



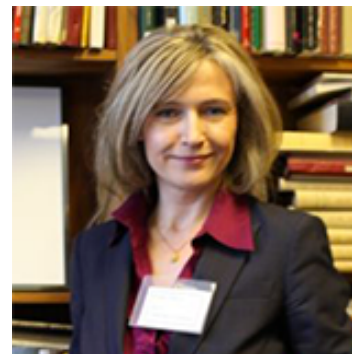
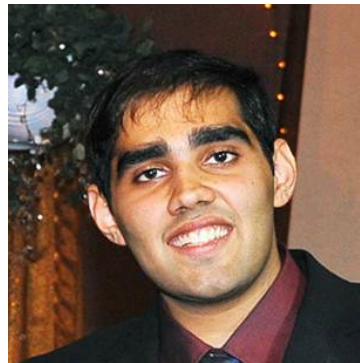
Network Dissection: Quantifying Interpretability of Deep Visual Representations

Bolei Zhou*

MIT

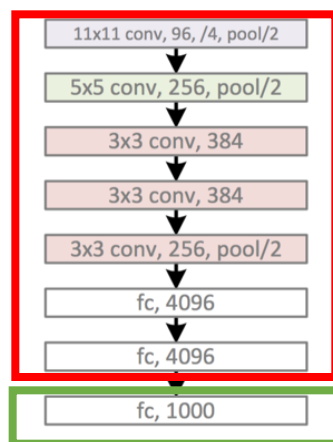
Joint work with David Bau*, Aditya Khosla, Aude Oliva, Antonio Torralba

* Indicates equal contribution



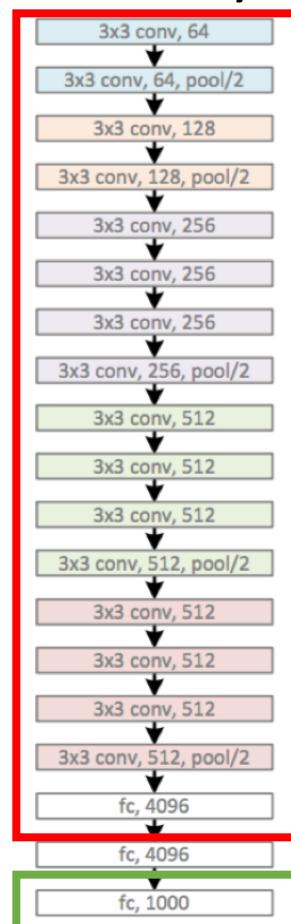
Deep ConvNet for Visual Recognition

2012: AlexNet
5 conv. layers



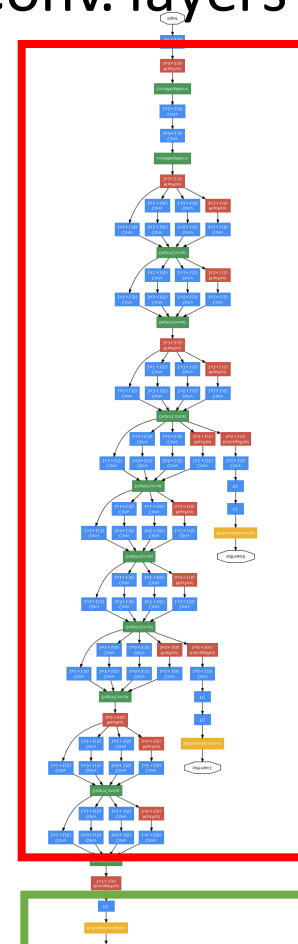
Error: 15.3%

2014: VGG
16 conv. layers



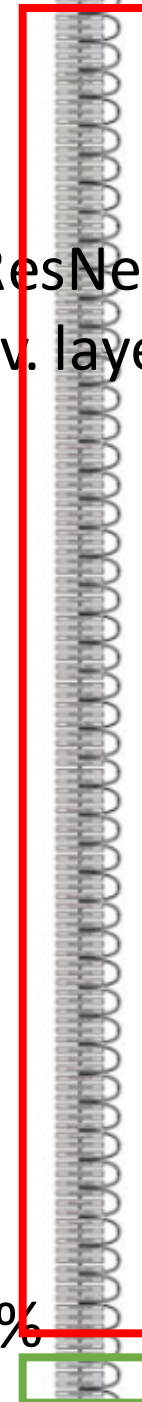
Error: 8.5%

2015: GoogLeNet
22 conv. layers



Error: 7.8%

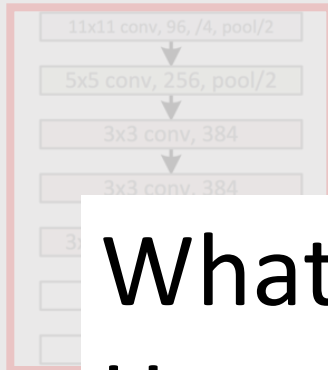
2016: ResNet
>100 conv. layers



Error: 4.4%

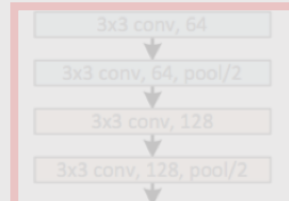
Deep ConvNet for Visual Recognition

2012: AlexNet
5 conv. layers



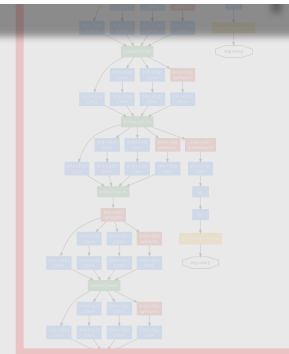
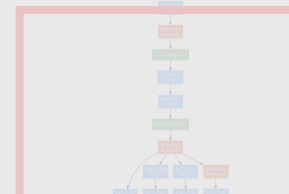
Error: 15.3%

2014: VGG
16 conv. layers



Error: 8.5%

2015: GoogLeNet
22 conv. layers



Error: 7.8%

2016: ResNet
>100 conv. layers

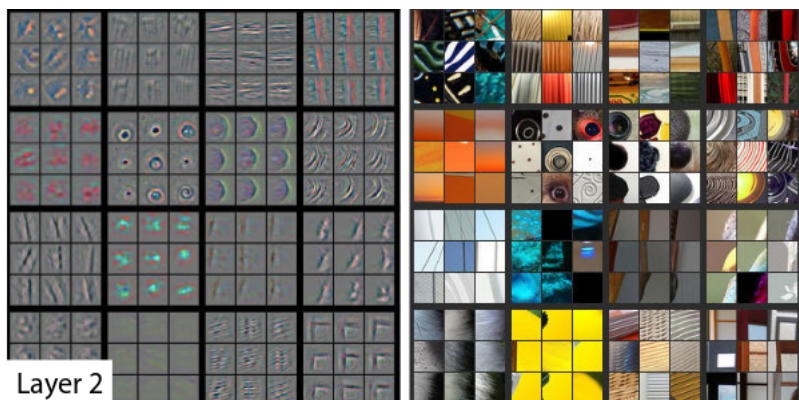
Error: 4.4%

What have been learned inside?

How to compare the internal representations?

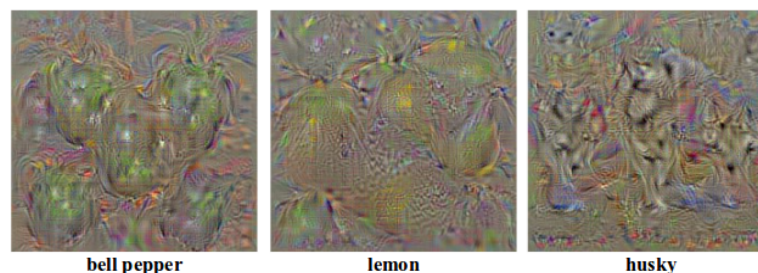
Previous Work on Network Visualization

Deconvolution

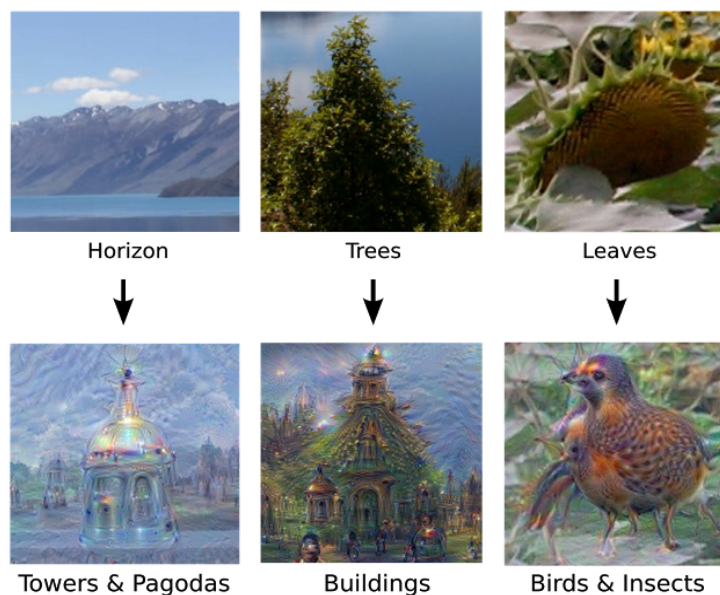


Zeiler et al., ECCV 2014.

