Visualizing and Interpreting Deep Neural Networks

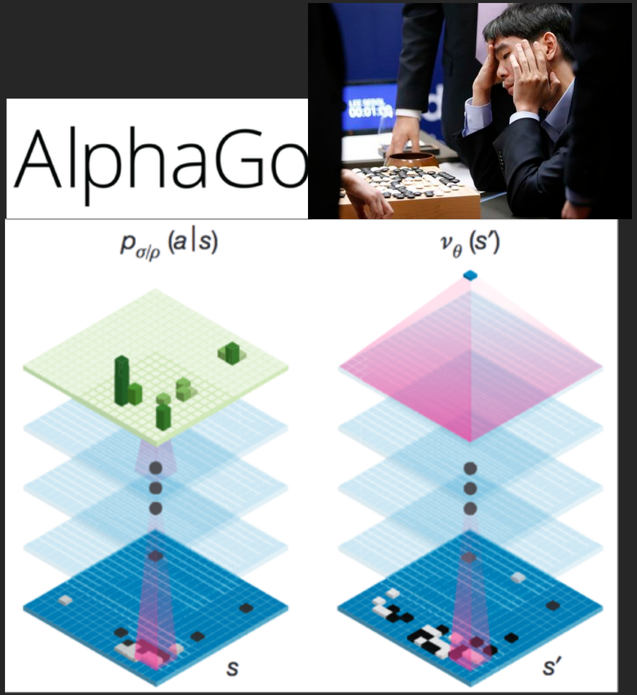
Bolei Zhou

Department of Information Engineering

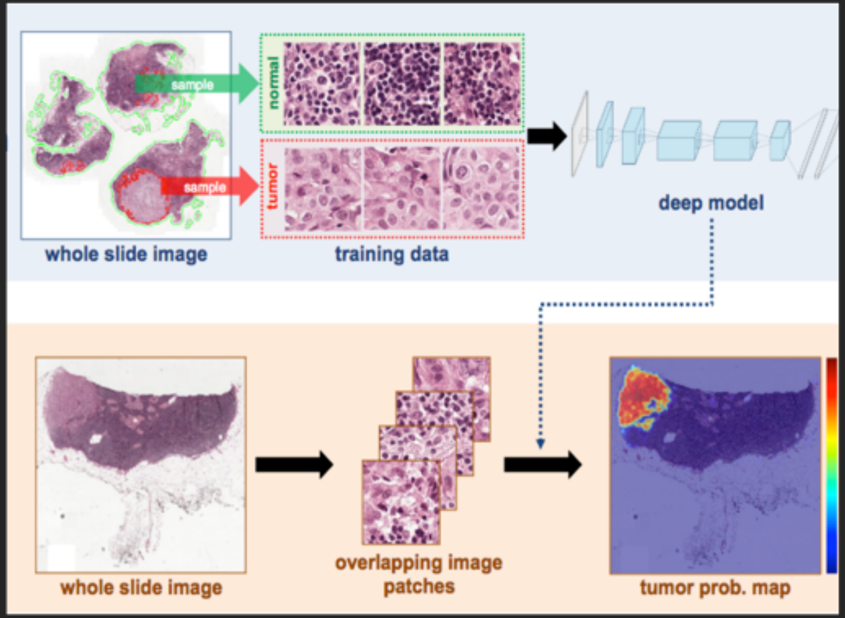
The Chinese University of Hong Kong

Deep Neural Networks are Everywhere

Playing Go



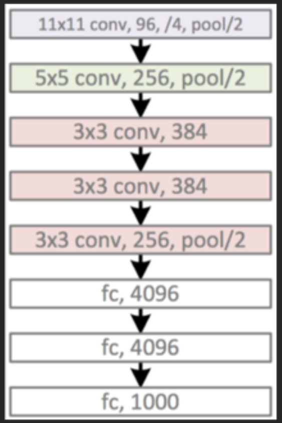
Making Medical Decision



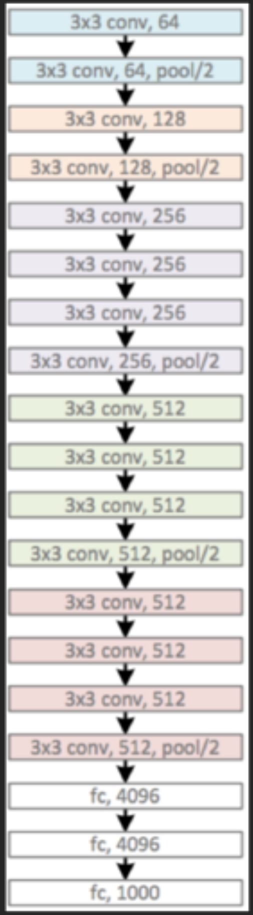
