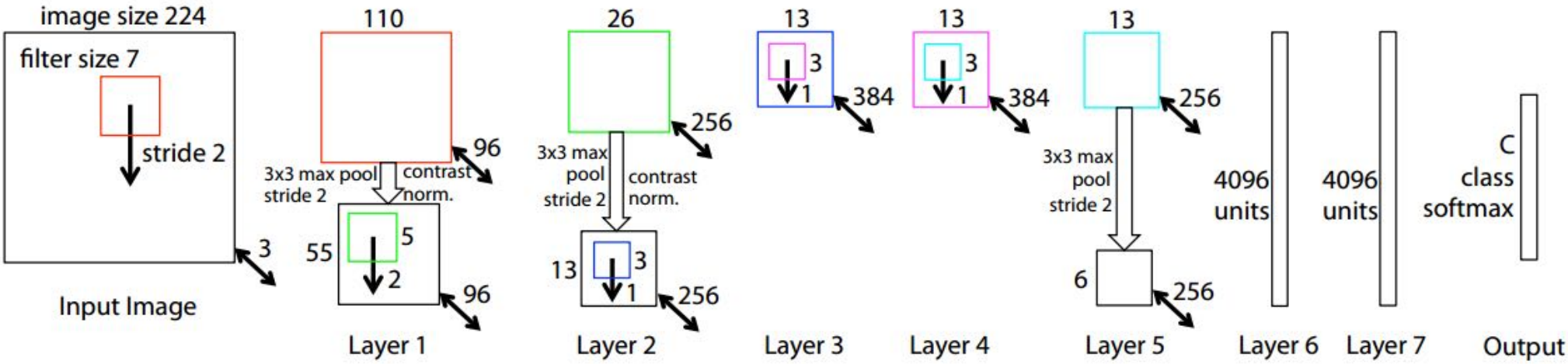
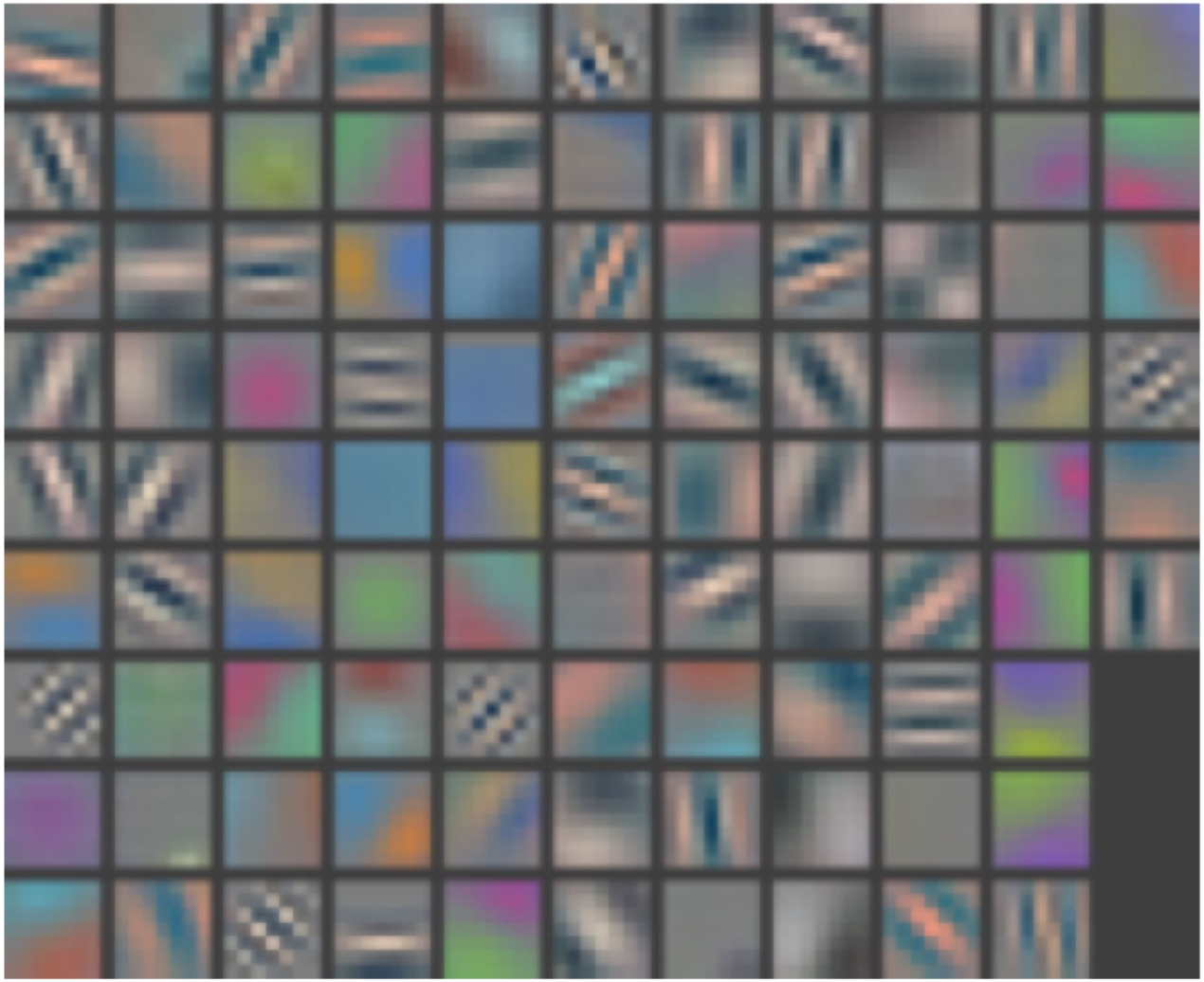


Figure 3. Architecture of our 8 layer convnet model. A 224 by 224 crop of an image (with 3 color planes) is presented as the input. This is convolved with 96 different 1st layer filters (red), each of size 7 by 7, using a stride of 2 in both x and y. The resulting feature maps are then: (i) passed through a rectified linear function (not shown), (ii) pooled (max within 3x3 regions, using stride 2) and (iii) contrast normalized across feature maps to give 96 different 55 by 55 element feature maps. Similar operations are repeated in layers 2,3,4,5. The last two layers are fully connected, taking features from the top convolutional layer as input in vector form ($6 \cdot 6 \cdot 256 = 9216$ dimensions). The final layer is a C-way softmax function, C being the number of classes. All filters and feature maps are square in shape.

M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, arXiv preprint, 2013

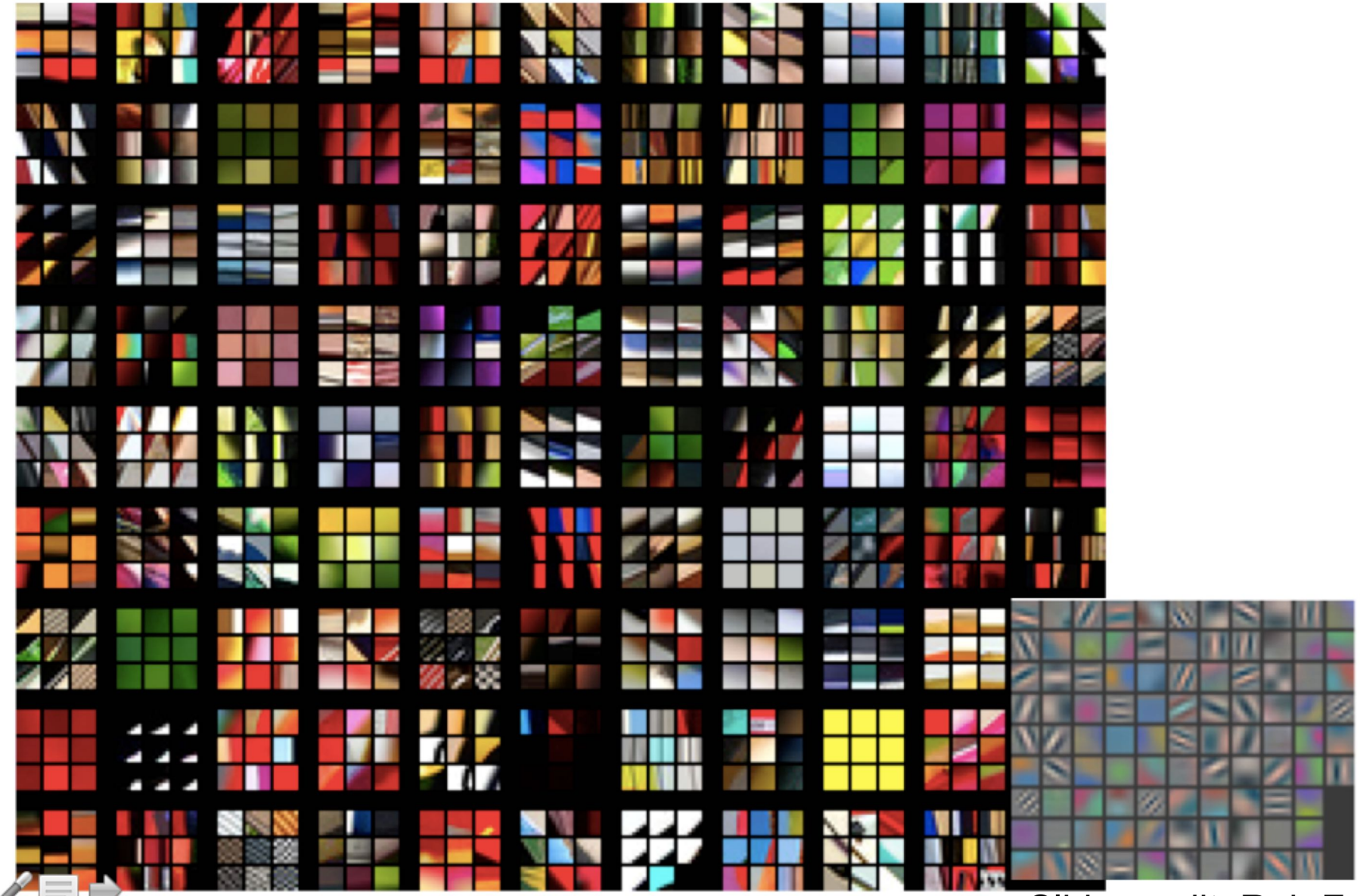
Slide credit: Rob Fergus

Layer 1 Filters

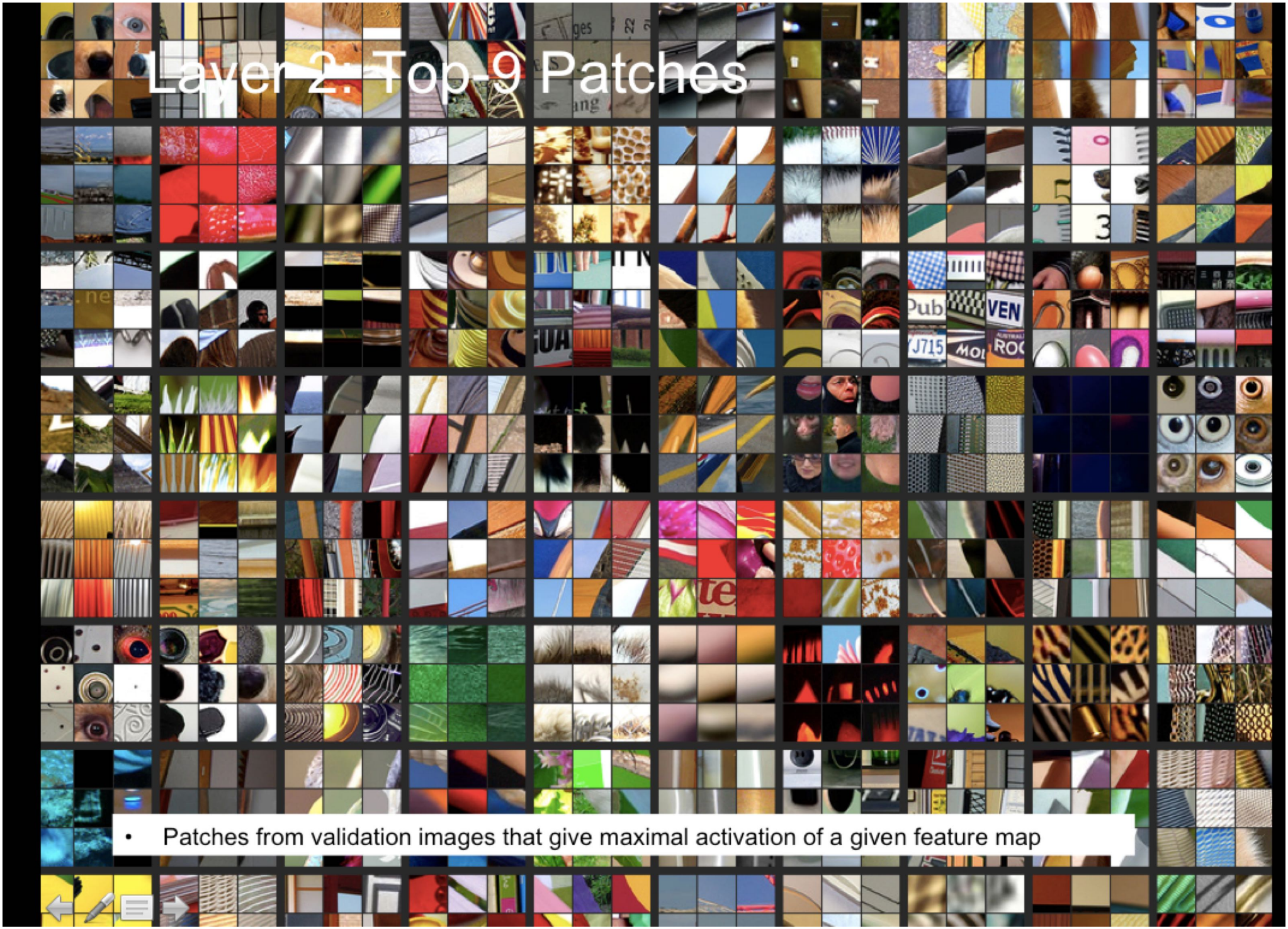


Slide credit: Rob Fergus

Layer 1: Top-9 Patches



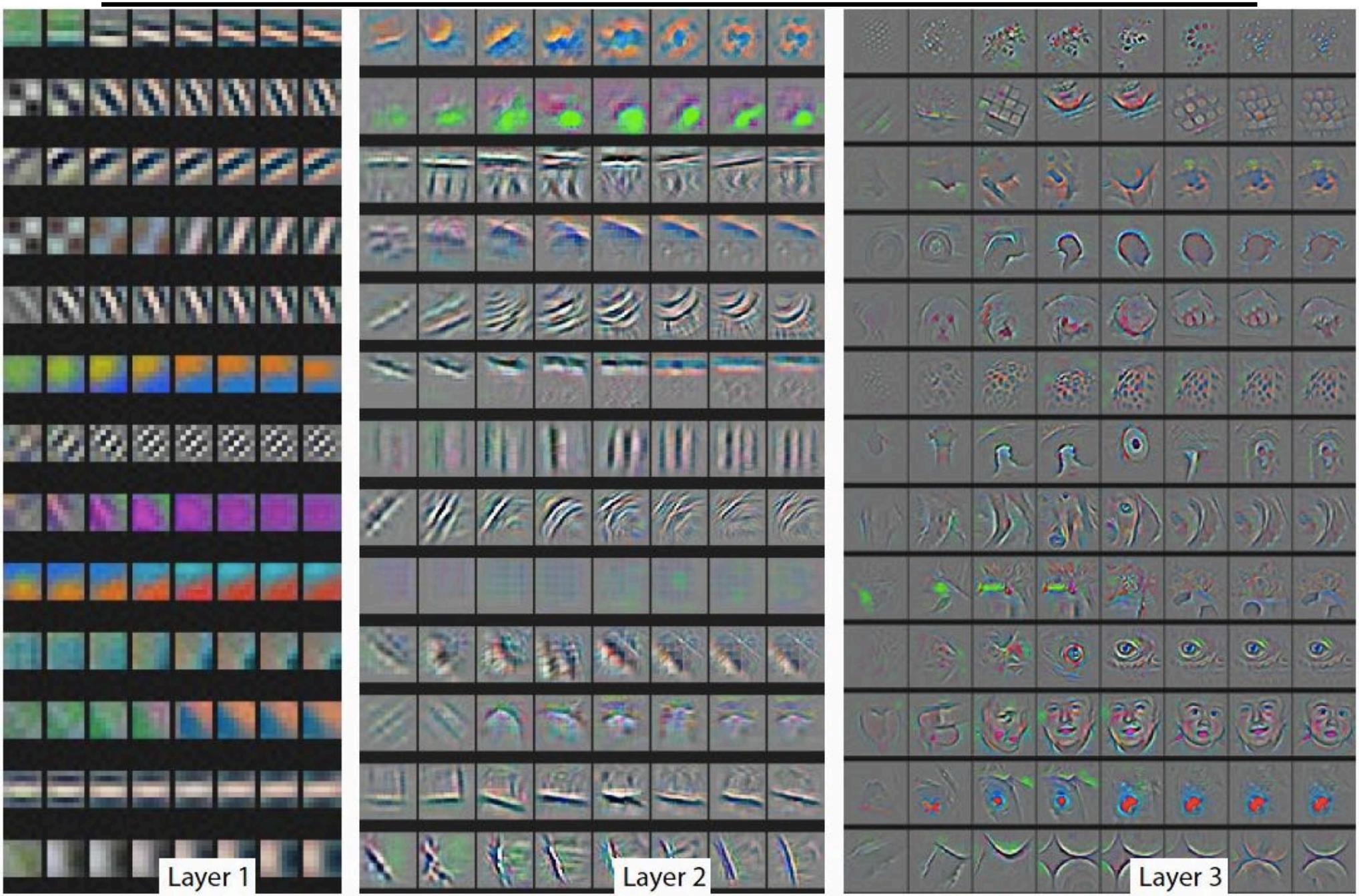
Slide credit: Rob Fergus


Layer 2: Top-9 Patches
• Patches from validation images that give maximal activation of a given feature map