Back-propagation

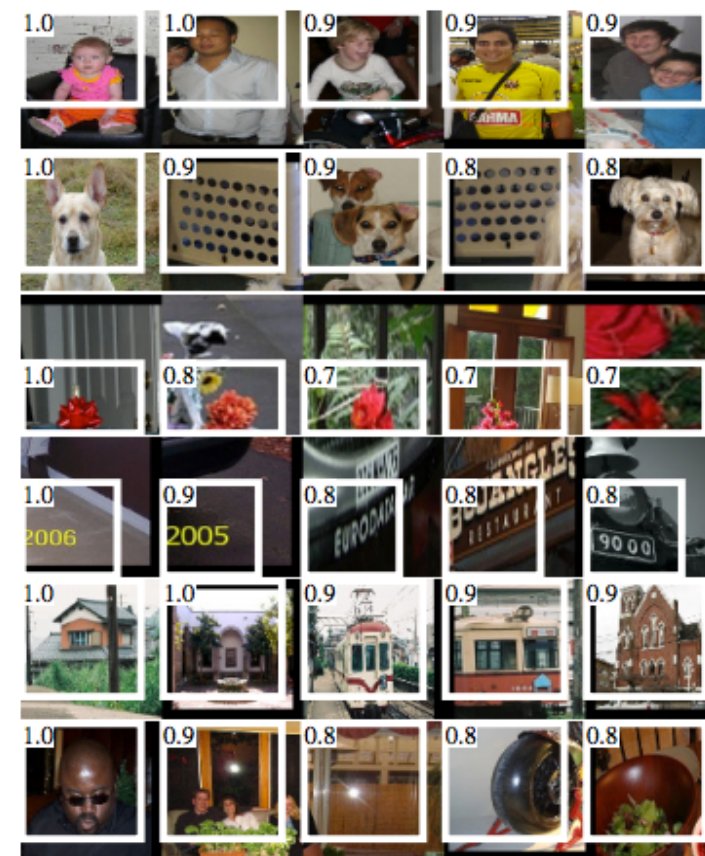


Simonyan et al., ICLR 2015



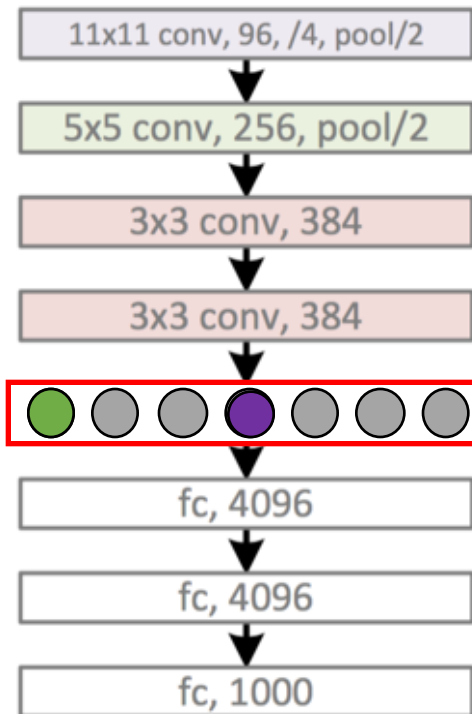
Inceptionism. Google Blog. June 2015

Top activated images



Girshick et al., CVPR 2014

Goal: From Visualization to Interpretation



Unit 1



Top Activated Images



Interpretation: lamp

Score: 0.15

Top Activated Images



Interpretation: car

Score: 0.02

Approach: Test units for semantic segmentation

Unit 1

Top activated images



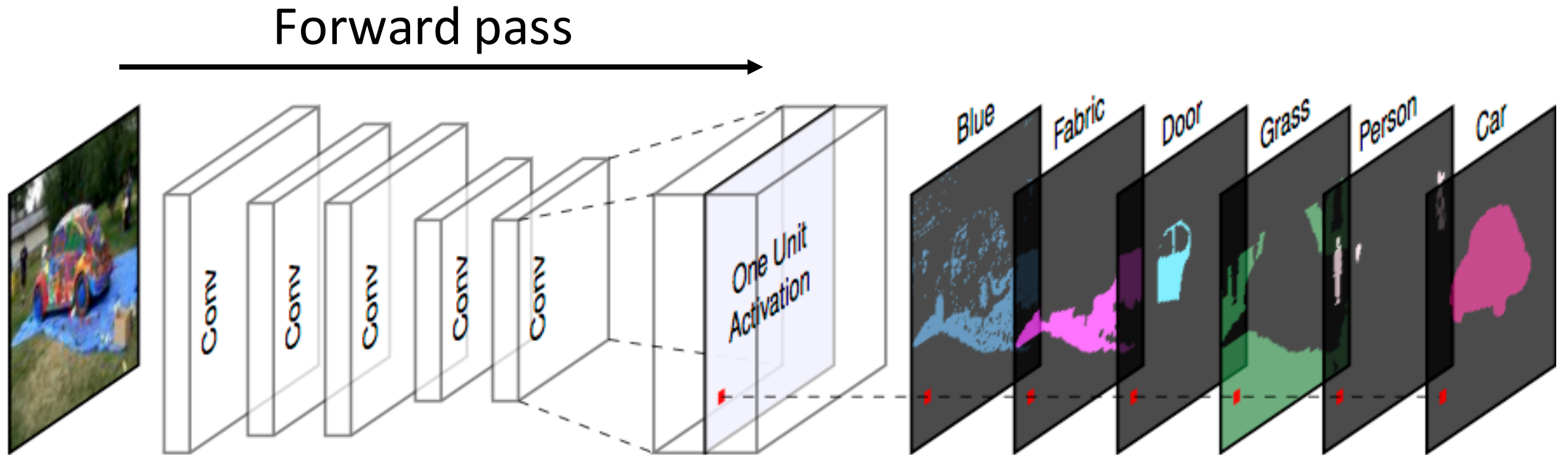
Lamp

Intersection over Union (IoU)= 0.12



Network Dissection

Quantifying the interpretability of units through segmentation



Broadly and Densely (**Broden**) Annotated Dataset

ADE20K

Zhou et al, CVPR'17

Pascal Context

Mottaghi et al, CVPR'14

Pascal Part

Chen et al, CVPR'14

Open-Surfaces

Bell et al, SIGGRAPH'14

Describable Textures

Cimpoi et al, CVPR'14

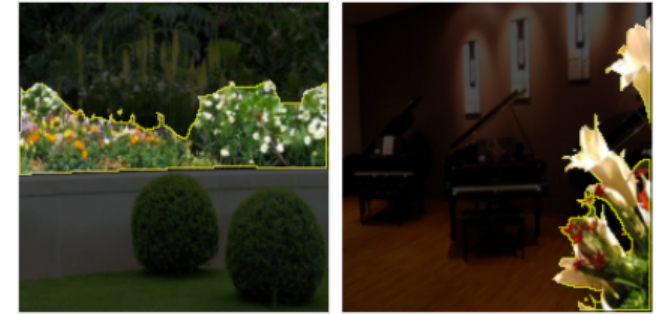
Colors

Total = **63,305** images
1,197 visual concepts

street (scene)



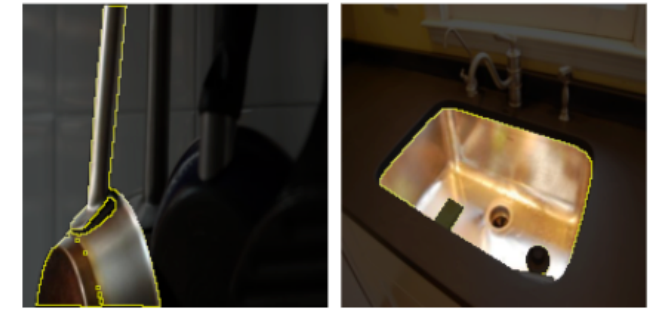
flower (object)



headboard (part)



metal (material)



swirly (texture)



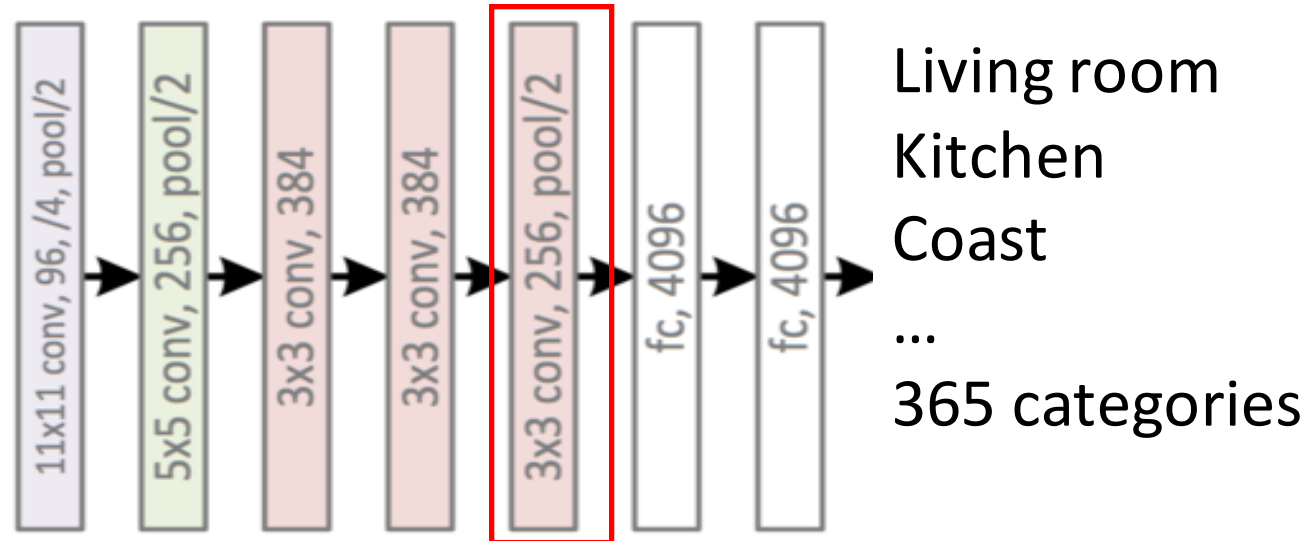
pink (color)