Understanding Scenes

Deep Neural Networks for Visual Recognition

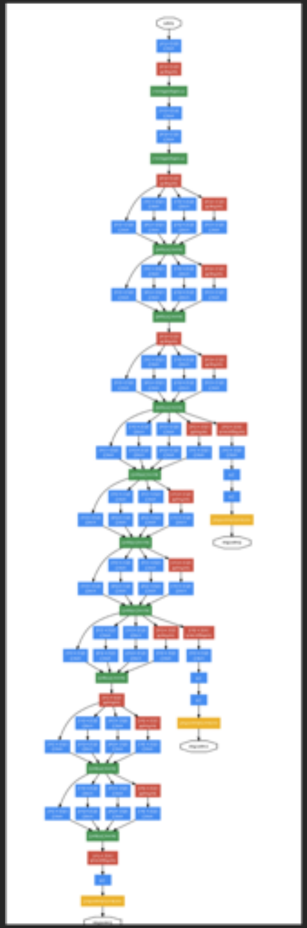
AlexNet


11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

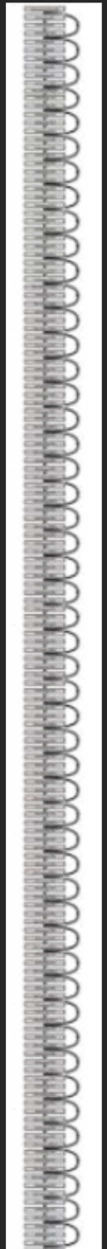
VGG


3x3 conv, 64
3x3 conv, 64, pool/2
3x3 conv, 128
3x3 conv, 128, pool/2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
fc, 4096
fc, 4096
fc, 1000