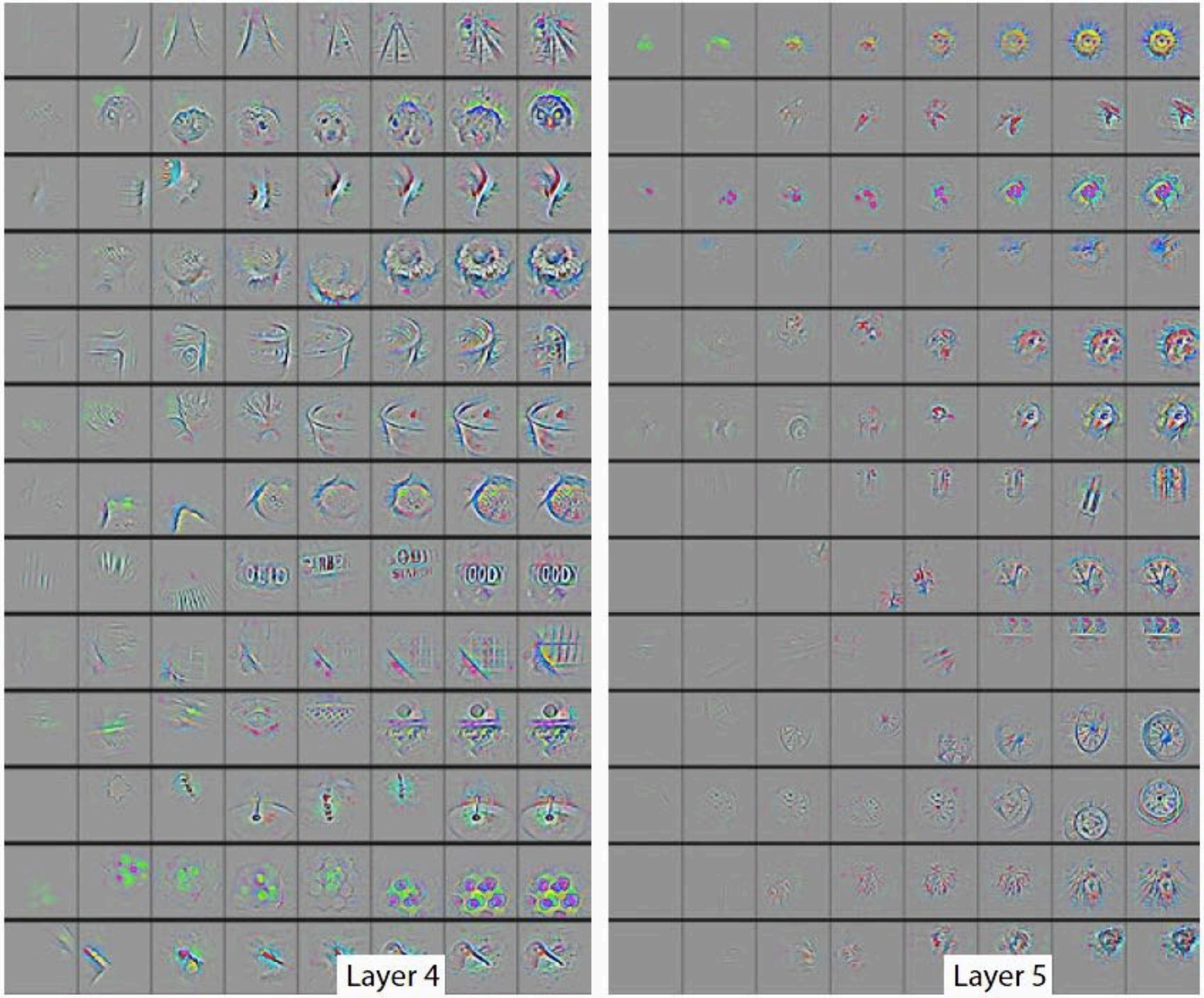

Layer 3: Top-9 Patches

Layer 5: Top-9 Patches

Evolution of Features During Training

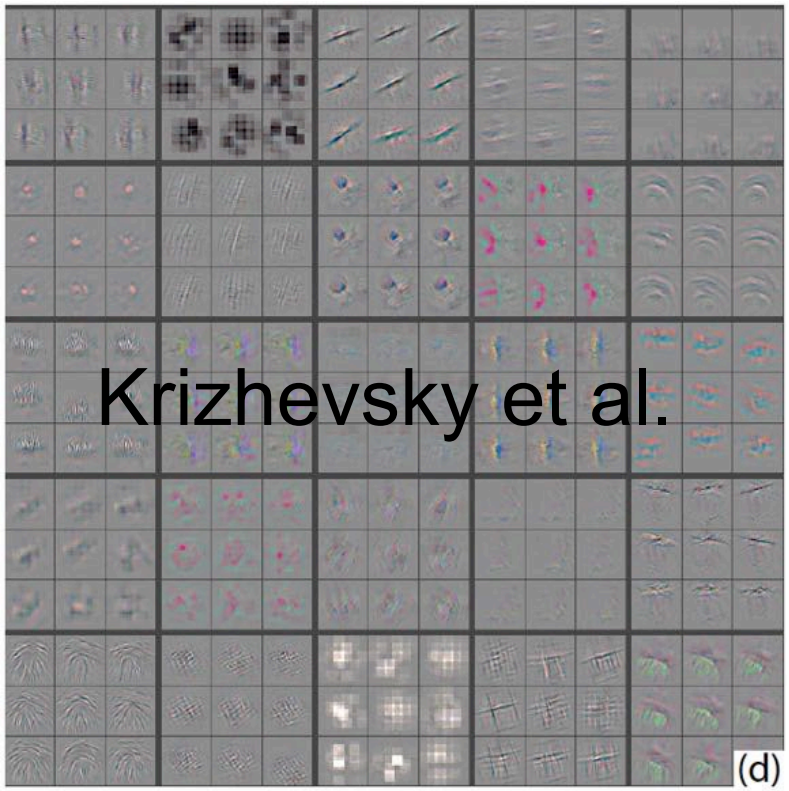


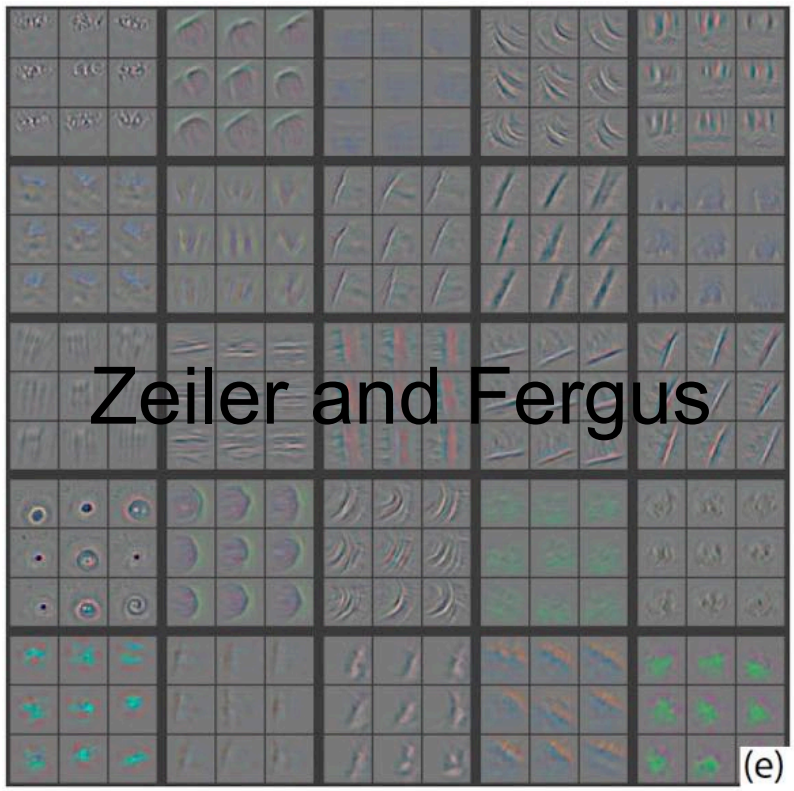
Evolution of Features During Training



Diagnosing Problems

- Visualization of Krizhevsky et al.'s architecture showed some problems with layers 1 and 2
  - Large stride of 4 used
- Alter architecture: smaller stride & filter size
  - Visualizations look better
  - Performance improves



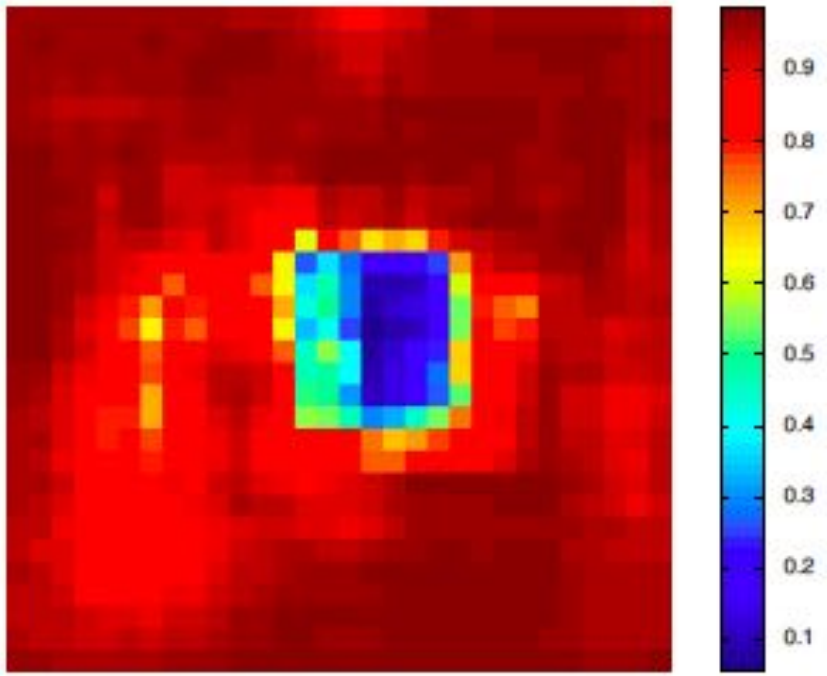


Occlusion Experiment

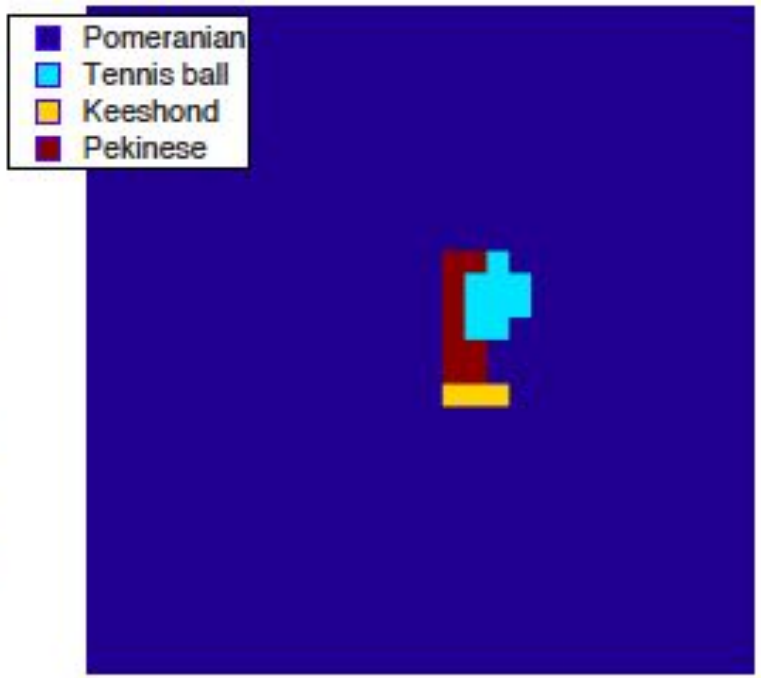
- Mask parts of input with occluding square
- Monitor output

Input image

p(True class)

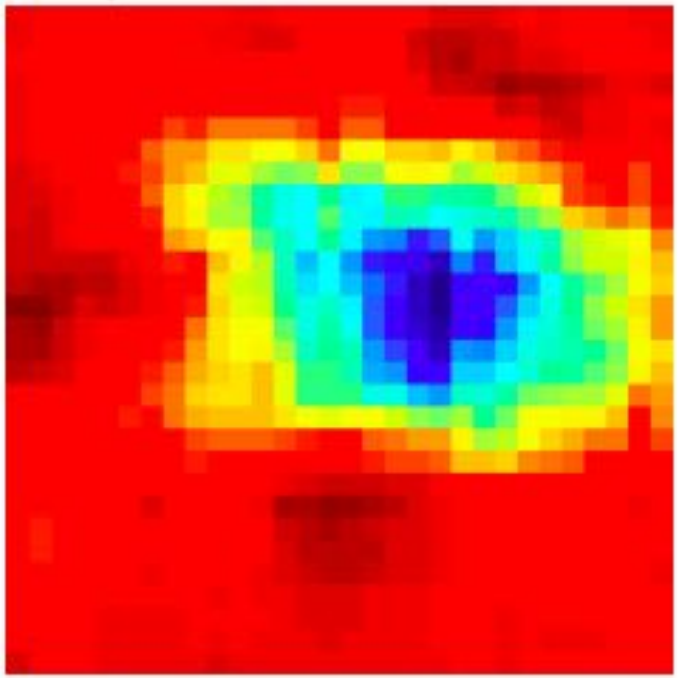


Most probable class

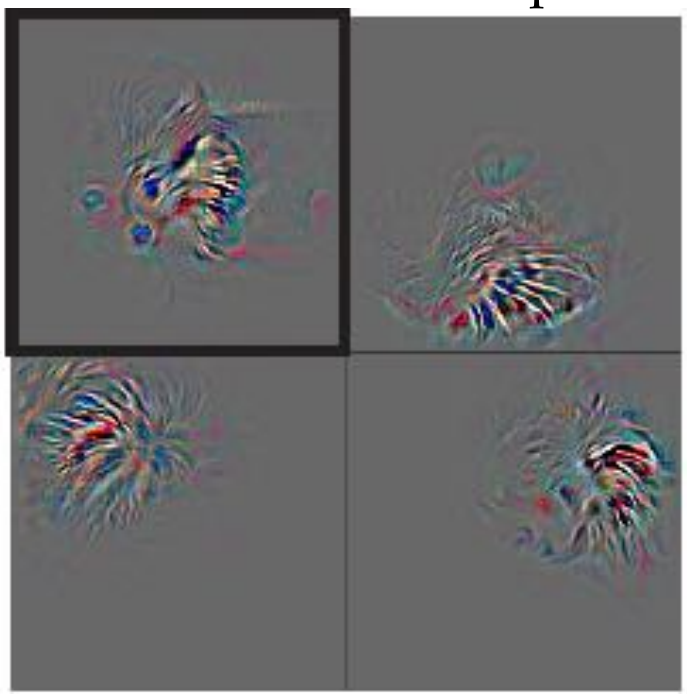


Input image

Total activation in most active 5th layer feature map

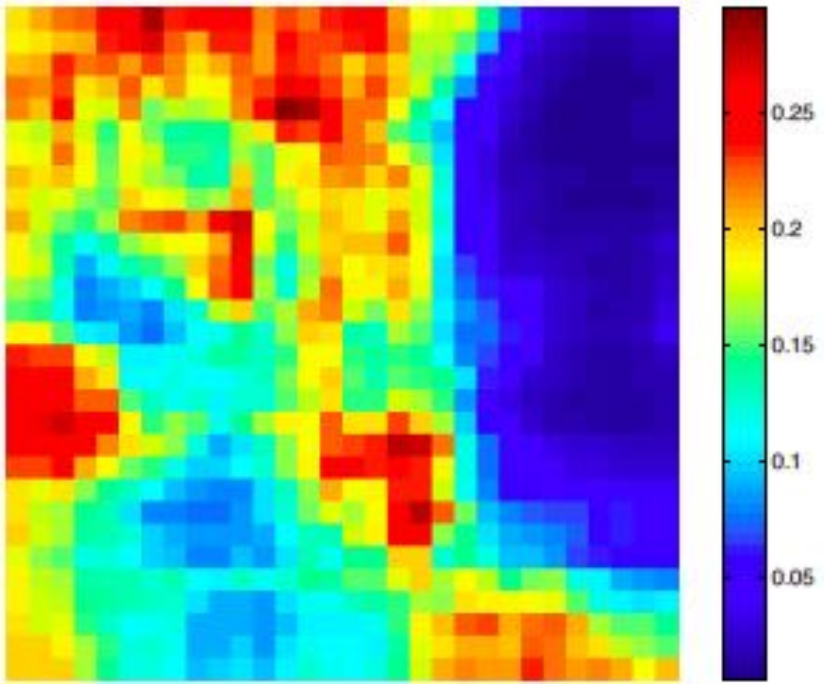


Other activations from same feature map

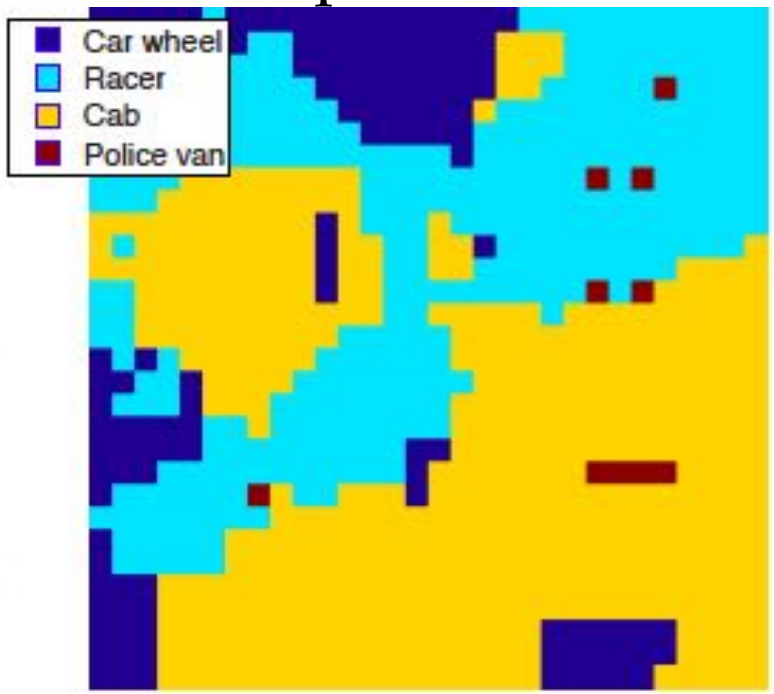


Input image

p(True class)