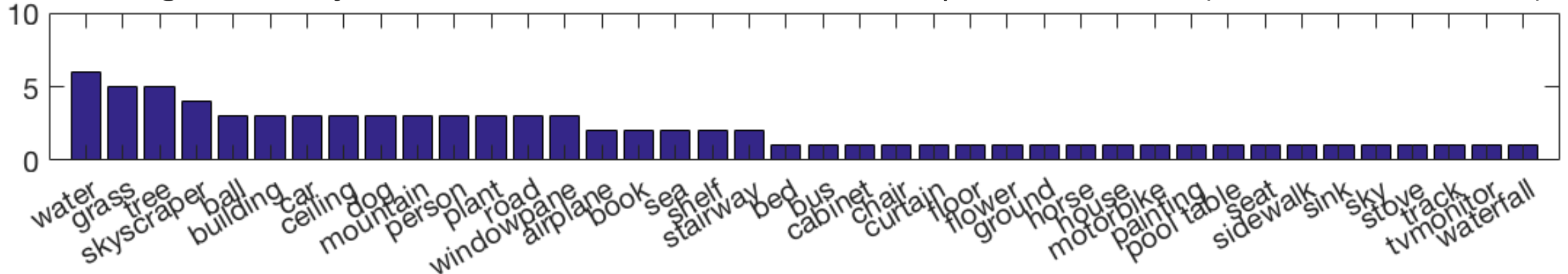
AlexNet trained on places

THE SCENE RECOGNITION DATABASE

conv5, 256 units



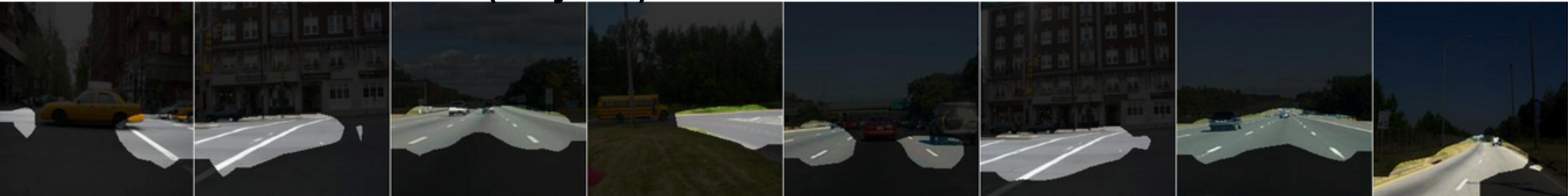
Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)



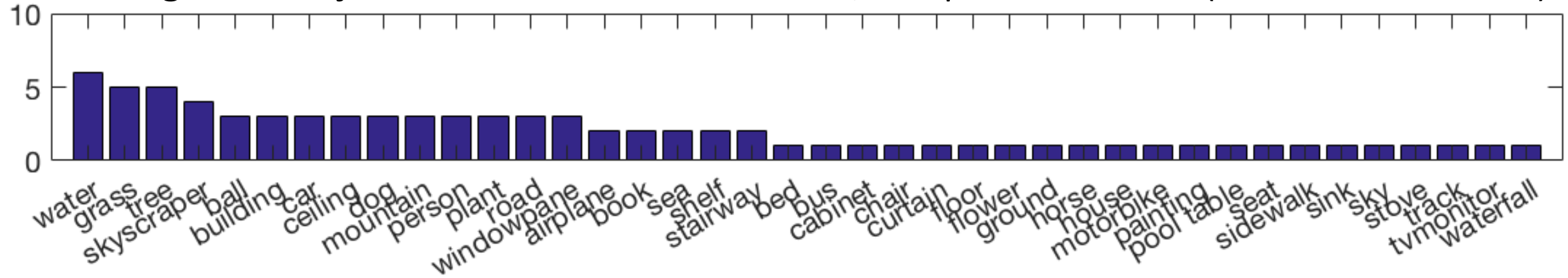
conv5 unit 79 car (object) IoU=0.13



conv5 unit 107 road (object) IoU=0.15



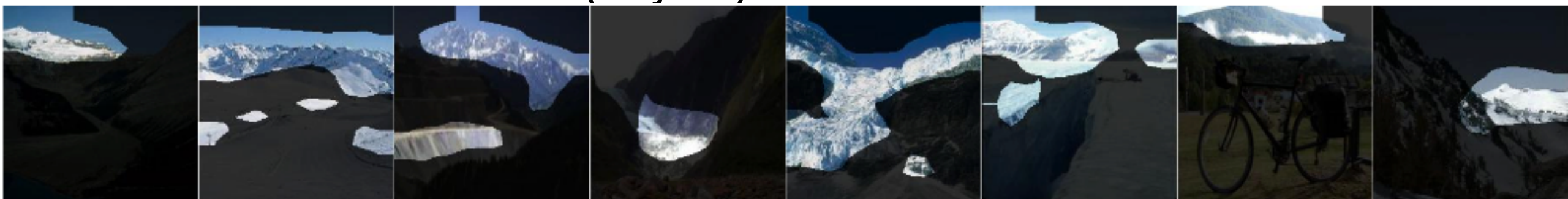
Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)



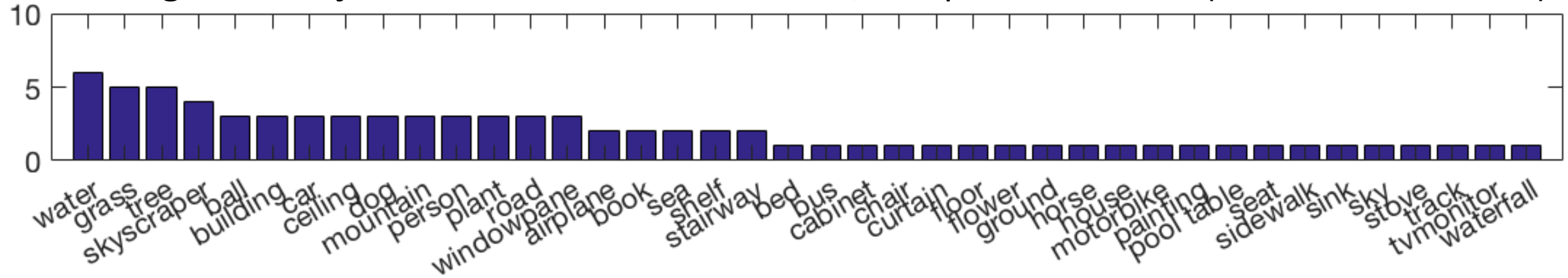
conv5 unit 144 mountain (object) IoU=0.13



conv5 unit 200 mountain (object) IoU=0.11

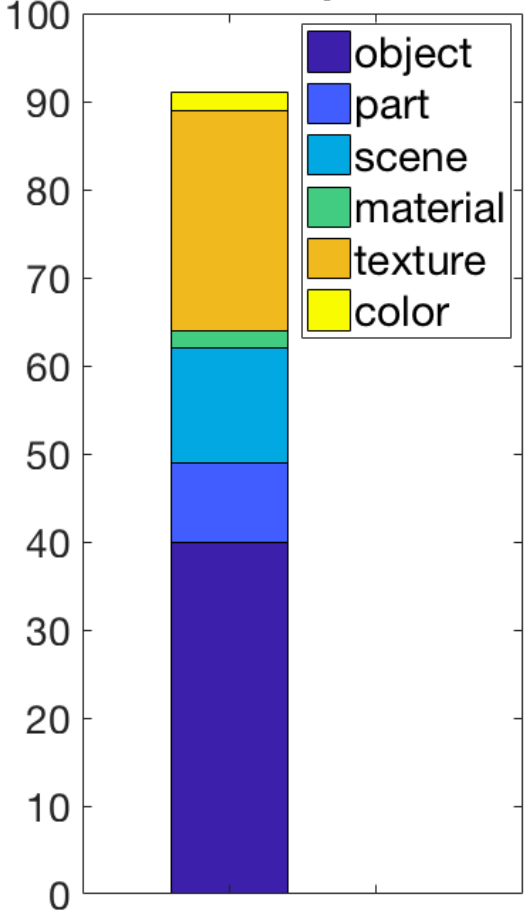


Histogram of object detectors: Detector:81/256, Unique Detector:40 (Units with IoU>0.04)