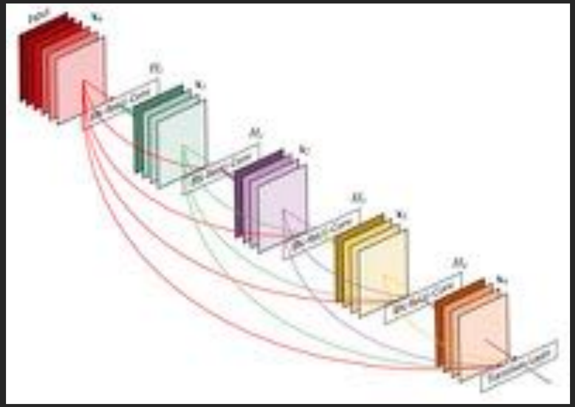
GoogLeNet



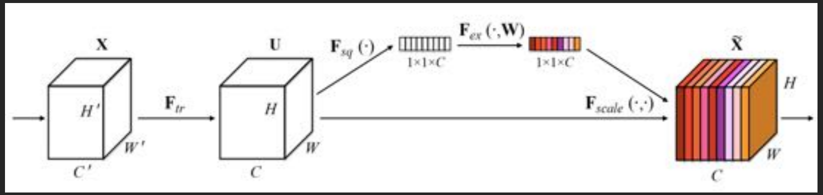
ResNet
>100 layers



DenseNet >250 layers

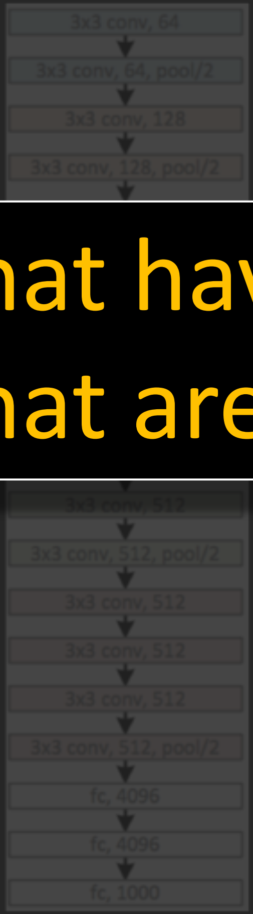


SE Net > 100 layers

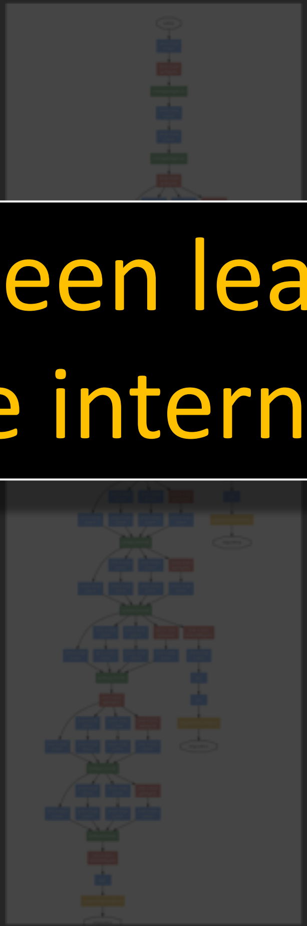


Deep Neural Networks for Visual Recognition

VGG

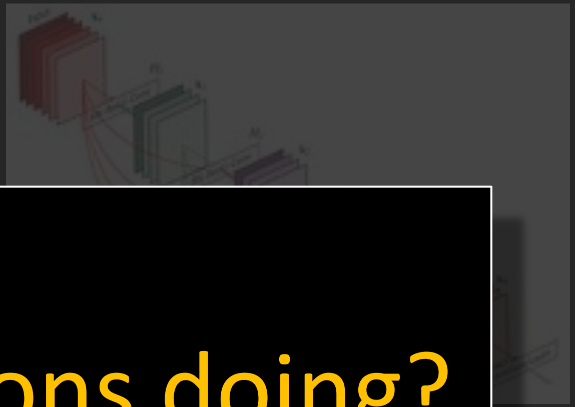


GoogLeNet



ResNet
>100 layers

DenseNet >250 layers



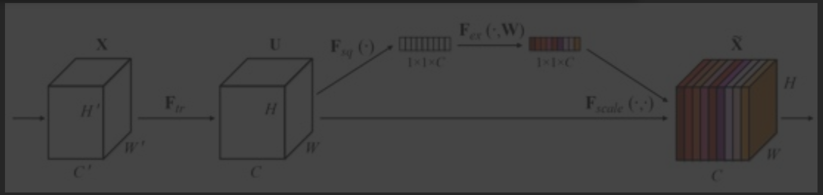
What have been learned inside?

What are the internal representations doing?