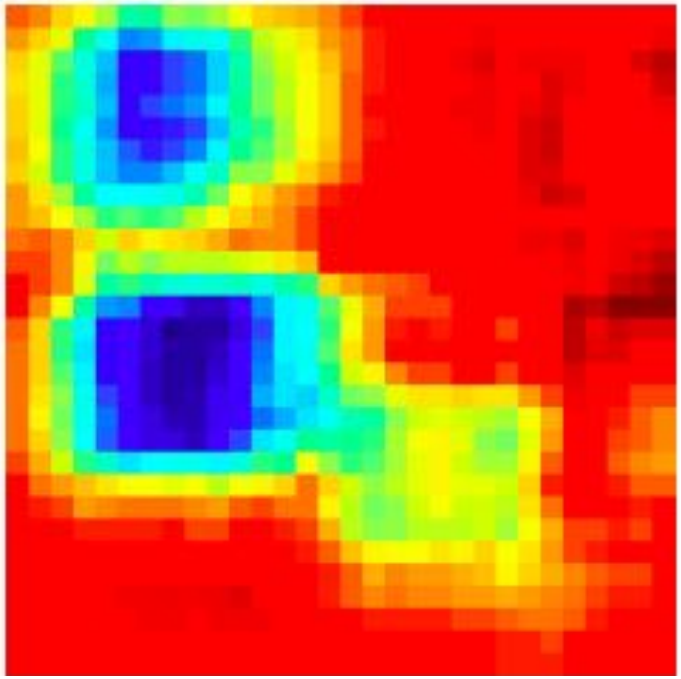


Most probable class



Input image

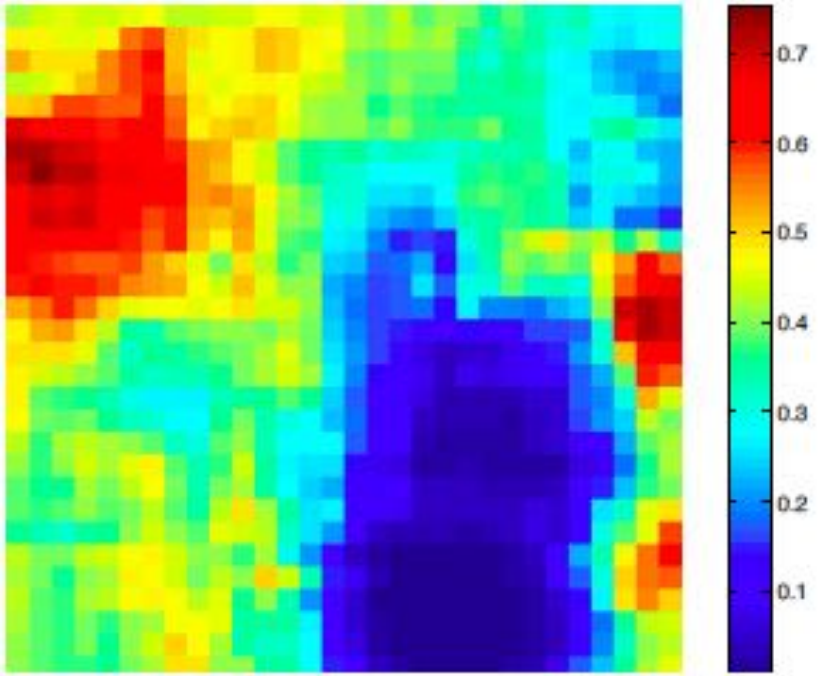
Total activation in most active 5th layer feature map



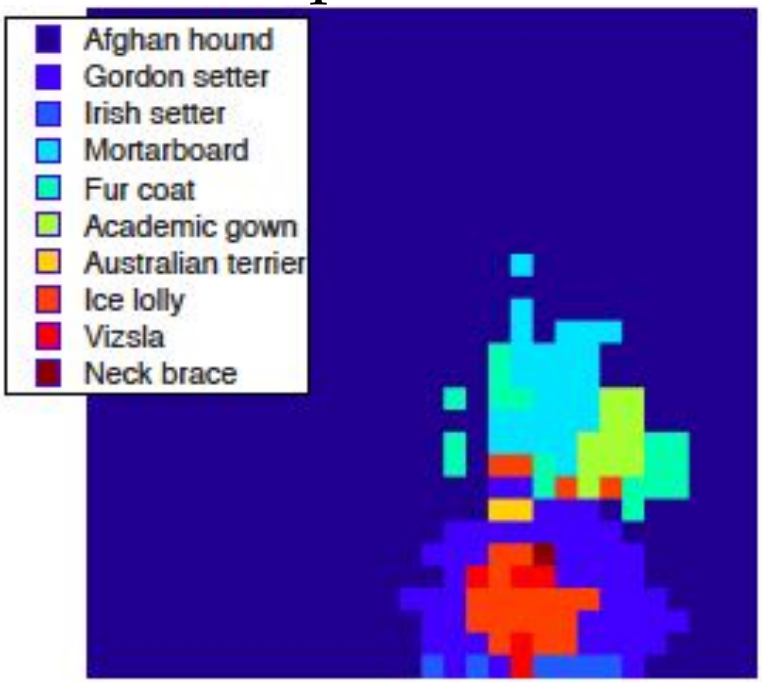
Other activations from same feature map

Input image

p(True class)

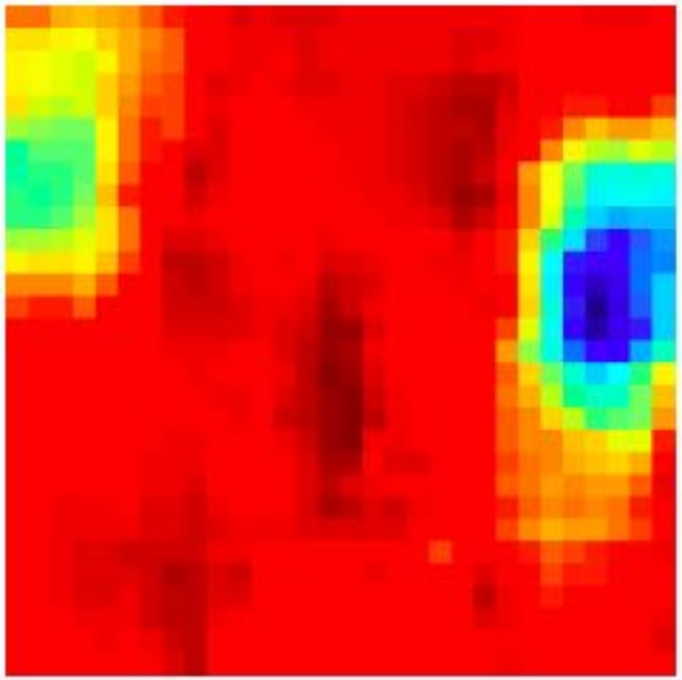


Most probable class

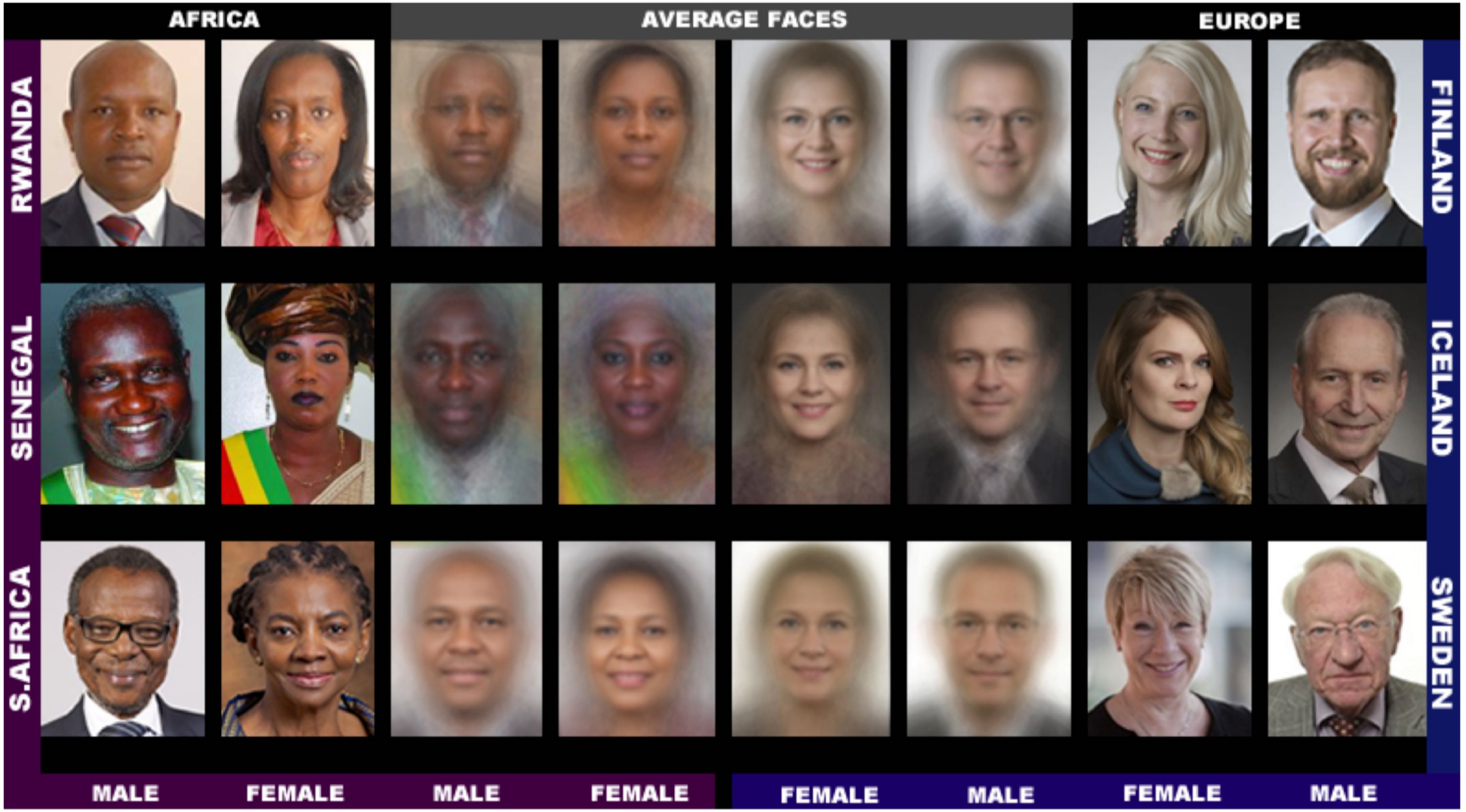


Input image

Total activation in most active 5th layer feature map



Other activations from same feature map



Buolamwini 2018

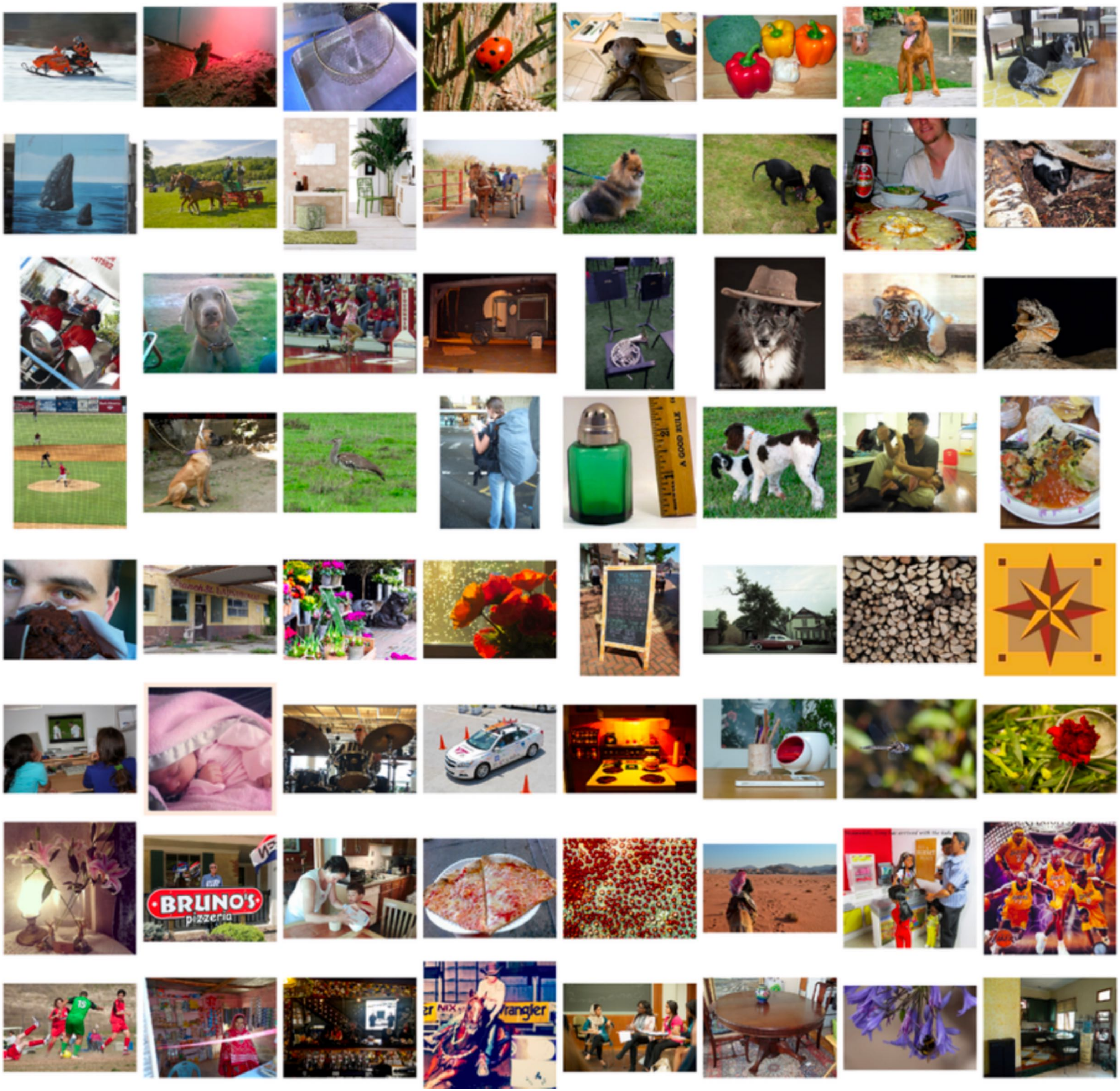
Classifier	Metric	DF	DM	LF	LM
MSFT	PPV(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	TPR(%)	**100**	84.2	**100**	**100**
	FPR(%)	**15.8**	0.0	0.0	0.0
Face++	PPV(%)	64.0	99.5	**100**	**100**
	Error Rate(%)	**36.0**	0.5	0.0	0.0
	TPR(%)	99.0	77.8	**100**	96.9
	FPR(%)	**22.2**	1.03	3.08	0.0
IBM	PPV(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	TPR(%)	90.4	78.0	96.4	**100**
	FPR(%)	**22.0**	9.7	0.0	3.6

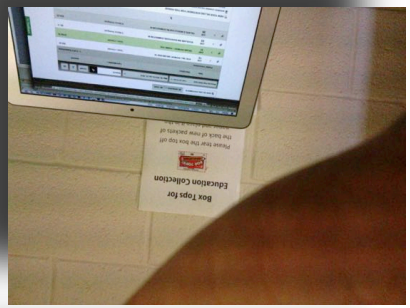
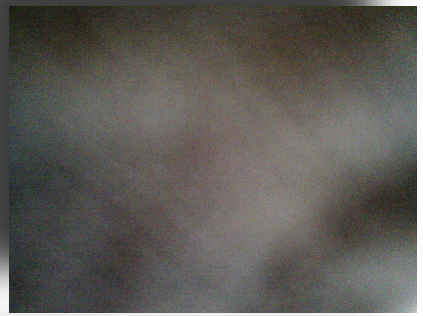
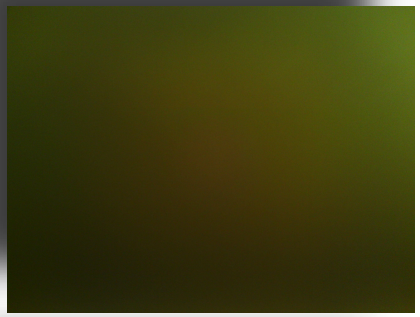
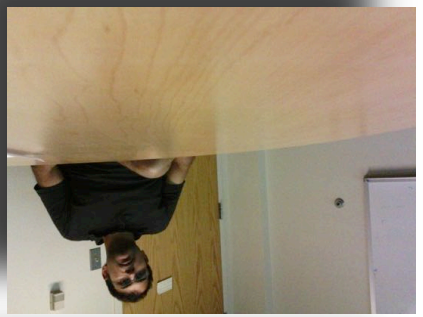
Table 5: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the South African subset of the PPB dataset. Results for South Africa follow the overall trend with the highest error rates seen on darker-skinned females.

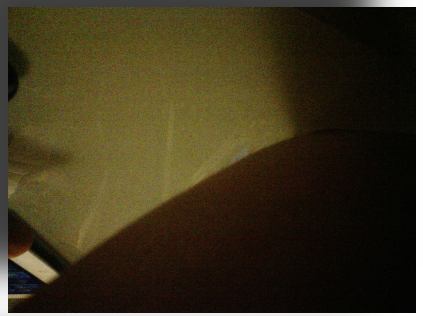
Moral:

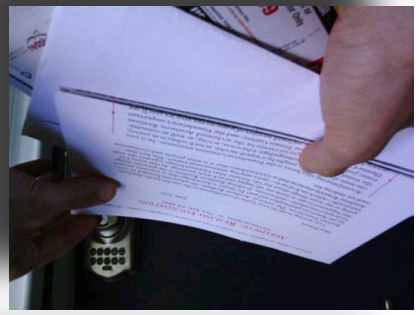
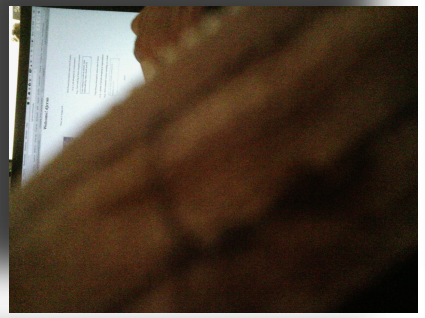
Algorithms are important.