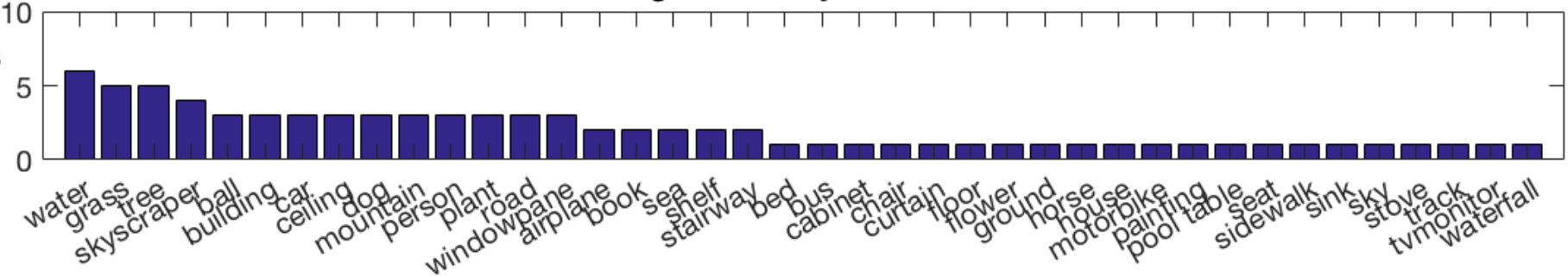


Dissection Report

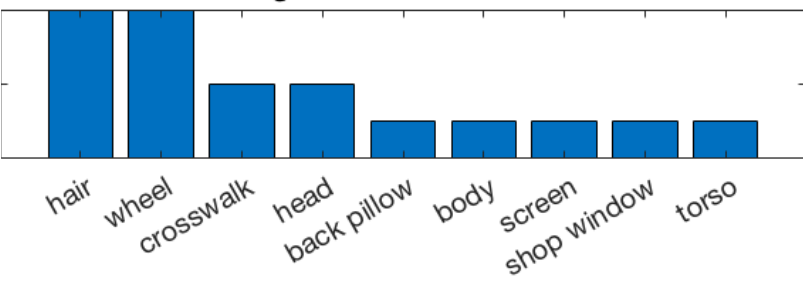
Number of Unique Detectors



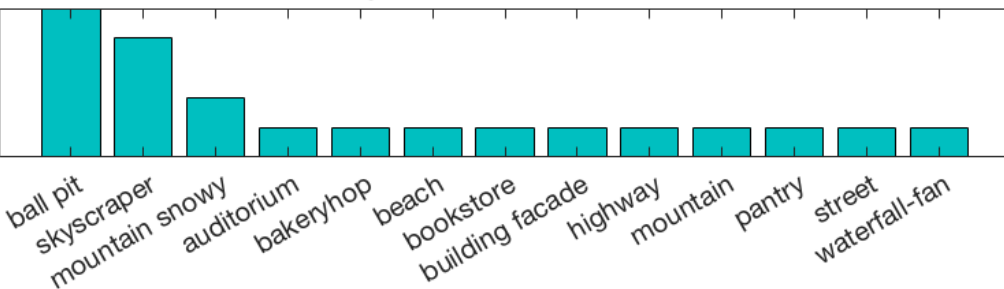
Histogram of Object Detectors



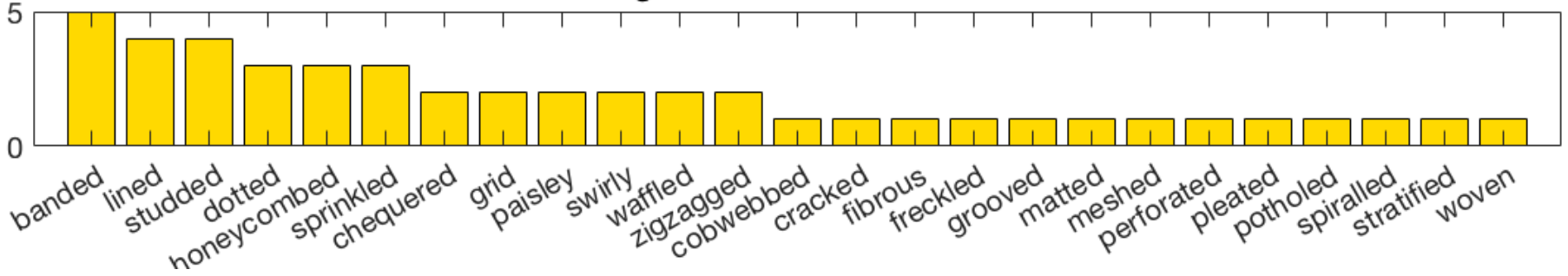
Histogram of Part Detectors



Histogram of Scene Detectors

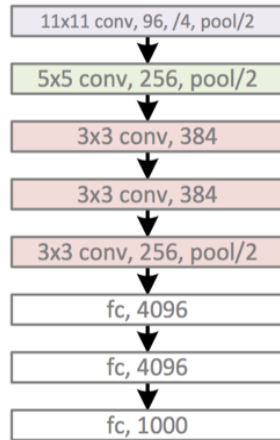


Histogram of Texture Detectors

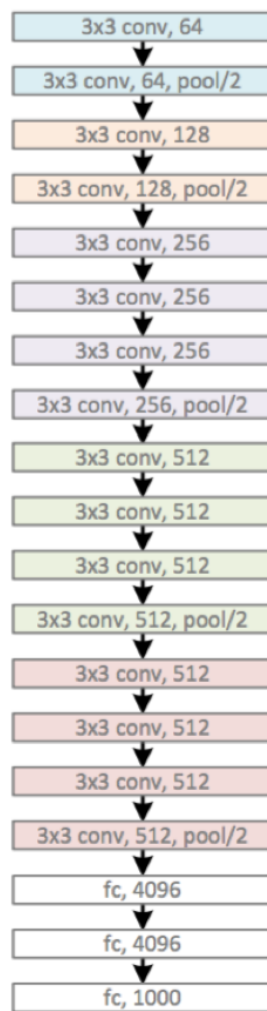


Architectures

AlexNet



VGG



GoogLeNet

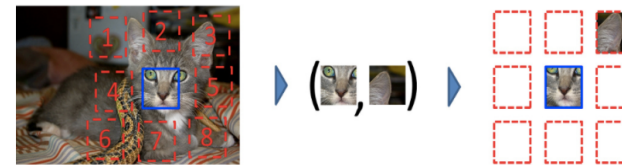