AlexNe

11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

SE Net > 100 layers

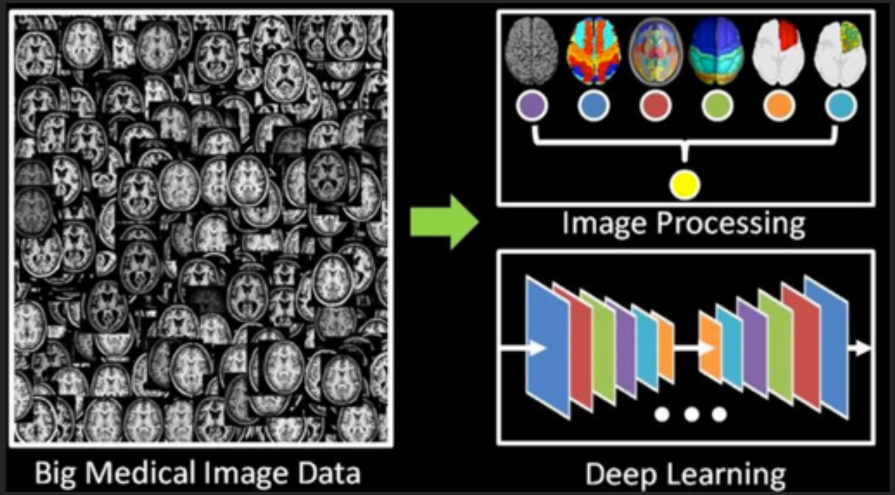


Interpretability of Deep Neural Networks

Safety of AI models

Autonomous Driving

Trust of AI decision



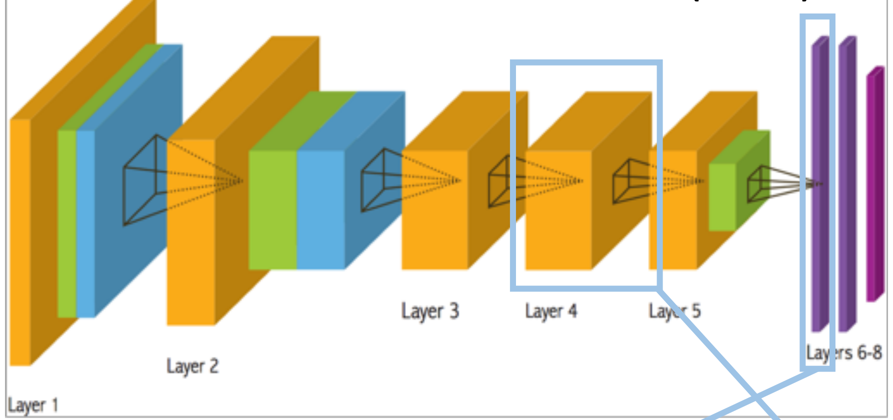
Medical Diagnosis

Policy and Regulation

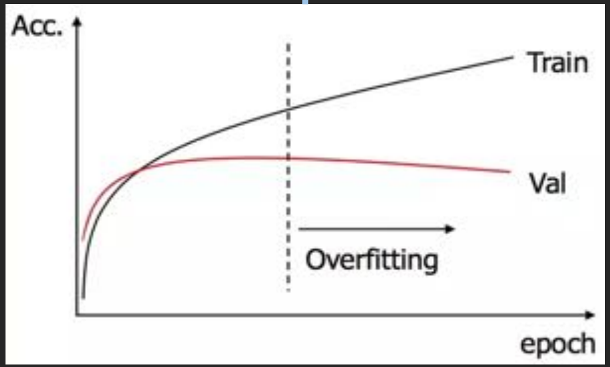
Right to the explanation
for algorithmic decisions

Understanding Networks at Different Granularity

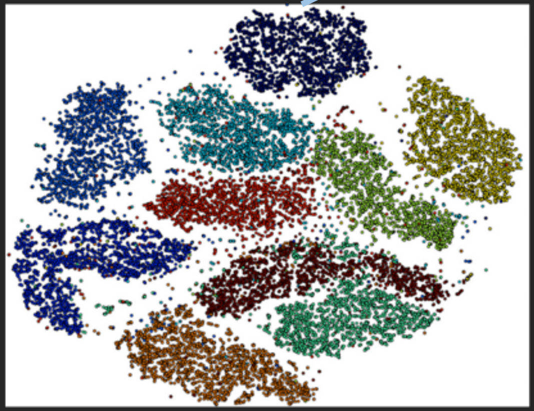
Convolutional Neural Network (CNN)



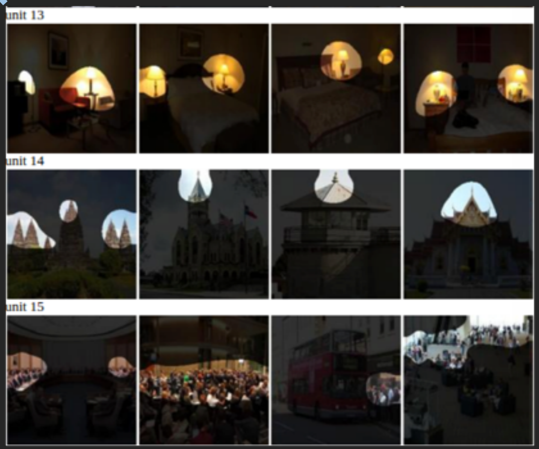
Cafeteria (0.9)



Network as a Whole



Feature Space



Individual Units

Outline

- What is a unit doing?
- What are all the units doing?
- How units are relevant to prediction?
- What’s inside generative model?