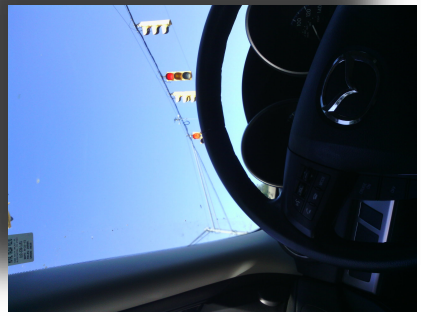
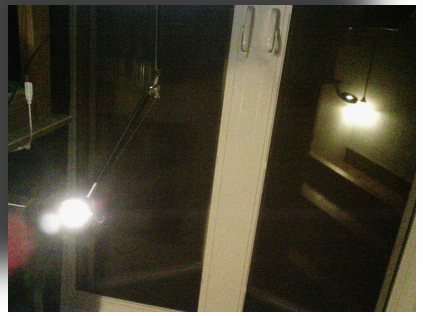
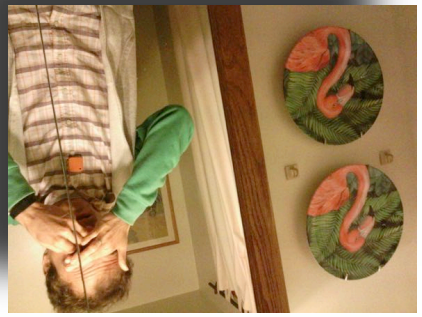
But so is the training data!


BRUNO'S
pizzeria
A GOOD RULE
Wrangler


Box Tops for
Education Collection




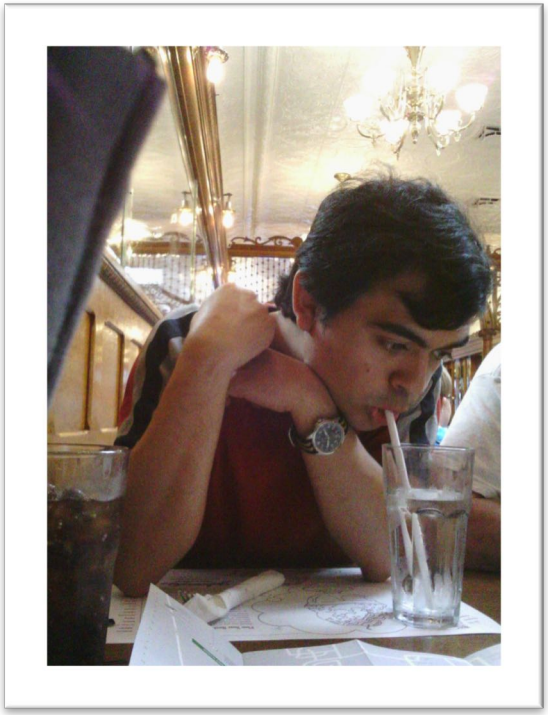





Lite



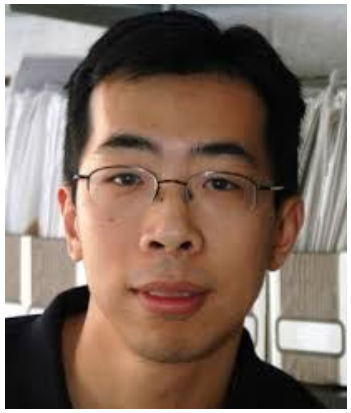
Studying visual learning in children using computer vision

(and vice-versa, maybe, eventually?)

David Crandall
School of Informatics, Computing, and Engineering
Indiana University
Bloomington

David Crandall

Linda Smith



Chen Yu

Sven Bambach

Elizabeth Clerkin

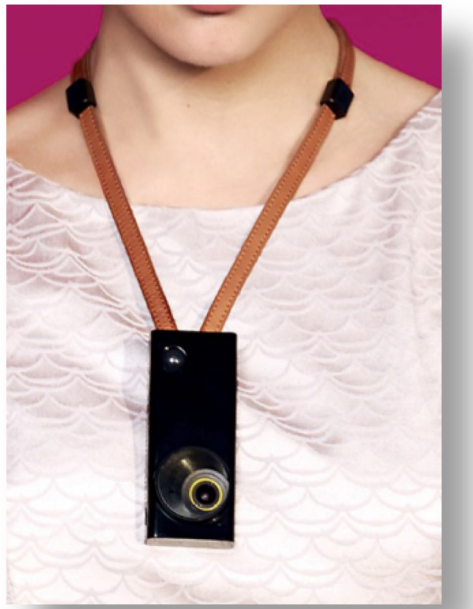
Christina DeSerio

Swapnaa Jayaraman

Zehua Zhang

- Google

- Narrative

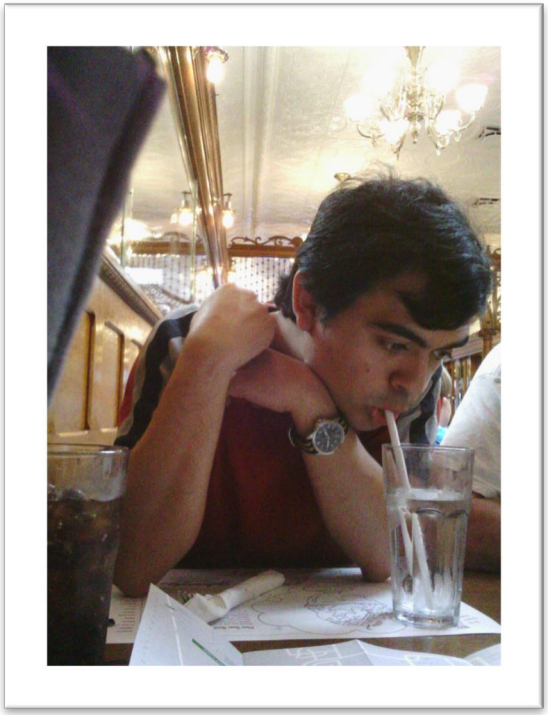


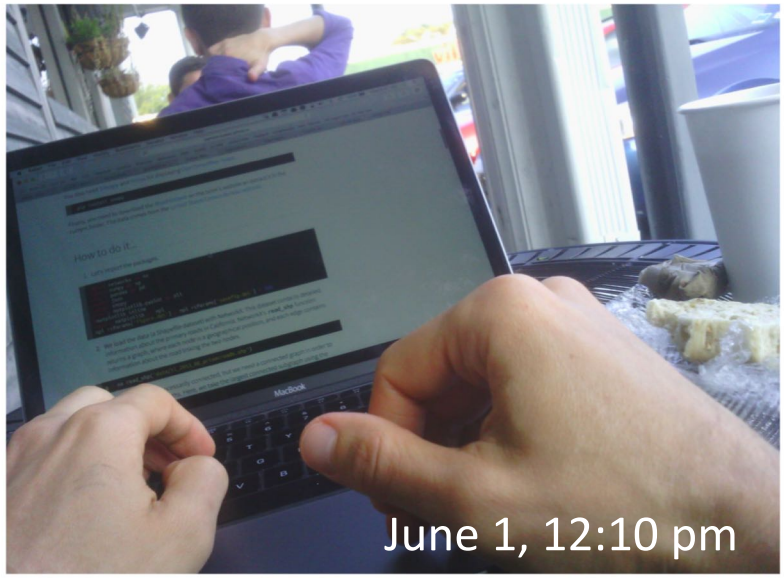
- Autographer

- GoPro

- Snapchat


Lite

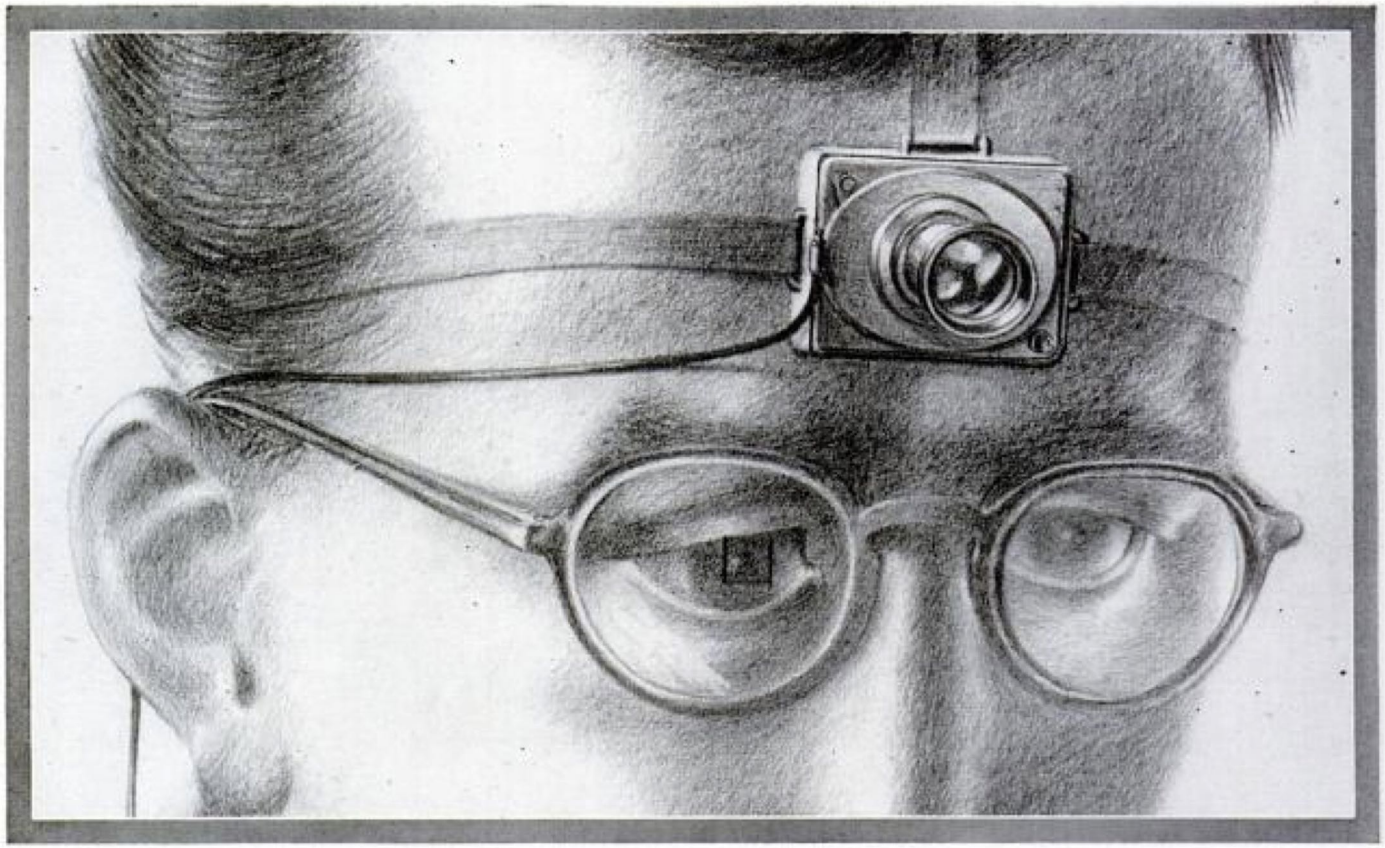



June 1, 12:10 pm


July 1, 12:10 pm


Aug 1, 12:10 pm


Sept 1, 12:10 pm



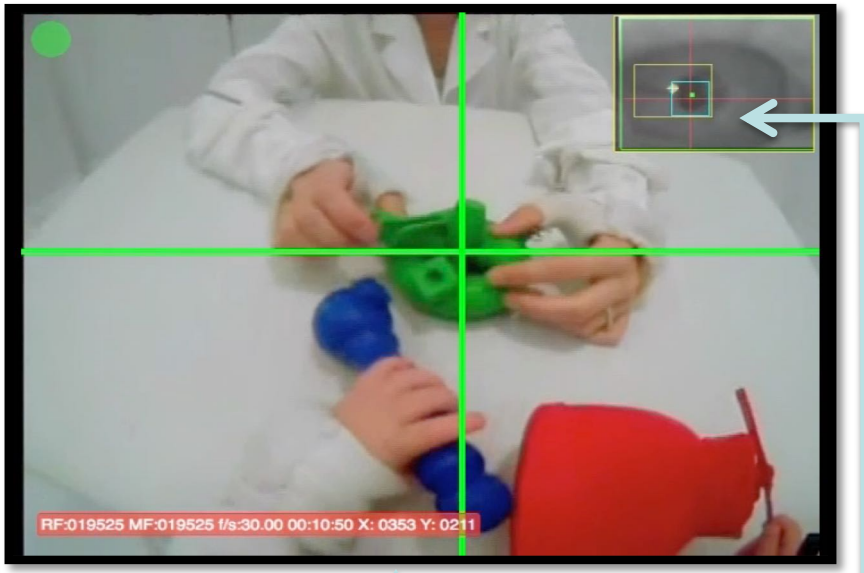
A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJECT

AS WE MAY THINK

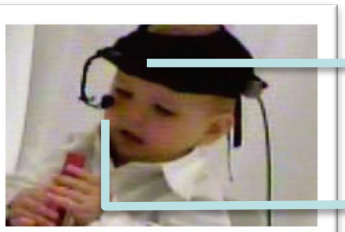
A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD

Vannevar Bush, *The Atlantic,* 1945

child's first person view

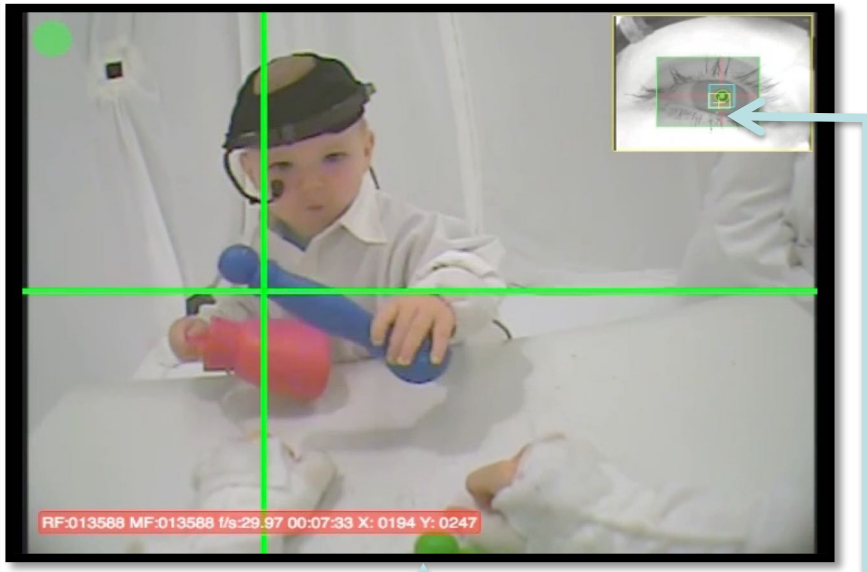


from eye camera

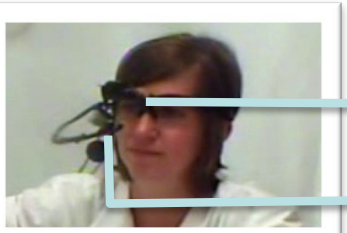


from head camera

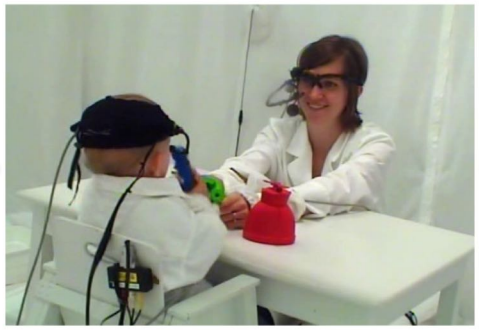
parent's first person view

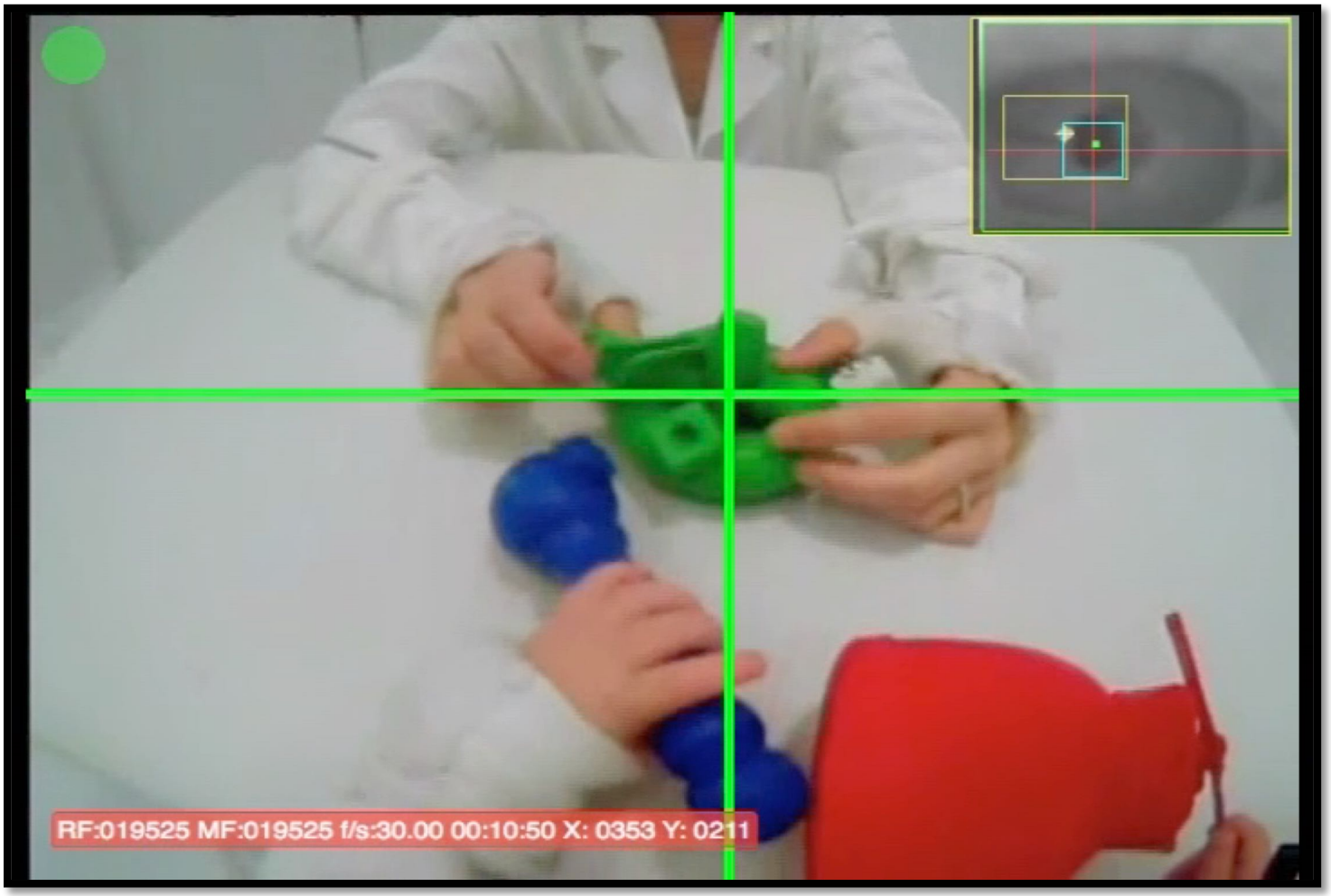


from eye camera

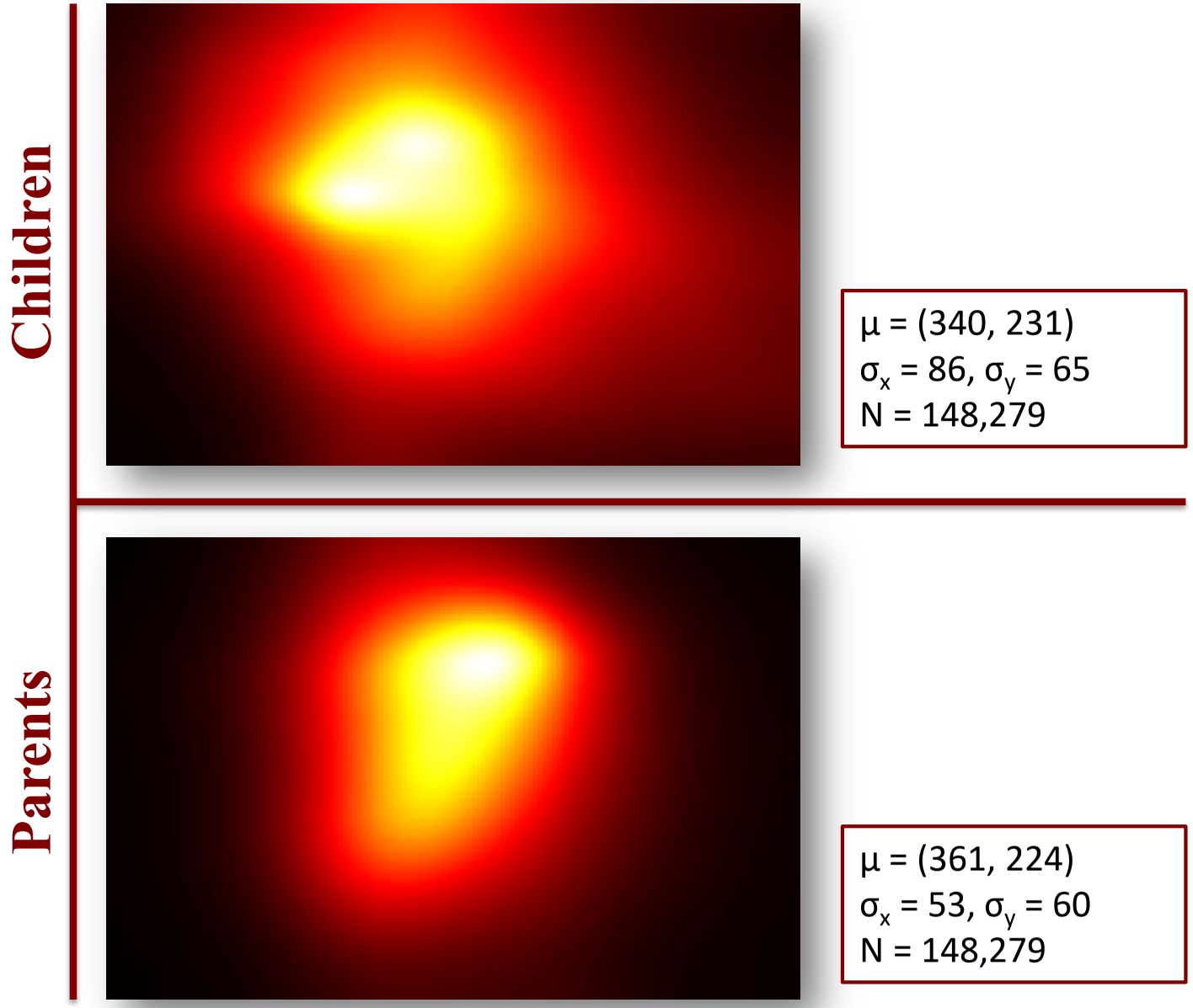


from head camera




RF:019525 MF:019525 f/s:30.00 00:10:50 X: 0353 Y: 0211

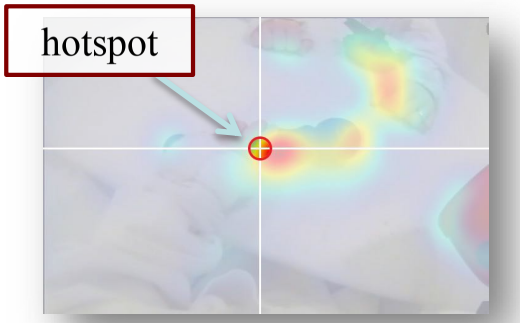
Eye gaze ***within the visual field***



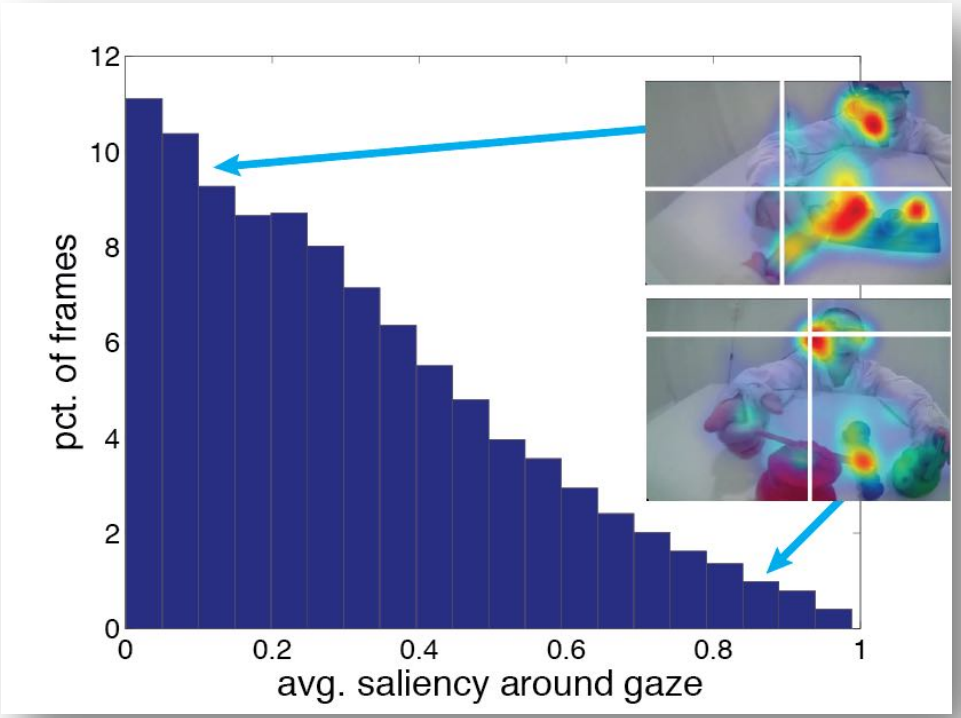
S. Bambach, D. Crandall, C. Yu. Understanding embodied visual attention in child-parent interactions, *IEEE International Conference on Development and Learning,* 2013.

Visual saliency in first-person views

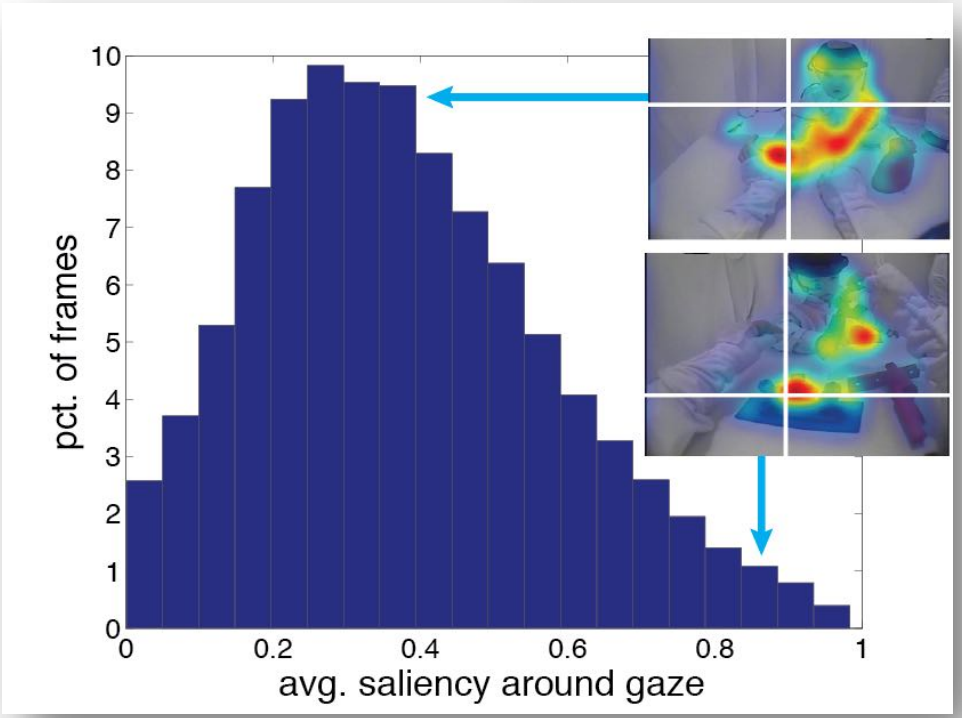
Comparison of average saliency within hotspot around gaze location on 148,000 frames



→ Gaze predictiveness differs significantly



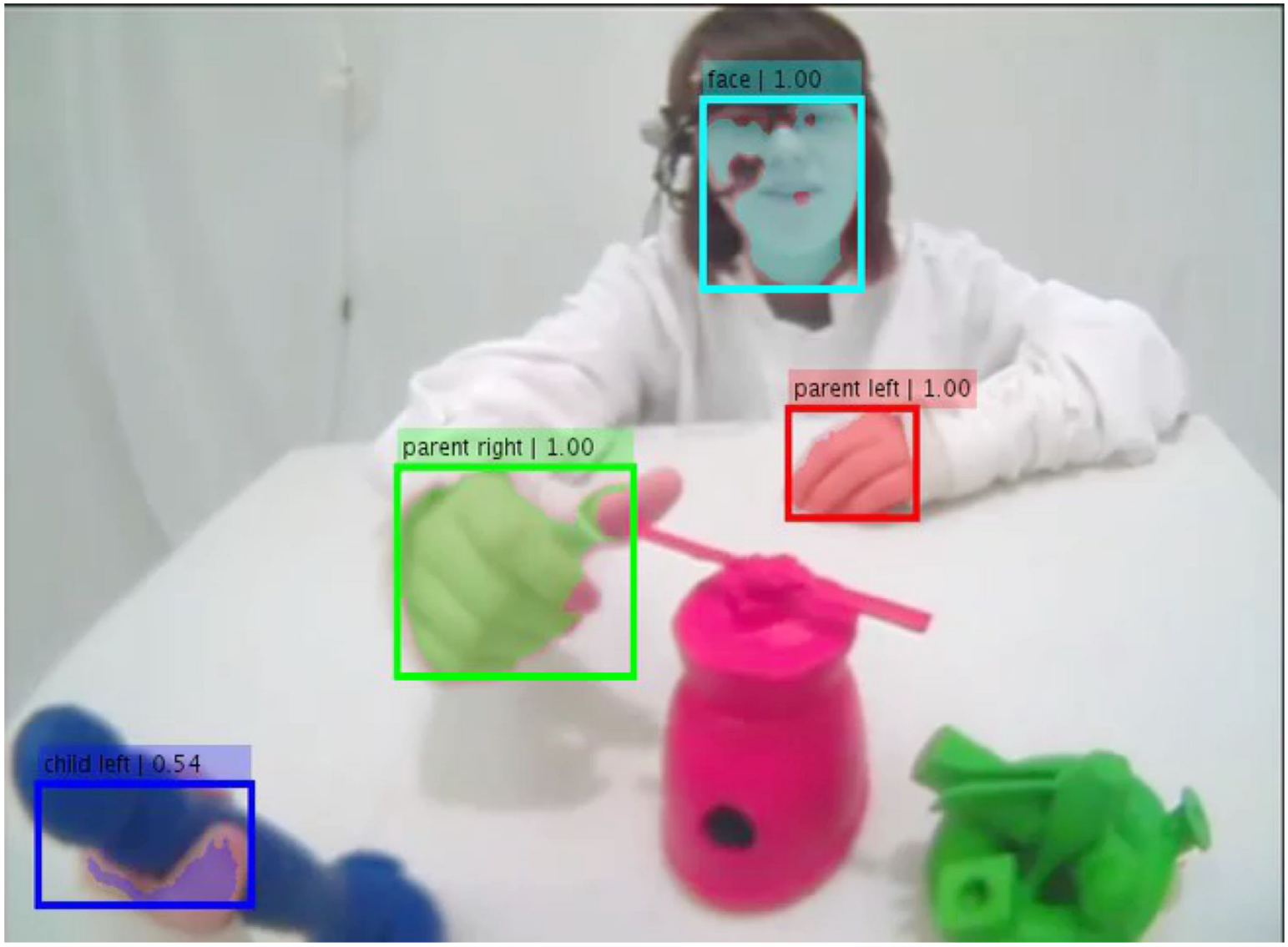
Child



Parent

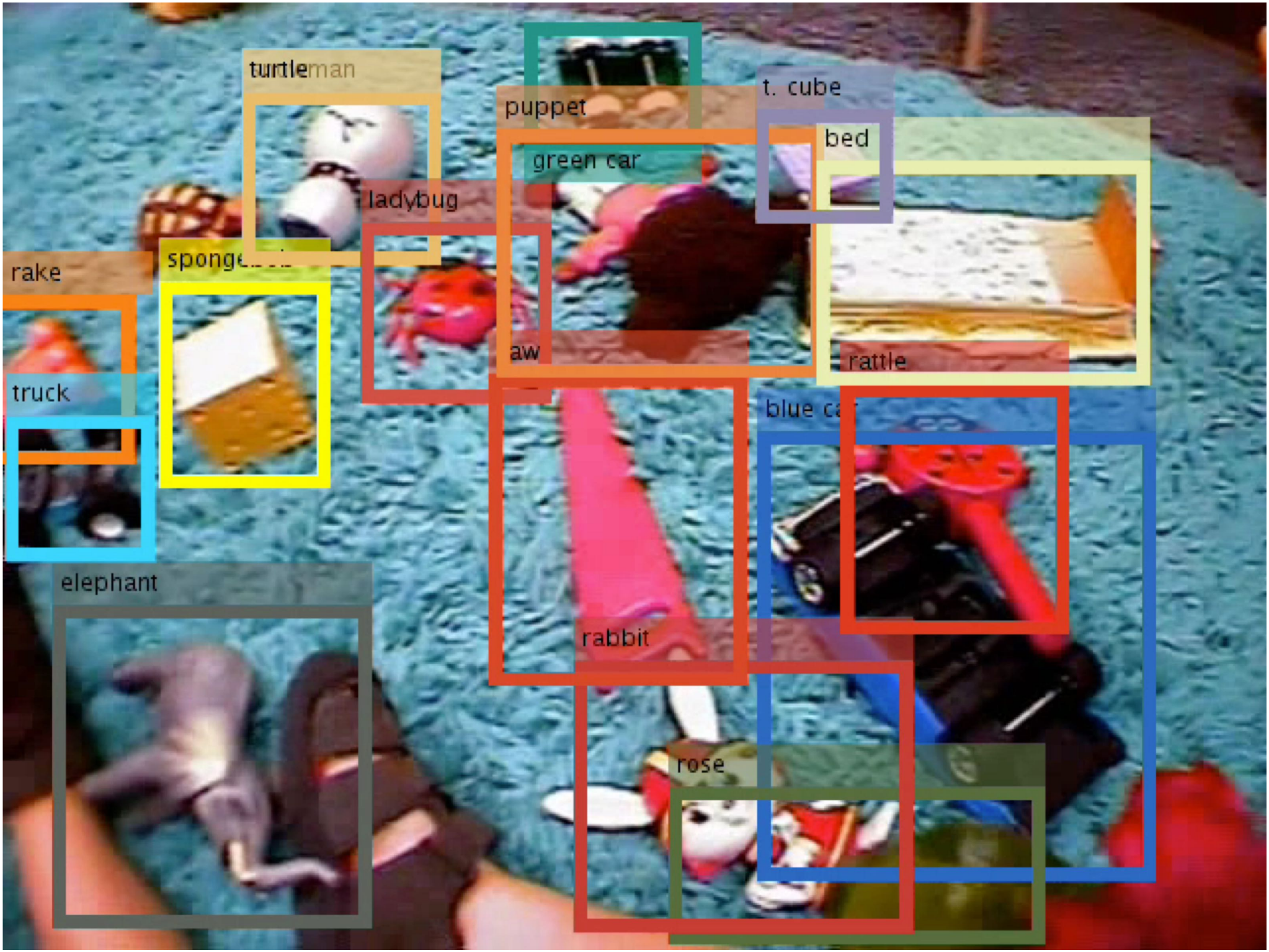
S. Bambach, D. Crandall, C. Yu. Understanding embodied visual attention in child-parent interactions, *IEEE International Conference on Development and Learning,* 2013.

Hand detection & segmentation



S. Lee, S. Bambach, C. Yu, D. Crandall. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video, *CVPR Egovision,* 2014.

S. Bambach, S. Lee, D. Crandall, C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. *ICCV*, 2015.


turtleman
puppet
t. cube
bed
green car
ladybug
rake
aw
rattle
truck
blue ca
rabbit
elephant
rose

How do children's amazing visual learning systems work?