ResNet

Supervised Learning



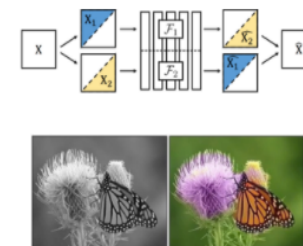
Self-Supervised Learning



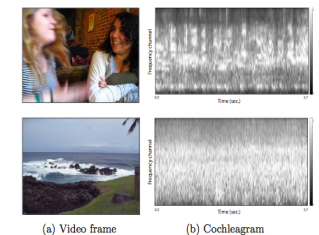
Context prediction, ICCV'15



Solving puzzle, ECCV'16

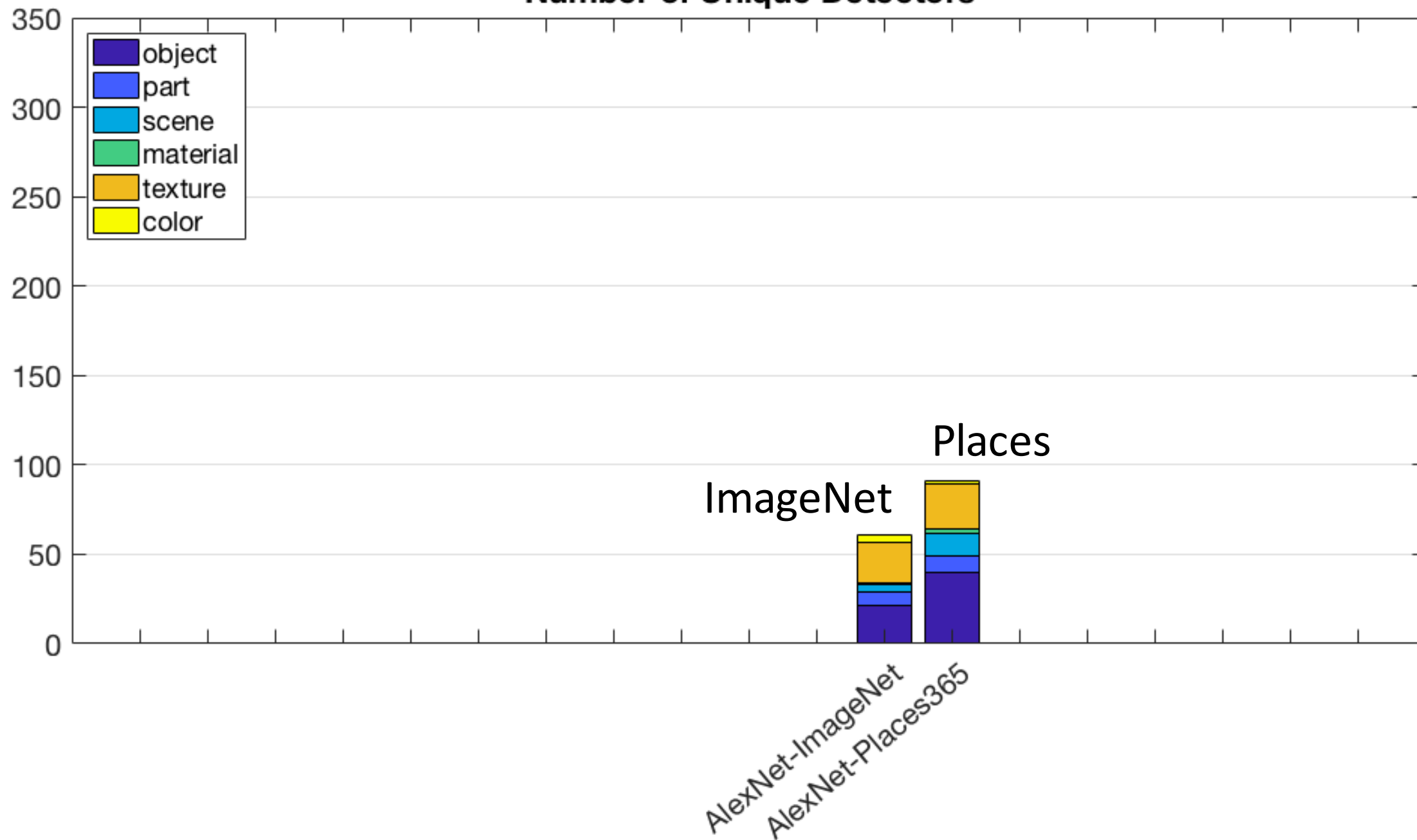


Colorization, ECCV'16

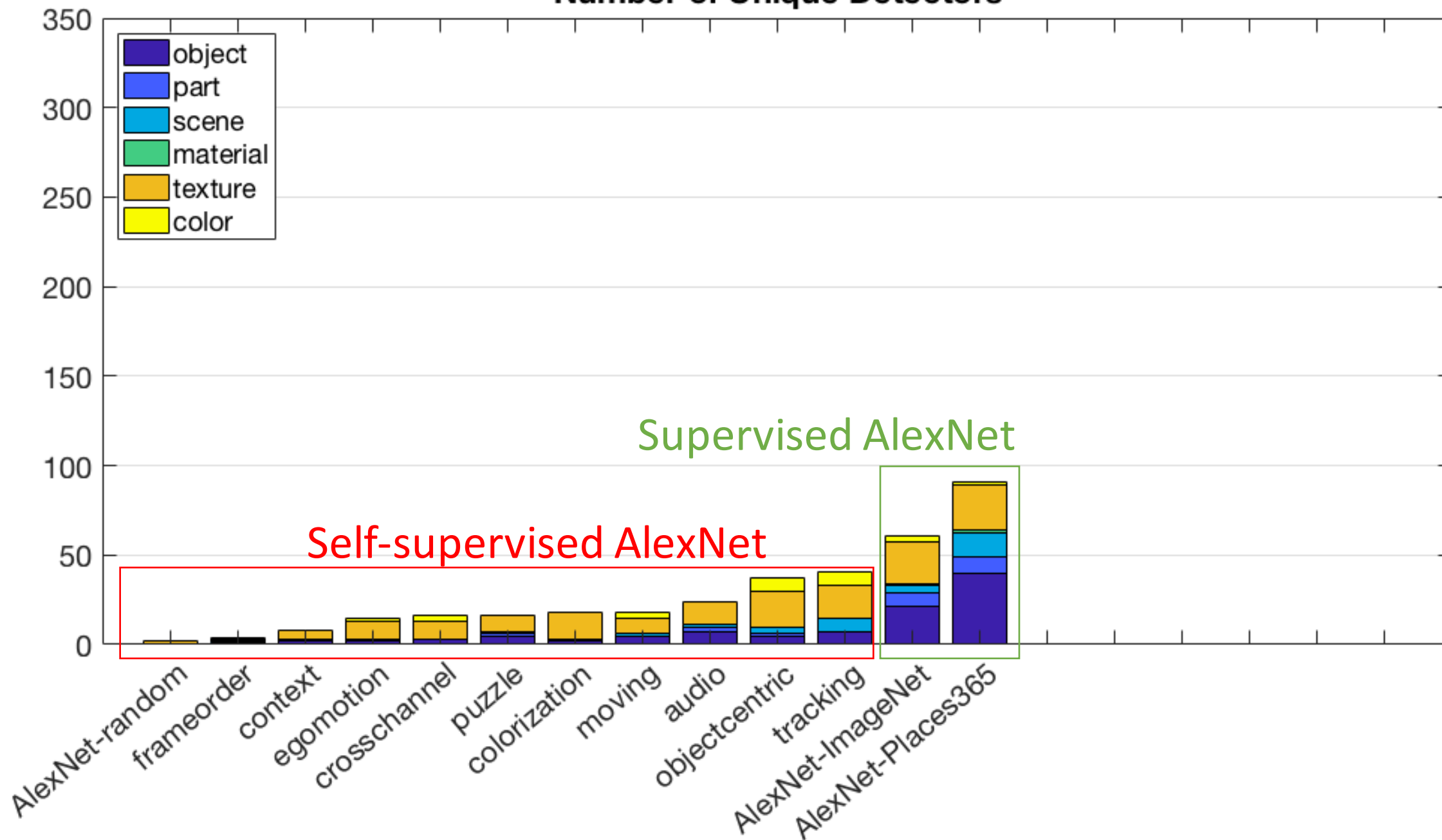


Audio prediction, ECCV'16

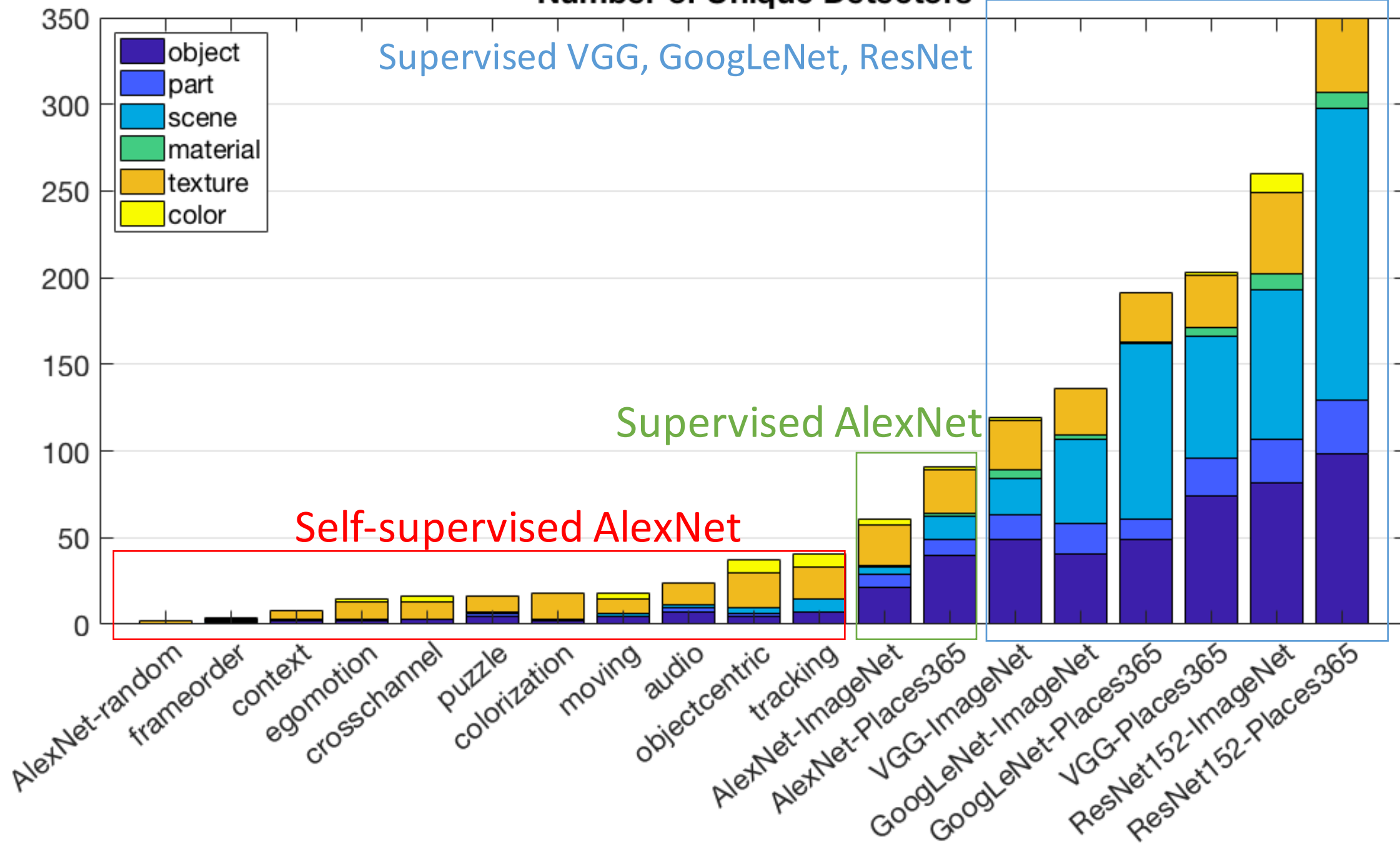
Number of Unique Detectors



Number of Unique Detectors



Number of Unique Detectors

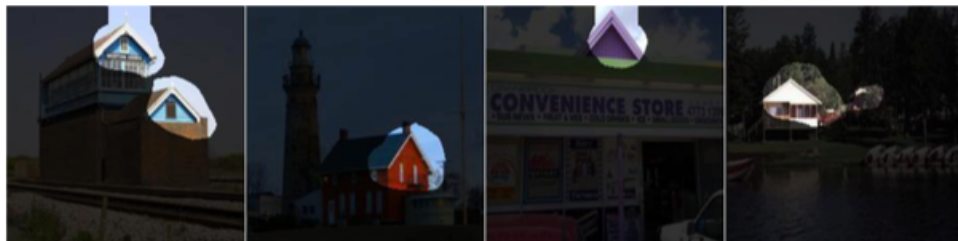