- Probably not with CNNs
- Probably not trained on ImageNet
  - Interact with objects
  - See objects in natural contexts
  - Their bodies impose constraints

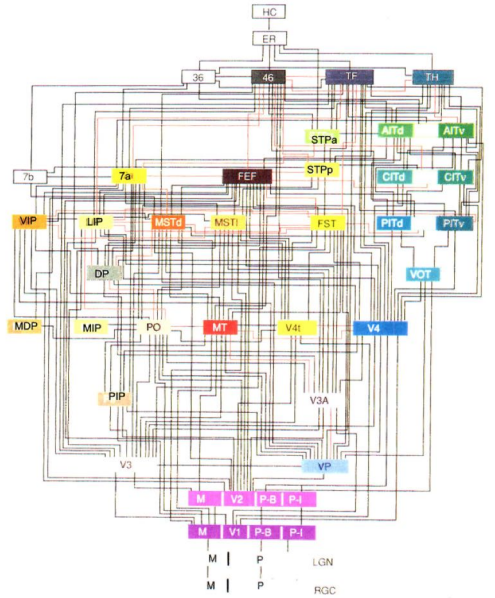
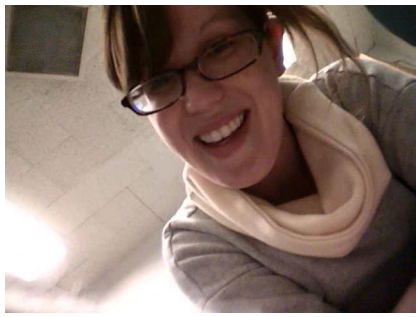


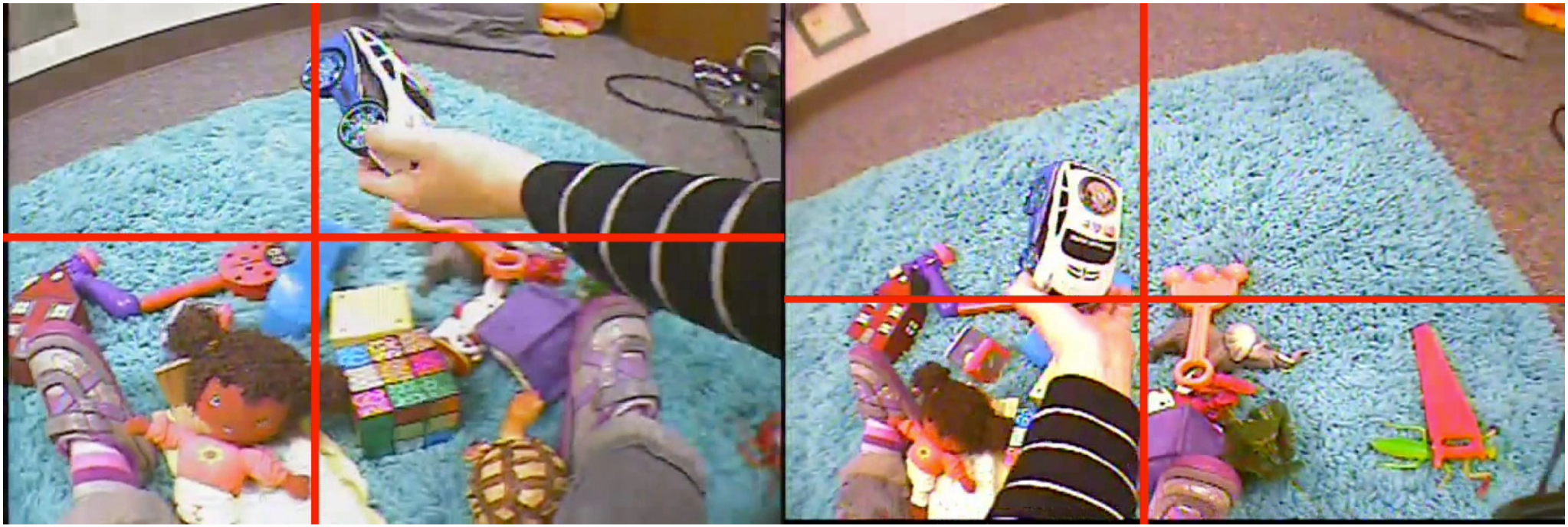
Figure 4. Hierarchy of visual areas. This hierarchy shows 32 visual cortical areas, shaded according to the same scheme as in Figure 2, 2 subcortical visual stages (the retinal ganglion cell layer and the LGN), plus several nonvisual areas (area 7b of somatosensory cortex, perirhinal area 36, the ER, and the hippocampal complex). These areas are connected by 187 linkages, most of which have been demonstrated to be reciprocal pathways.




- Can we use wearable cameras to characterize the "training data" that they receive?

The “training data” each day at 1 year old:

- 16 hours of visual experience
- 20,000 word
- Self-generated through activity (10,000 steps)

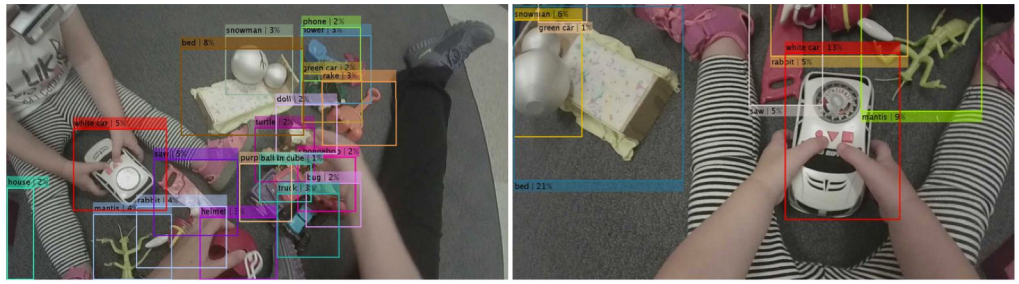


Child view

Parent view

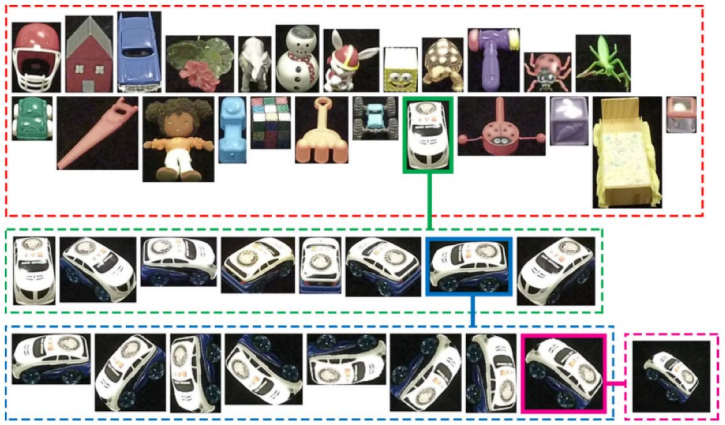
Whose view is better for object learning?

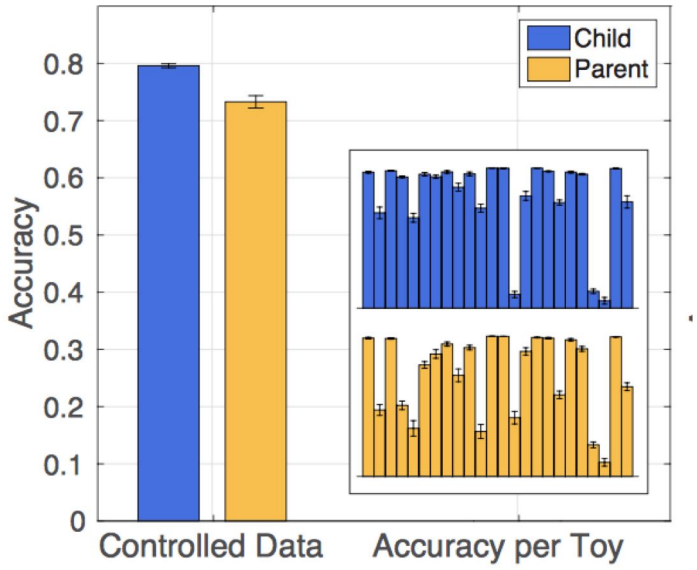
- Training data: ~100 minutes of video from 10 child-parent dyads playing with 24 toys in a free-form lab environment



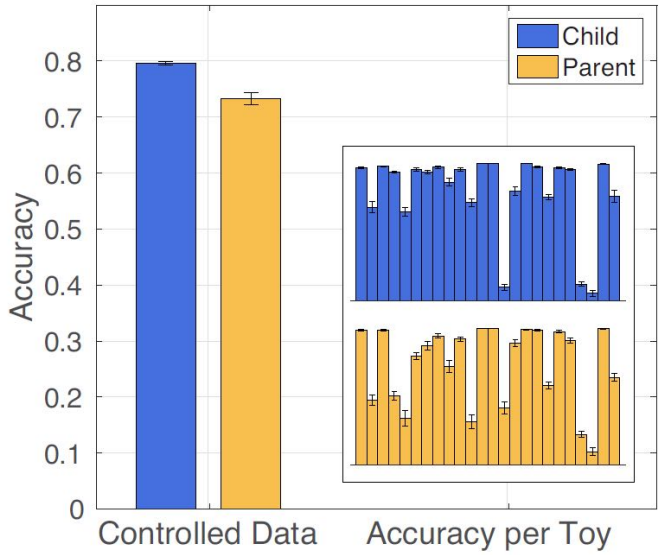
(a) Parent view (b) Child view

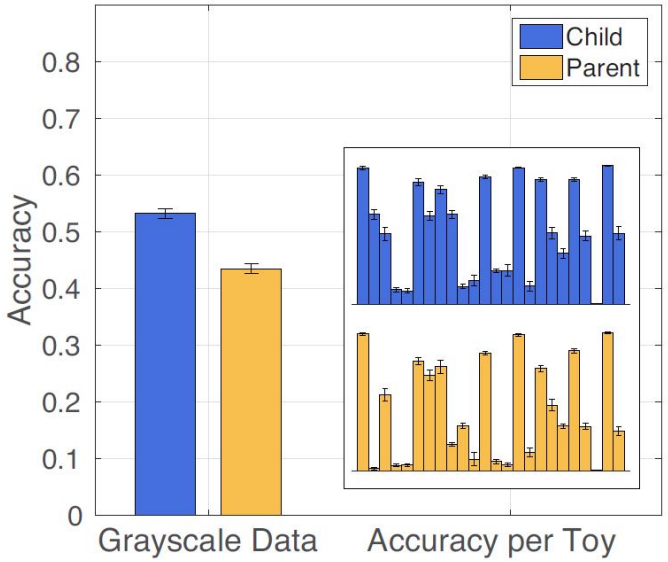
- Test data: Controlled object views on a turntable

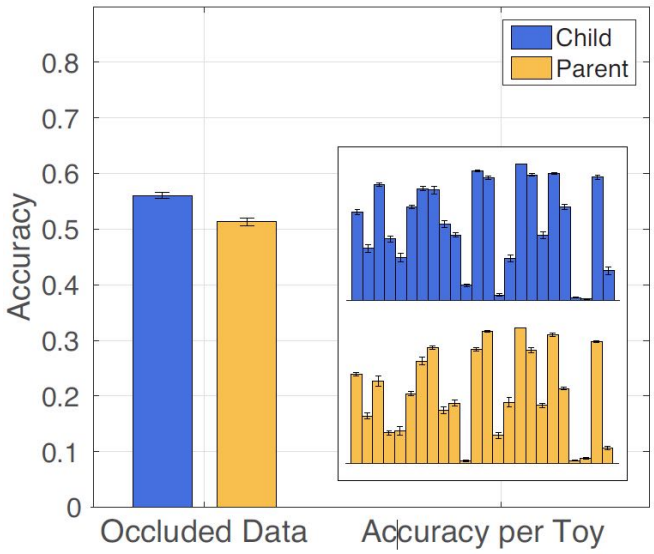


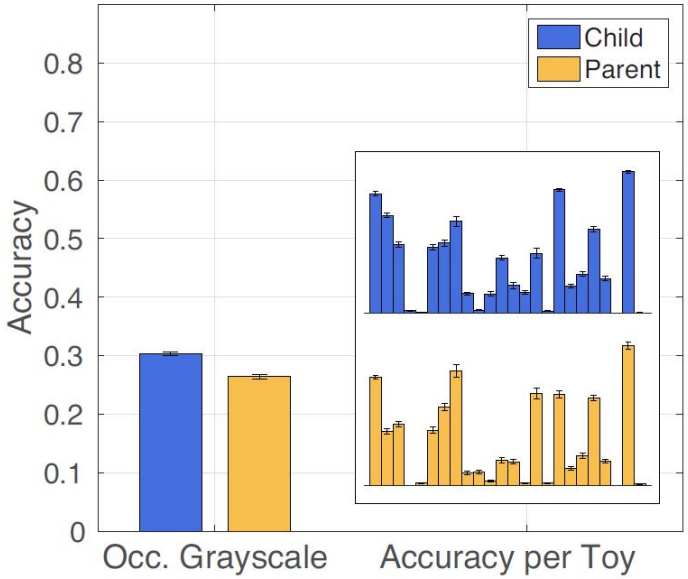


S. Bambach, D. Crandall, L. Smith, C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach, *CogSci,* 2016.



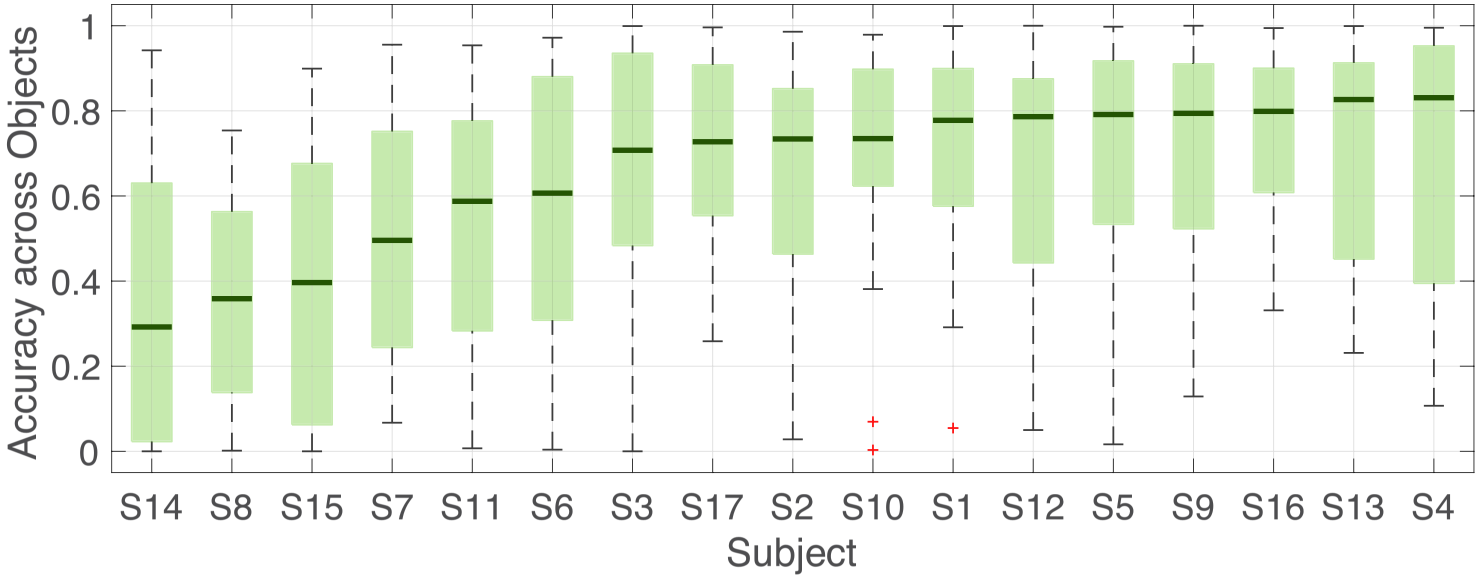




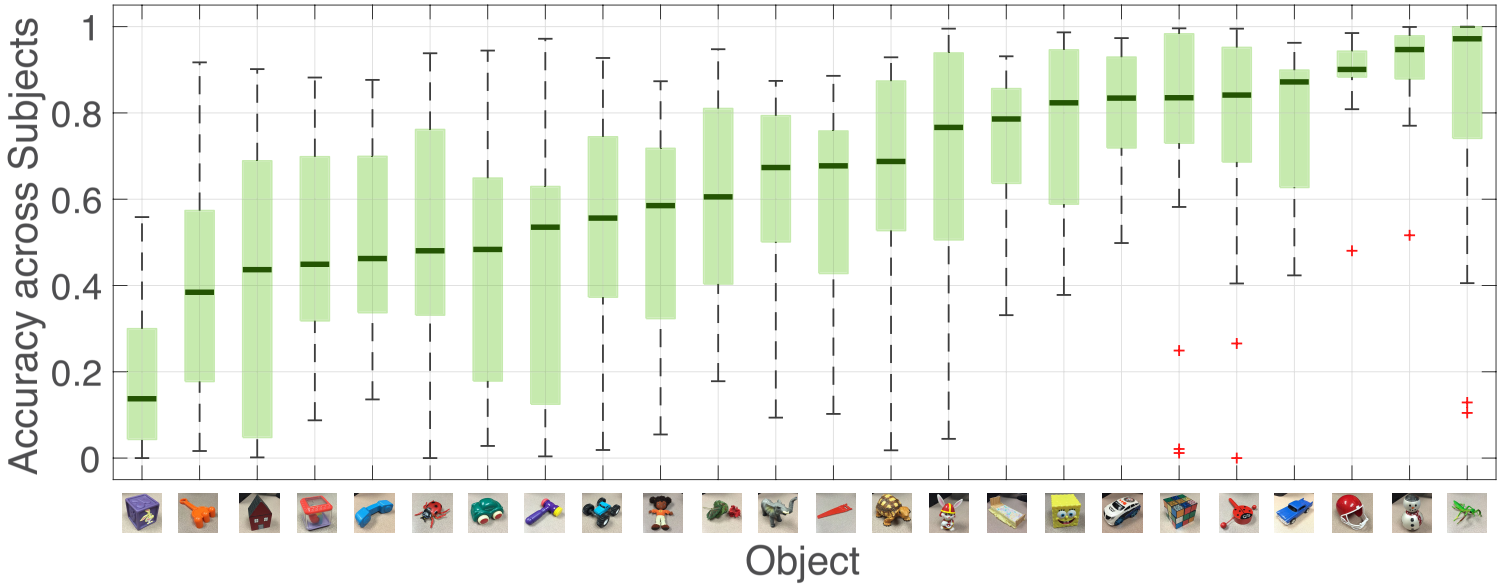


- CNNs can learn models of the toys in this first-person data;
- They generalize to different contexts and viewpoints;
- Toddler views seem to be of particularly high value for training.

Each Subject's Recognition Accuracies Across Objects



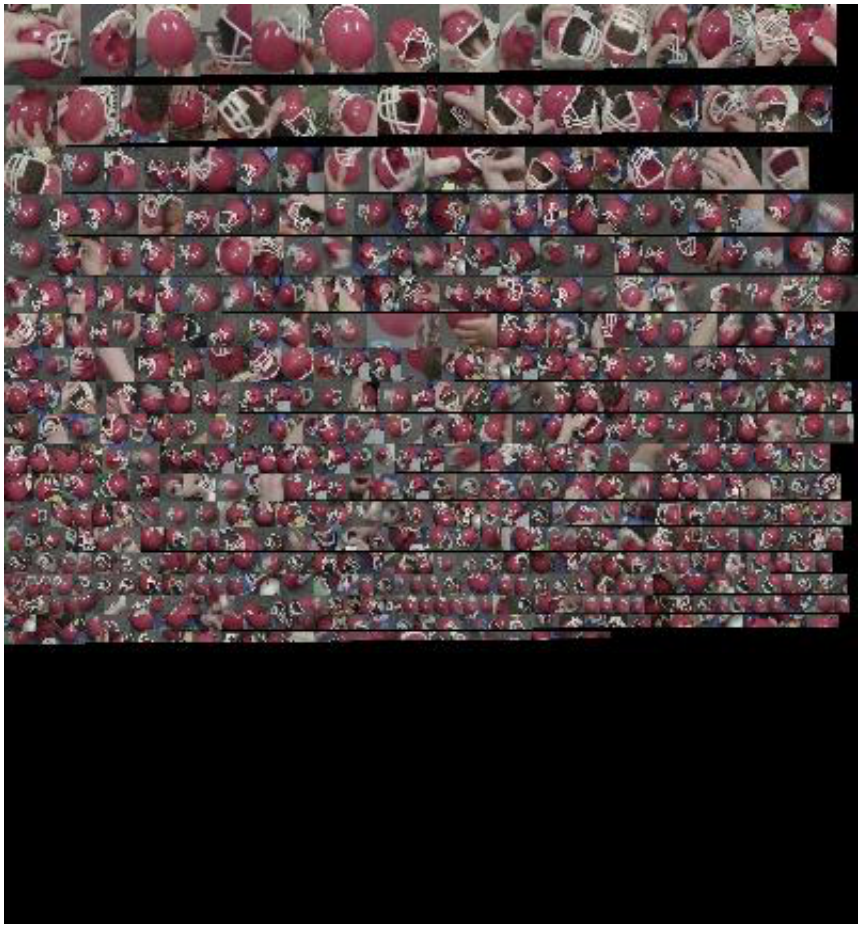
Each Object's Recognition Accuracies Across Subjects



S. Bambach, Z. Zhang, D. Crandall, C. Yu. Exploring inter-observer differences in first-person object views using deep learning models, *ICCV Workshop on Mutual Benefits of Cognitive and Computer Vision,* 2017.

Different views of the world

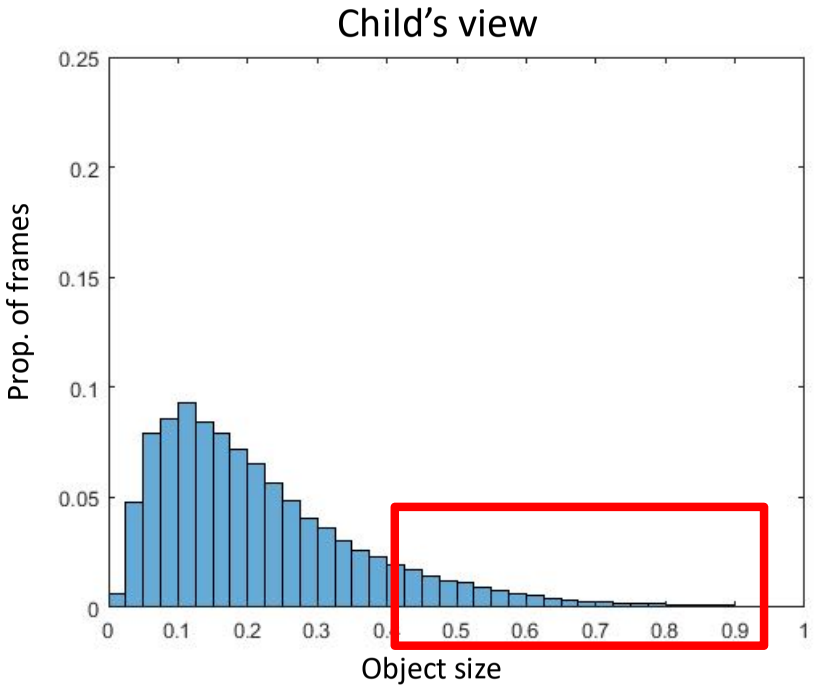
child's view

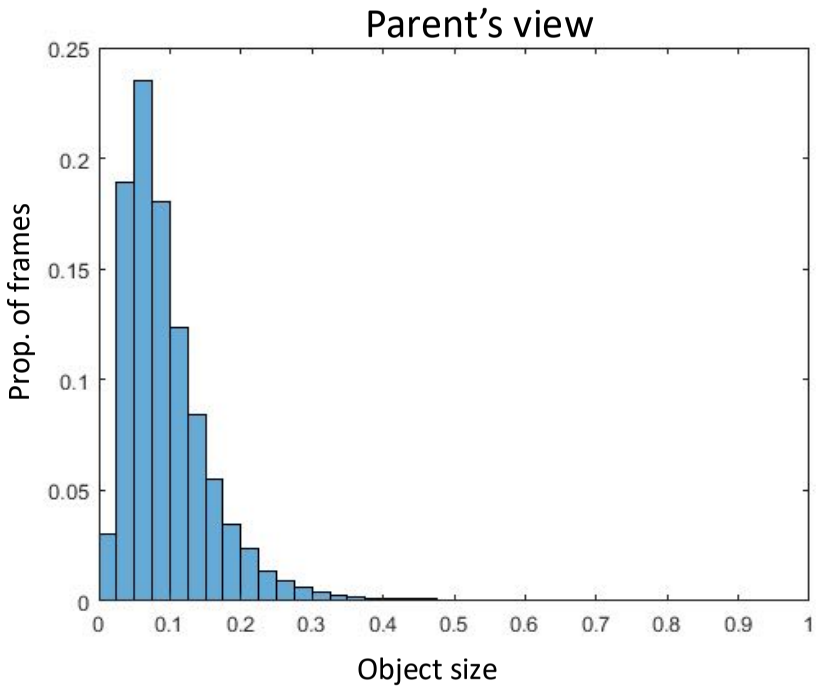


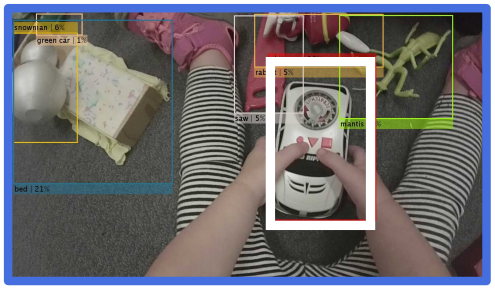
parent's view

S. Bambach, D. Crandall, L. Smith, C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach, *CogSci,* 2016.

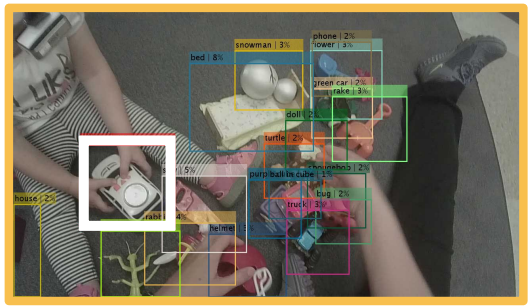
Object Size







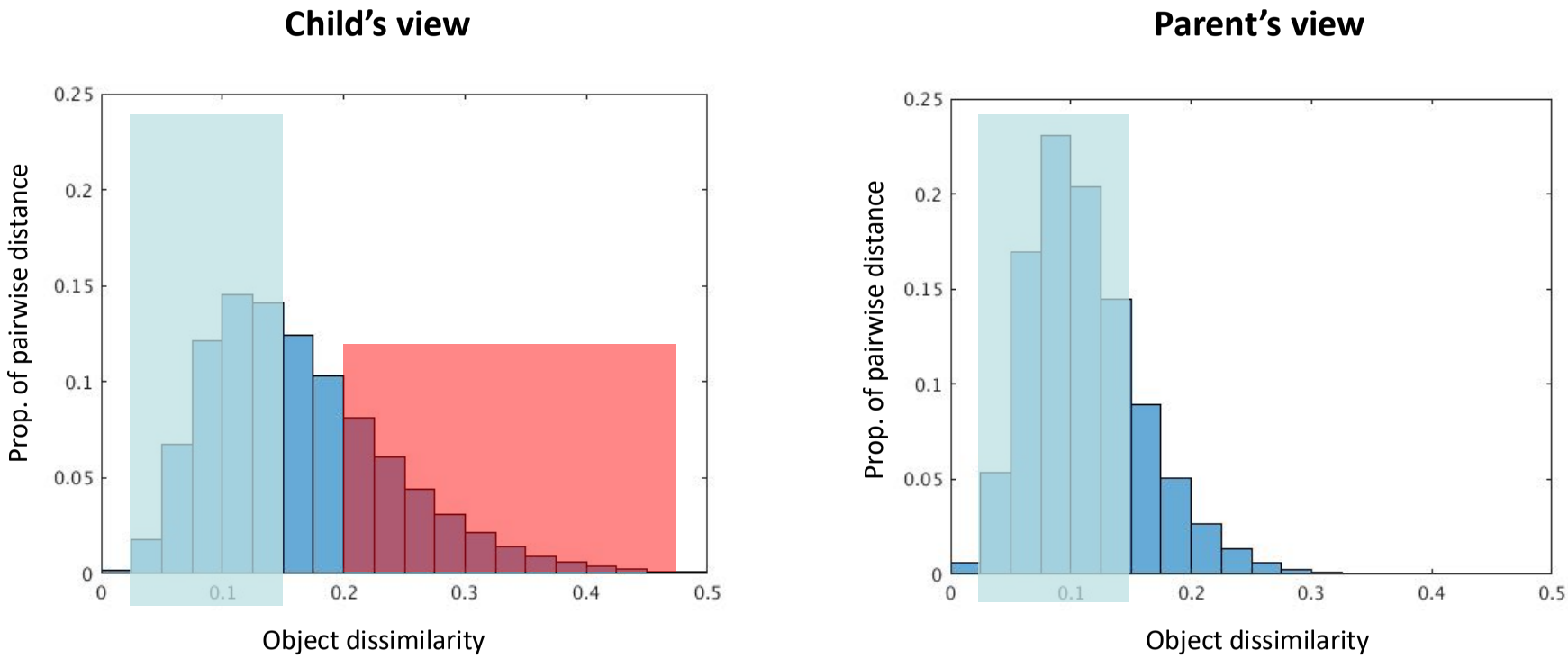
13%



5%

S. Bambach, D. Crandall, L. Smith, C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach, *CogSci,* 2016.

Diversity of object views

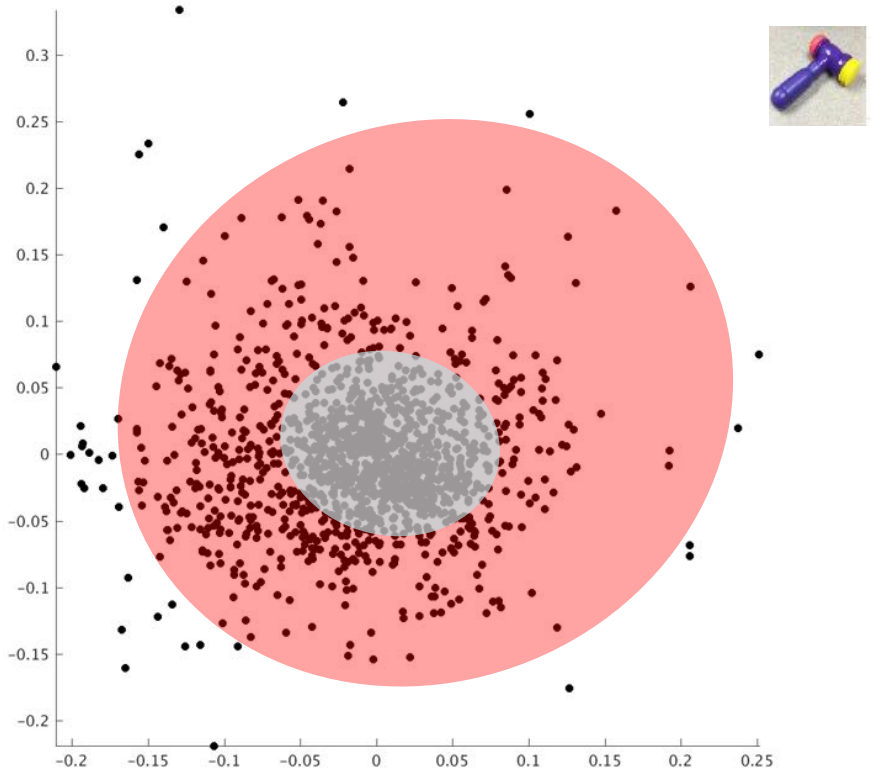


Toddlers spend hours every day playing with toys, actively manipulating them and creating "training data" by self-selecting object views. This creates more diverse views for children than the parents.

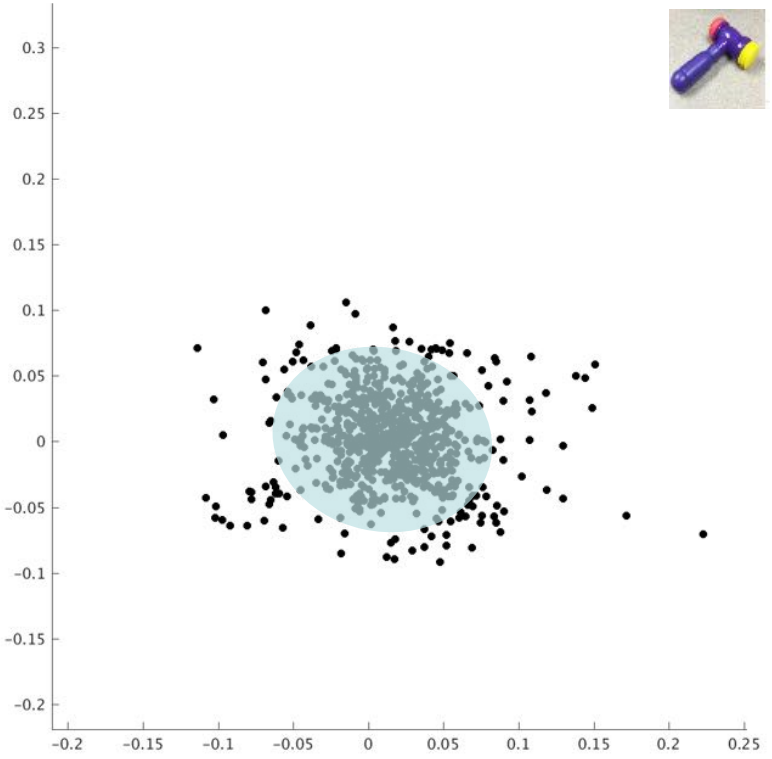
S. Bambach, D. Crandall, L. Smith, C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach, *CogSci,* 2016.

Consistency and Variability

Child's view

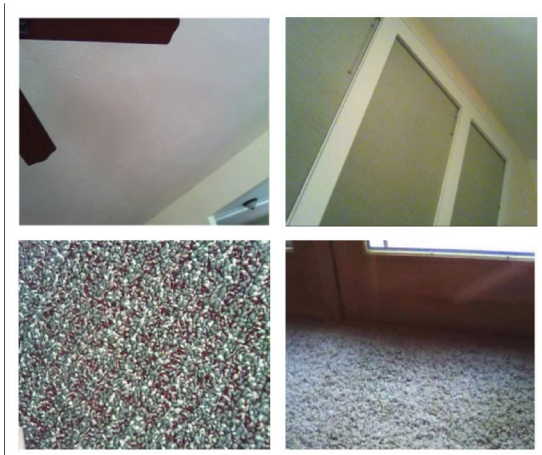


Parent's view



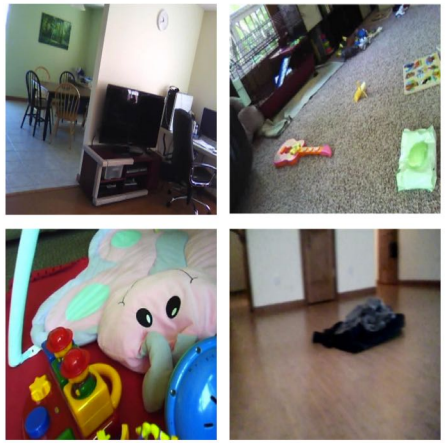
S. Bambach, D. Crandall, L. Smith, C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach, *CogSci,* 2016.