House

AlexNet

conv5 unit 36

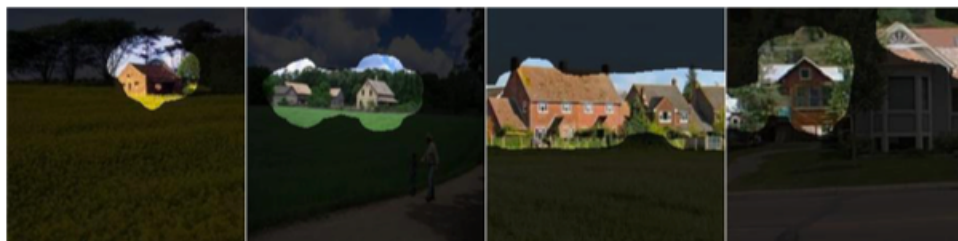
IoU=0.053



VGG

conv5_3 unit 243

IoU=0.070



GoogLeNet

inception_4e unit 789

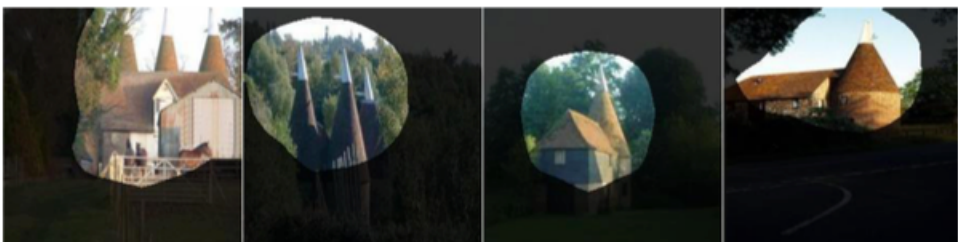
IoU=0.137



ResNet

res5c unit 1410

IoU=0.142



Airplane

conv5 unit 13

IoU=0.101



conv5_3 unit 151

IoU=0.150



inception_4e unit 92

IoU=0.164

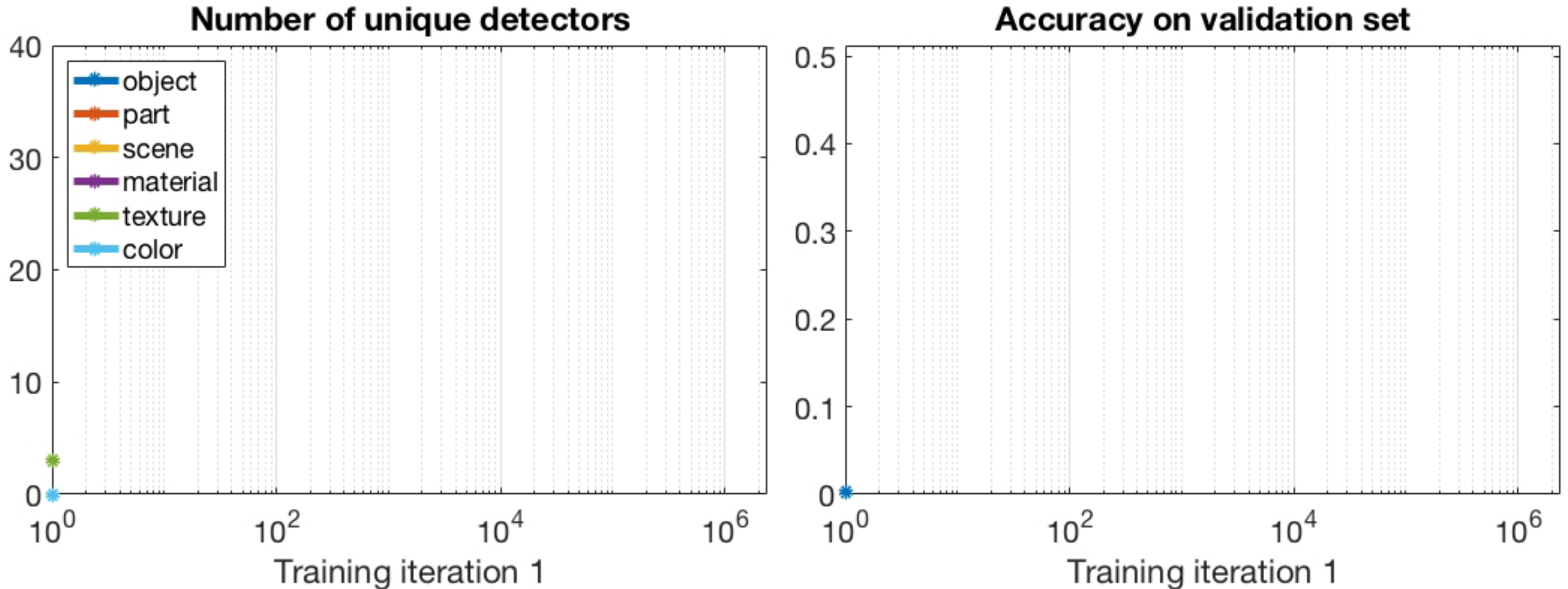


res5c unit 1243

IoU=0.172

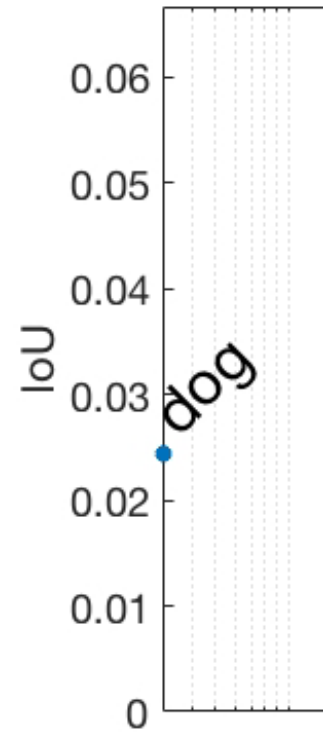
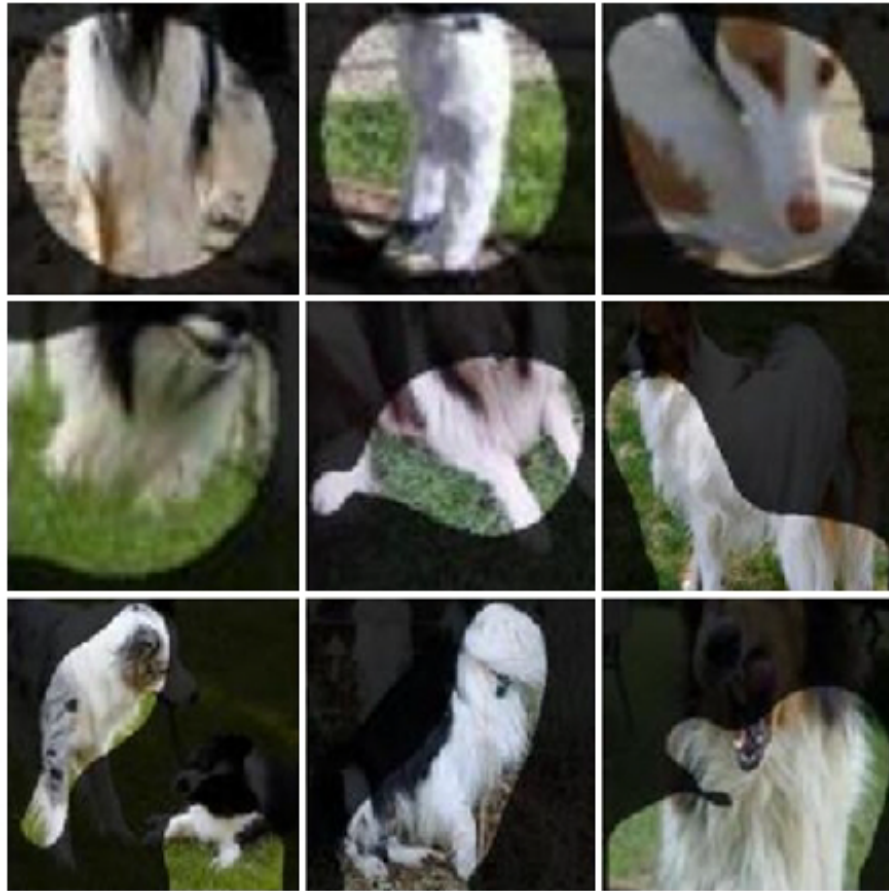


Emergence of Interpretable Units during Training



Fine-tuning from ImageNet to Places

Unit 8 at conv5 layer

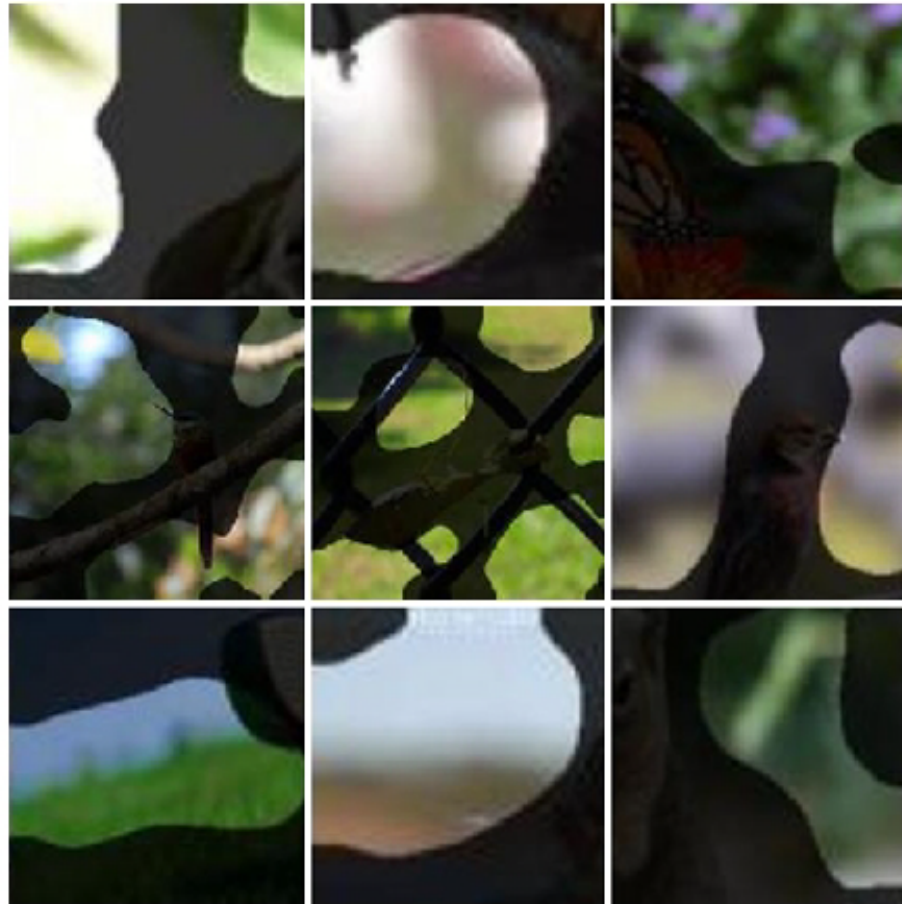


Before fine-tuning

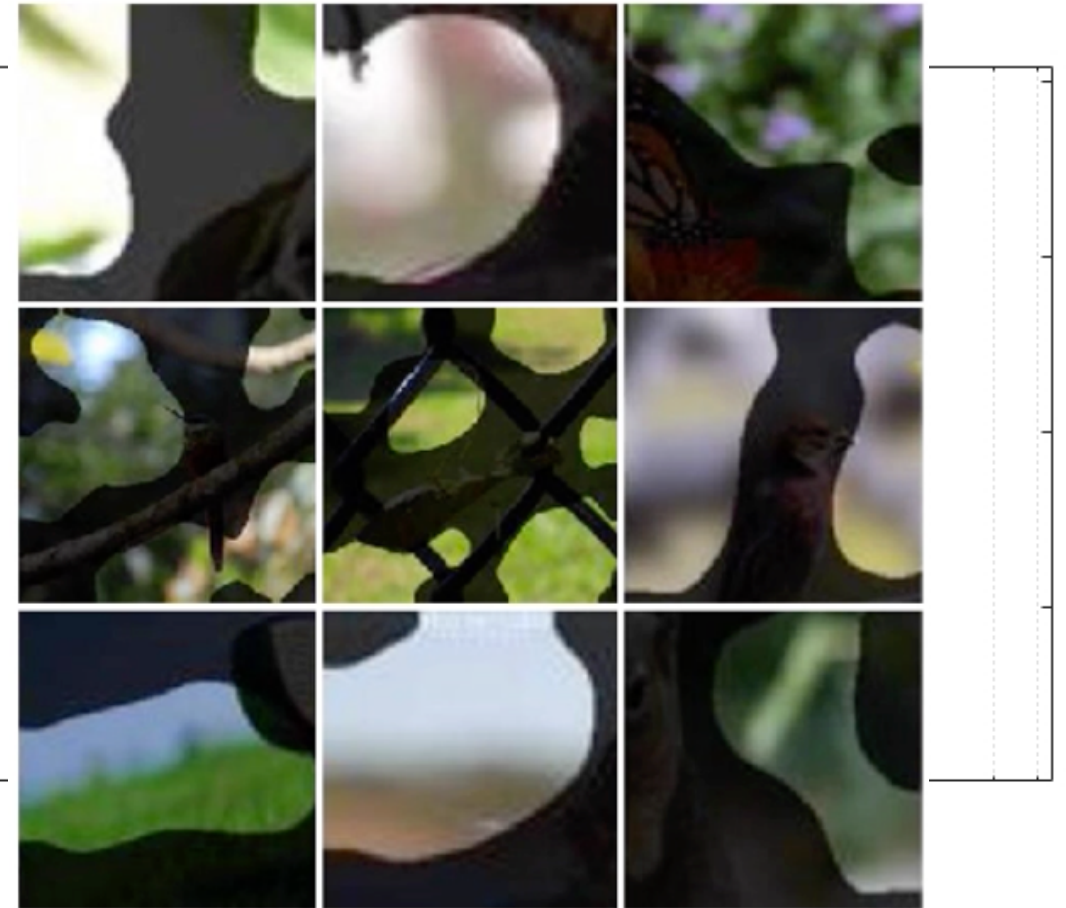
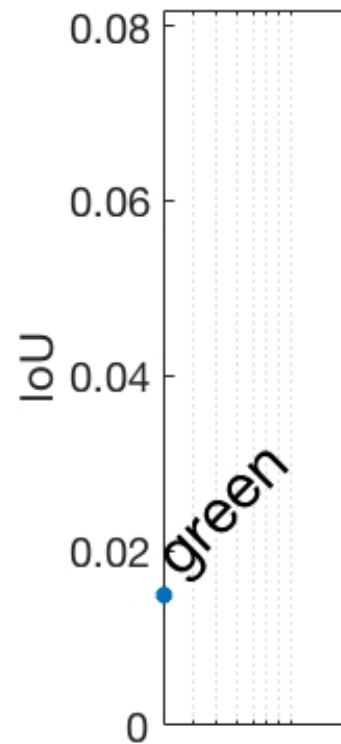