This “training data” changes over time

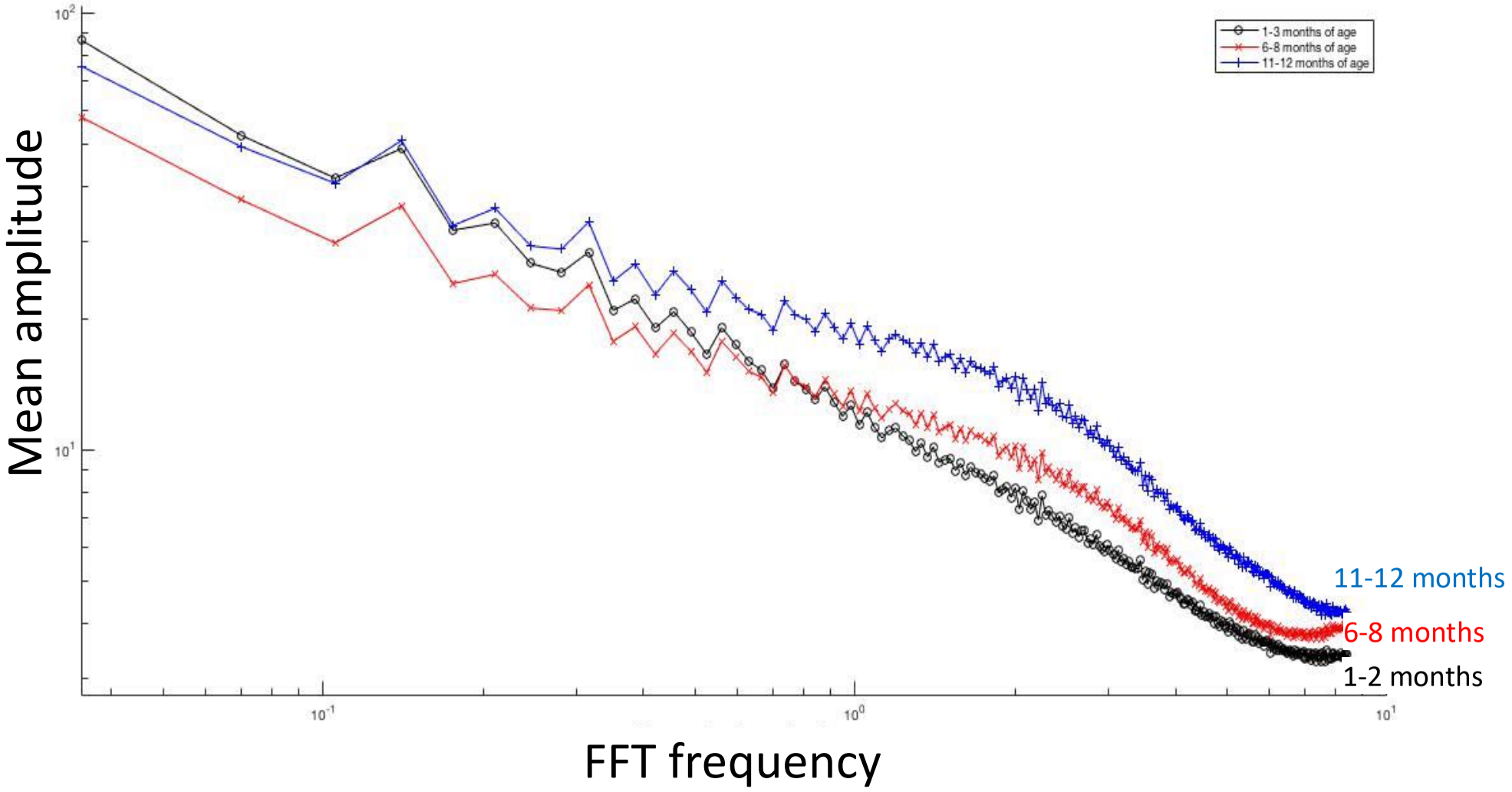


HomeView: dataset of developmentally indexed head-mounted camera data

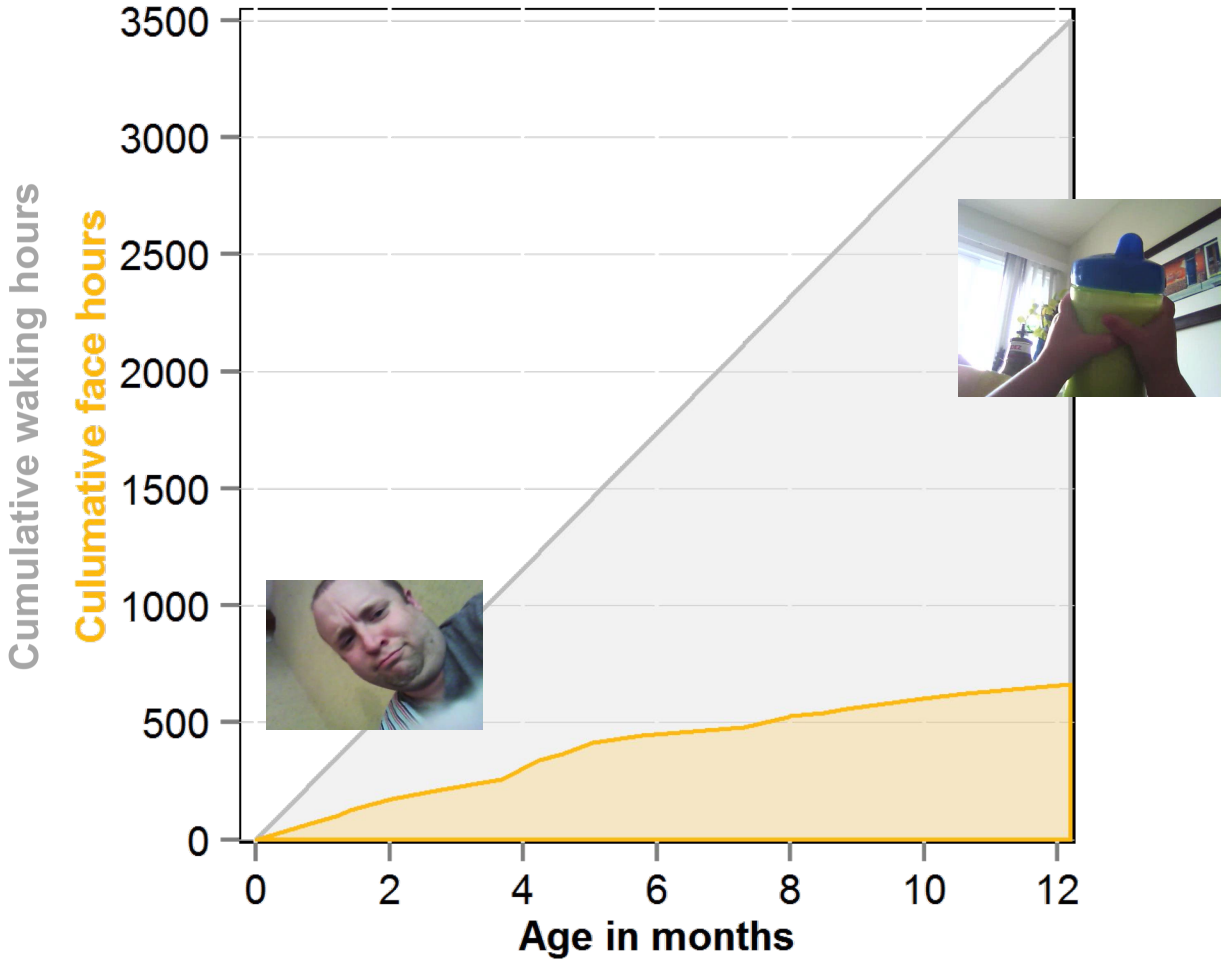
- 101 infants, 3 weeks – 24 months
- 4 to 6 hours of in-home video
- No experimenters present
- 30 Hz, 750 million frames



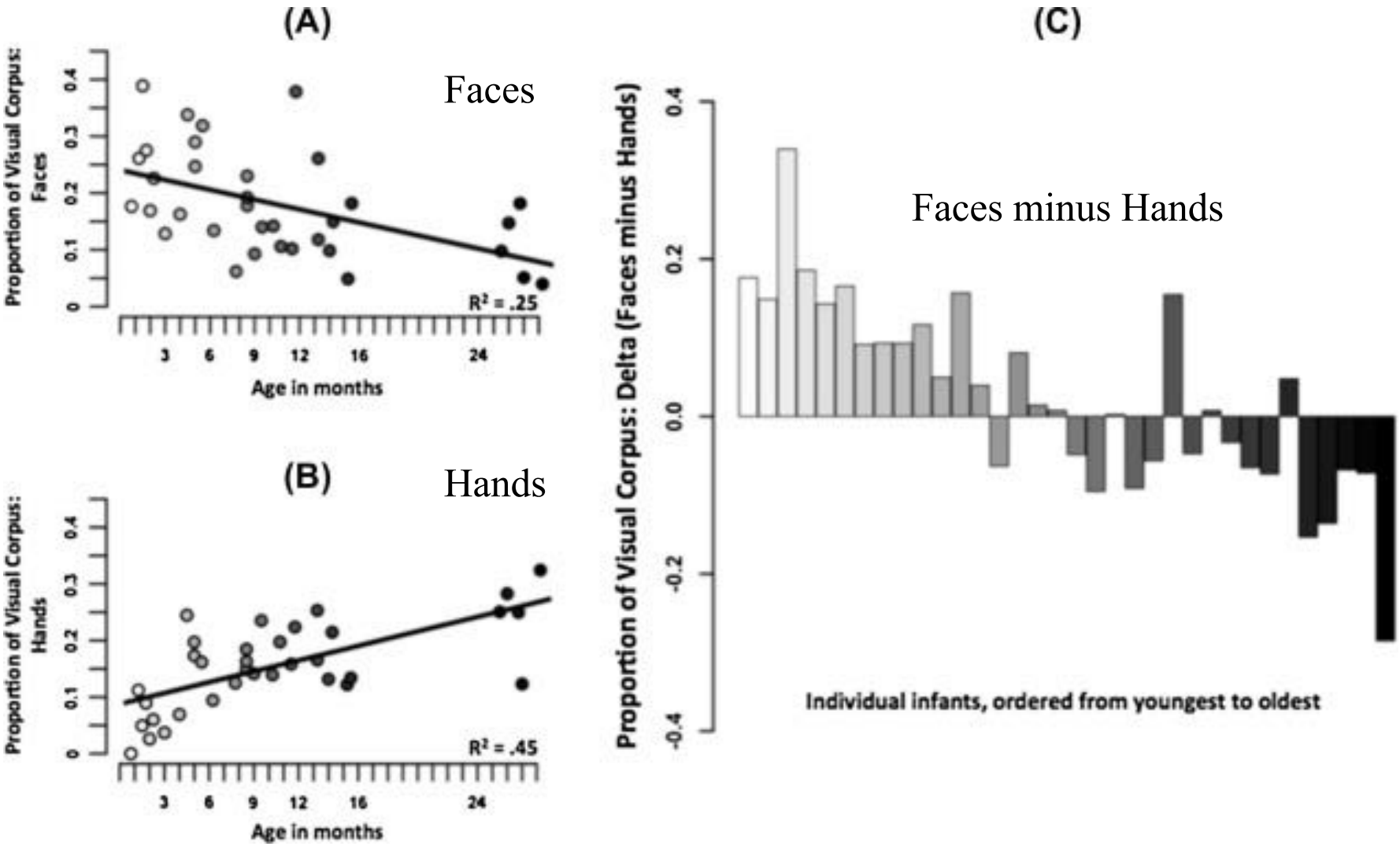
Low-level image stats change over time



Frequency of objects in view also change



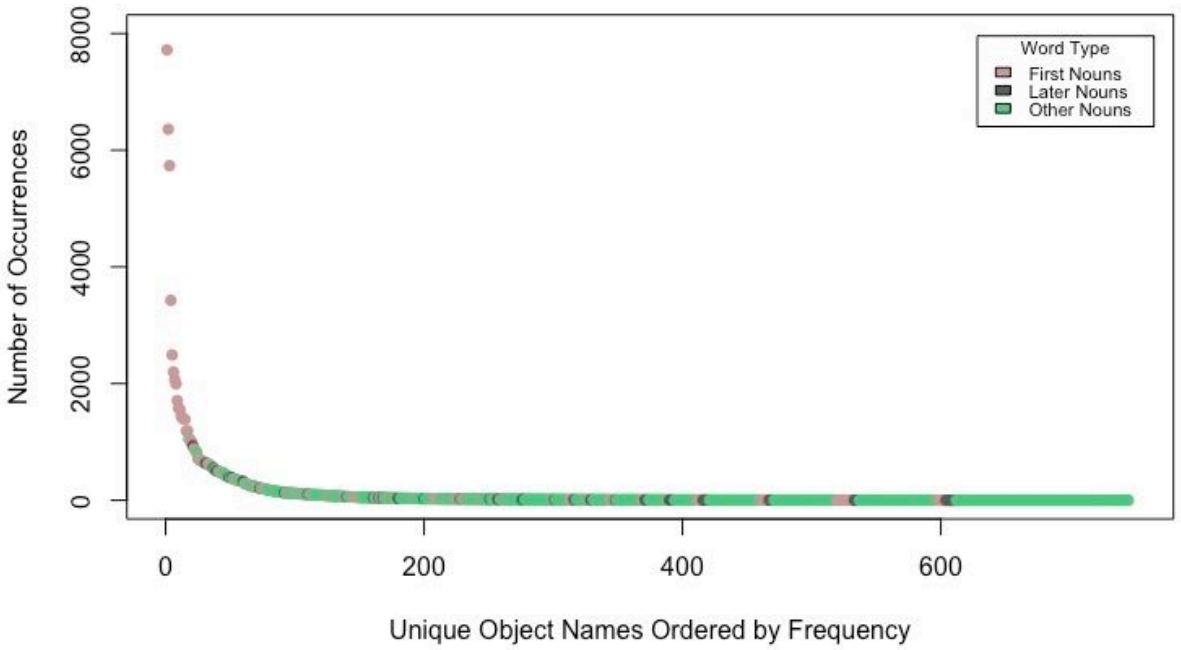
Jayaraman, Fausey & Smith (2015) *PLoS One*



Fausey, Jayaraman & Smith, (2016) *Cognition*

Word learning

- Labeled instances of 745 objects in HomeView data
- **Key finding:** The visual frequency of objects in scenes at 8-10 months – not the frequency of their names -- predicts age of acquisition of object names.



- This suggests early visual word learning is largely "unsupervised;" around 1 year transitions to "supervised"

Clerkin, et al, Real-world visual statistics and infants' first-learned object names, *Phil Trans Royal Society B*, 2016.

Short term questions:

How can automated computer vision help analyze egocentric images?

Longer-term question:

Can we model head and hand pose, eye gaze, saliency, and activity to predict and explain human vision?

Long-term question:

What can machine learning (and deep learning in particular) reveal about human visual learning, and vice versa?

IU Bloomington awards up to $3 million to advance new approach to study of learning

Campus's Emerging Areas of Research program makes inaugural grant

Jan. 11, 2017
FOR IMMEDIATE RELEASE

BLOOMINGTON, Ind. -- Can machines learn to think like children? An interdisciplinary team of cognitive scientists, neuroscientists and computer scientists at Indiana University Bloomington has received the campus's inaugural Emerging Areas of Research funding award to explore that question.

Called "Learning: Brains, Machines and Children," the first Emerging Areas of Research initiative is led by Linda Smith, Distinguished Professor and Chancellor's Professor of psychological and brain sciences in the IU Bloomington College of Arts and Sciences. The research team will receive up to $3 million for the four-year project.

The Emerging Areas of Research program was proposed in the Bicentennial Strategic Plan for IU Bloomington as part of a suite of initiatives designed to invest in research innovation and excellence on campus.

"The Emerging Areas of Research funding program ensures that our campus is constantly thinking about the next frontier in knowledge and hiring new faculty to ensure we can reach that frontier," IU Bloomington Provost and Executive Vice President Lauren Robel said. "The stellar team led by Professor Smith epitomizes the strengths that allow us to build toward that future."

Representing expertise in developmental psychology, human learning, neuroscience and artificial intelligence, the inaugural Emerging Areas of Research team will pursue a comprehensive and unified theory of learning rooted in research about how infants and children learn to classify faces, objects, letters, numbers, etc.

"We are thrilled that IU Bloomington's first Emerging Areas of Research initiative represents such a profound area of focus and such an impressive group of researchers," said Rick Van Kooten, the vice provost for research at IU Bloomington, whose office oversees the Emerging Areas of Research program. "The proposal leverages exemplary areas of research on our campus, and the combination is certain to yield exciting results."

PRINT SHARE

Linda B. Smith | Photo by Indiana University

Print-Quality Photo

Media Contacts

Lauren Bryant

Office of the Vice Provost for Research
Office 812-855-4152
labryant@indiana.edu

Related Stories

New Emerging Areas of Research funding program launches at IU Bloomington

Workshop on

Egocentric Vision: From Science to Real-World Applications

Indiana University, Bloomington Indiana — June 3-5, 2017

Sponsored by the Indiana University Ostrom Grants Program, the College of Arts and Science, the School of Informatics and Computing, and the OVPR Emerging Areas of Research program.

http://vision.soic.indiana.edu/egocentric-workshop-2017

Organizers

David Crandall
School of Informatics and Computing
Indiana University

Michael Ryoo
School of Informatics and Computing
Indiana University

Linda Smith
Psychological and Brain Sciences
Indiana University

Chen Yu
Psychological and Brain Sciences
Indiana University

Speakers and panelists

Karen Adolph
Psychology
NYU

Dana Ballard
Computer Science
UT Austin

Derek Houston
Otolaryngology-Head and Neck Surgery
The Ohio State University

Michael Gaffrey
Psychiatry
Washington University in St. Louis

Mary Hayhoe
Center for Perceptual Systems
UT Austin

Dan Kennedy
Psychological and Brain Sciences
Indiana University

Maithilee Kunda
Electrical Engineering and Computer Science
Vanderbilt University

Hyun Soo Park
Computer Science and Engineering
University of Minnesota

Ruslan Salakhutdinov
Machine Learning
Carnegie Mellon University / Apple

Jim Rehg
School of Interactive Computing
Georgia Tech

Alexander Schwing
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

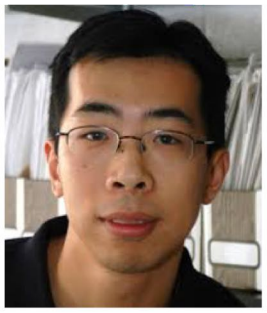
Talks and slides online!

Thanks!

http://vision.soic.indiana.edu/

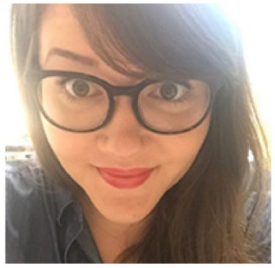
David Crandall

Linda Smith

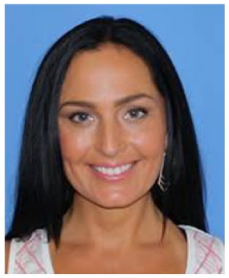


Chen Yu

Sven Bambach



Elizabeth Clerkin



Christina DeSerio

Swapnaa Jayaraman

Zehua Zhang

Sponsors: NSF, NIH, Google, Nvidia, IU EAR, IU FRSP, Lilly Endowment