Fine-tuning from ImageNet to Places

Unit 52 at conv5 layer



Before fine-tuning



Fine-tuning from Places to ImageNet

Unit 35 at conv5 layer

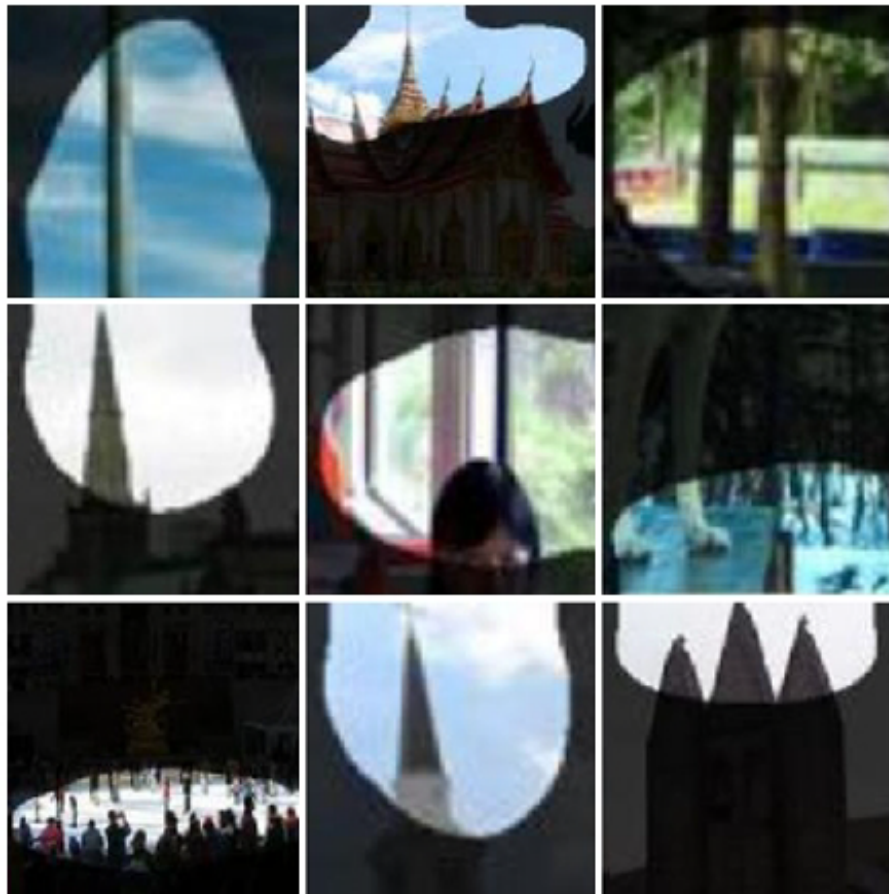


Before fine-tuning

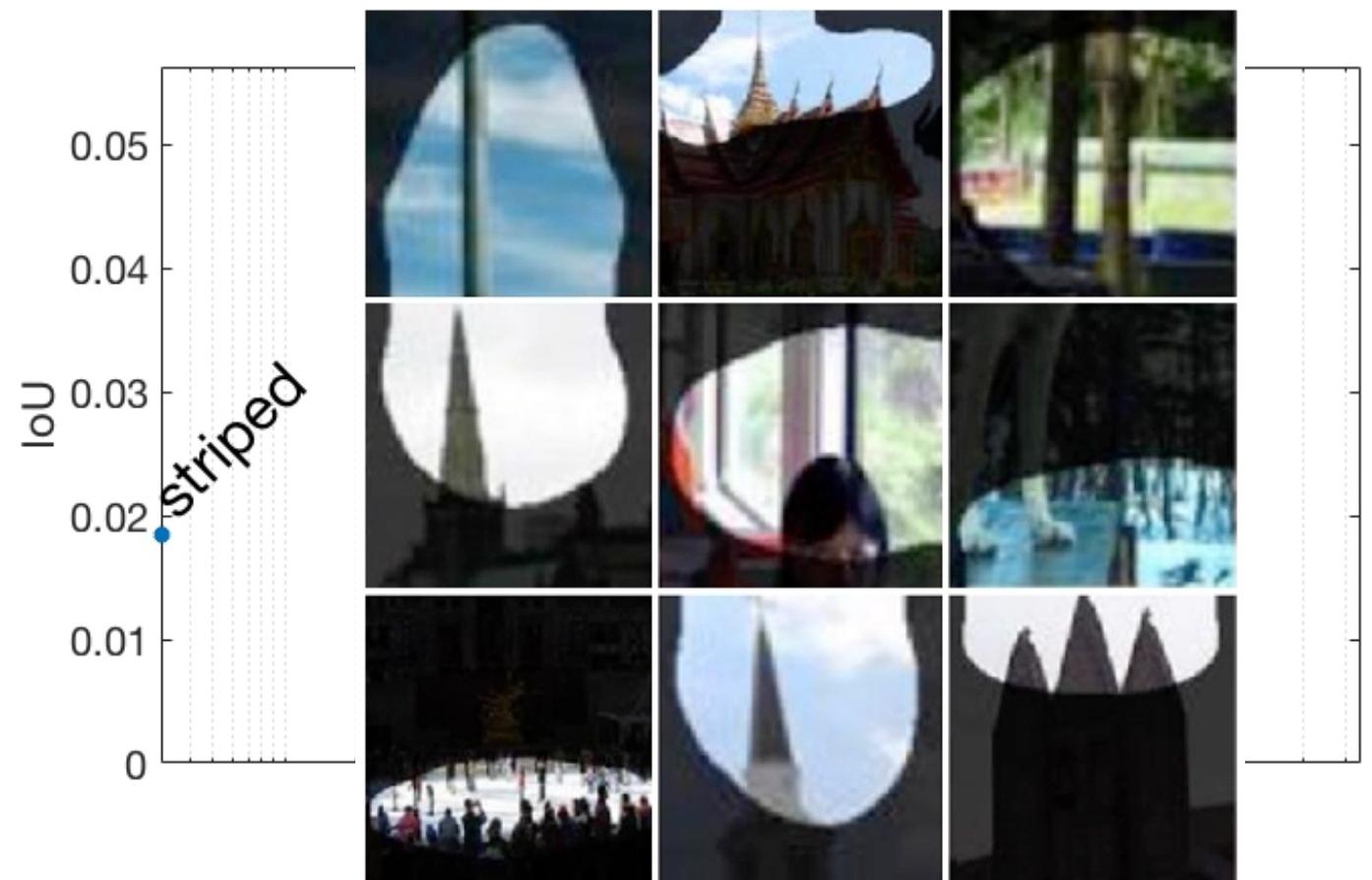


Fine-tuning from Places to ImageNet

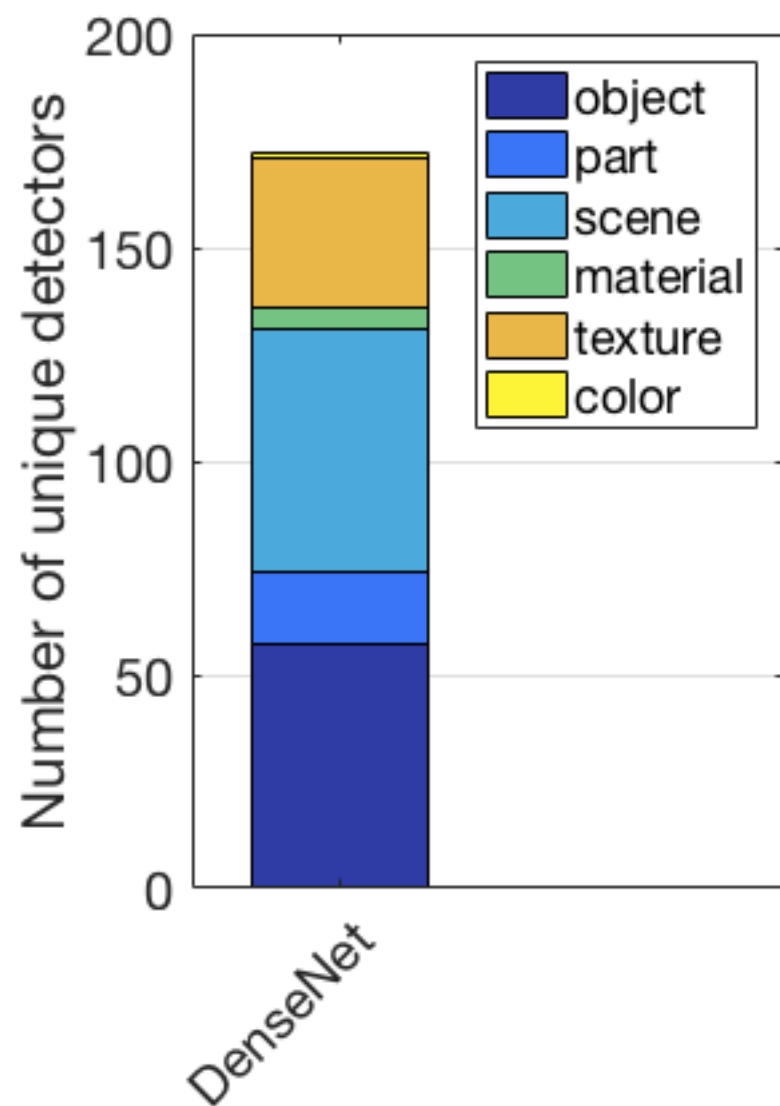
Unit 103 at conv5 layer



Before fine-tuning



Interpretable Units in DenseNet



Horse

unit 651



IoU=0.151

Bus

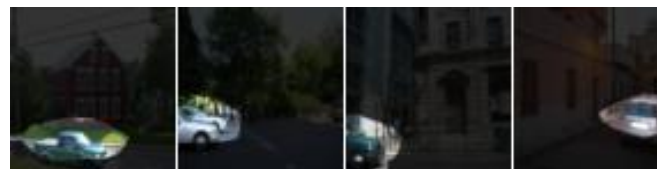
unit 740



IoU=0.152

Car

unit 529



IoU=0.150

Bed

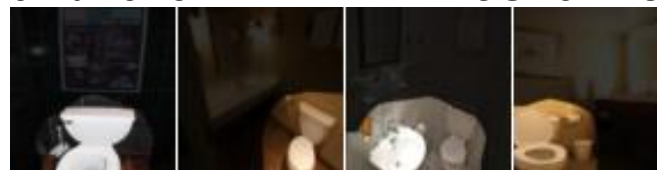
unit 953



IoU=0.133

Toilet

unit 1016



IoU=0.125

Bakery Shop

unit 926



IoU=0.167

Wheel

unit 648



IoU=0.165

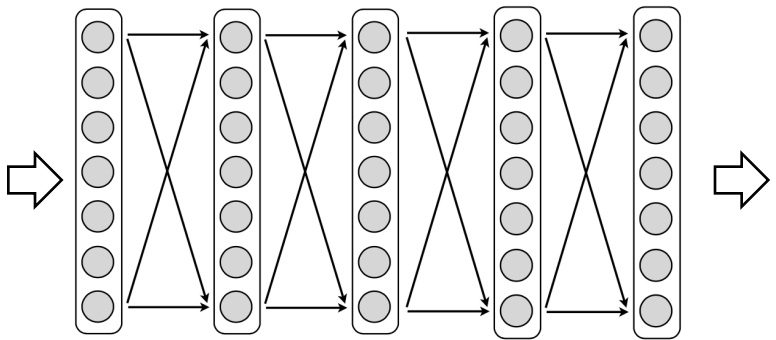
Cobwebbed

unit 700



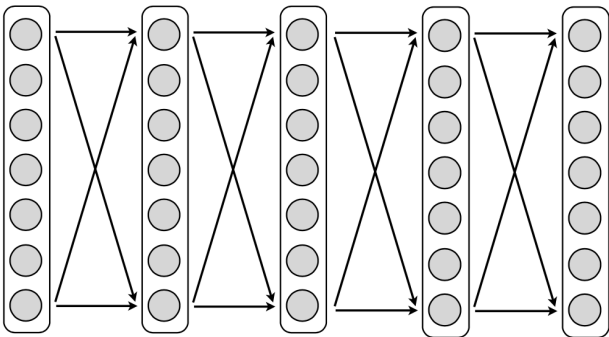
IoU=0.327

Conclusion

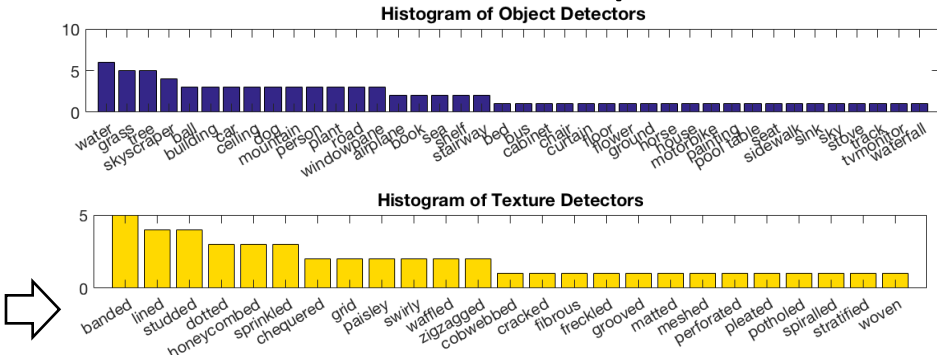


Living room
Kitchen
Coast
Theater
...

Network Dissection



Dissection Report



unit 79 car, IoU=0.13



unit 107 road, IoU=0.15