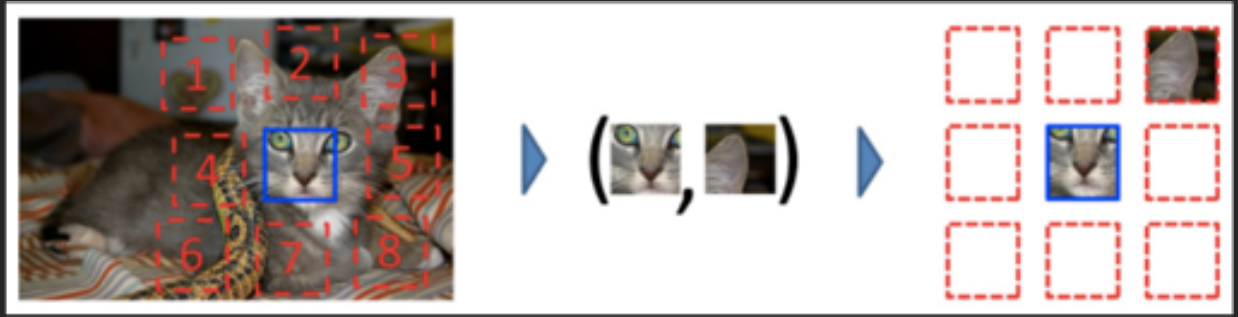
Sources of Deep Representations

Supervised Learning

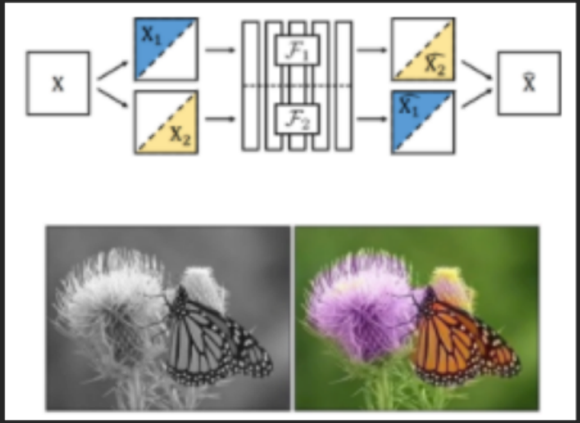
Object Recognition

Scene Recognition

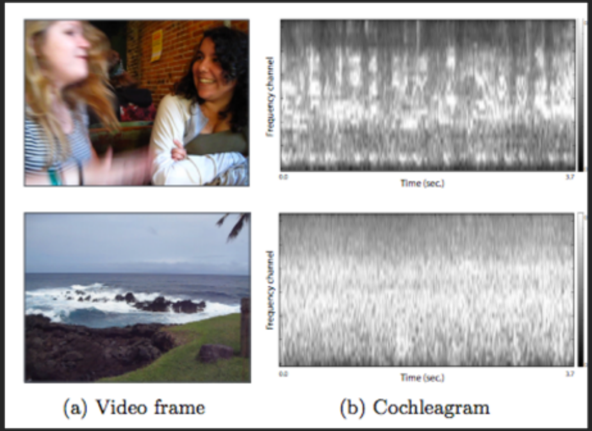
Self Supervised Learning



Context prediction, ICCV'15



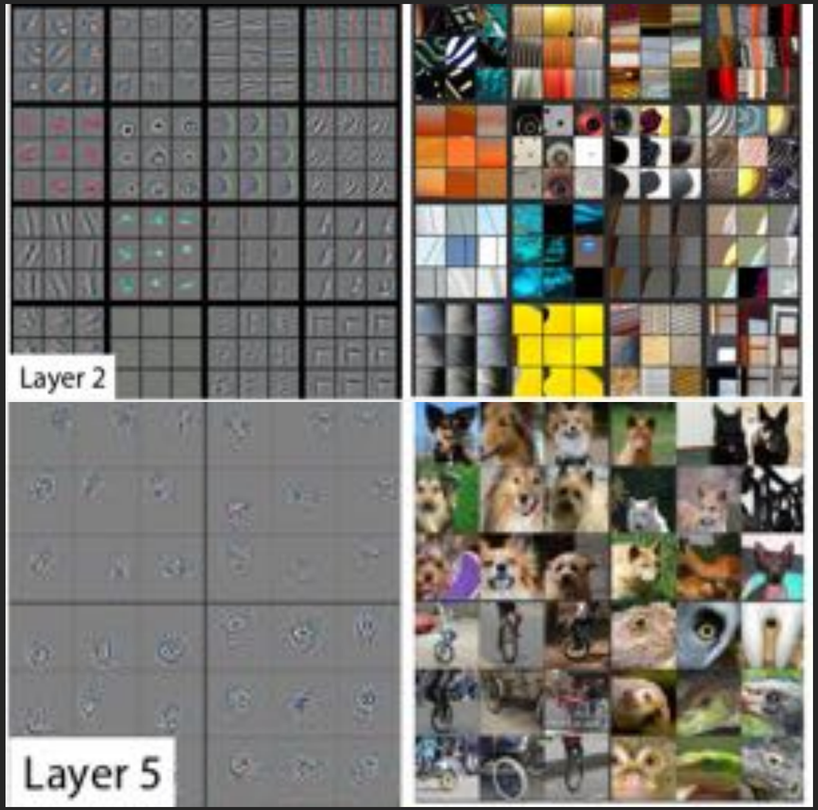
Colorization
ECCV'16 and CVPR'17



Audio prediction, ECCV'16

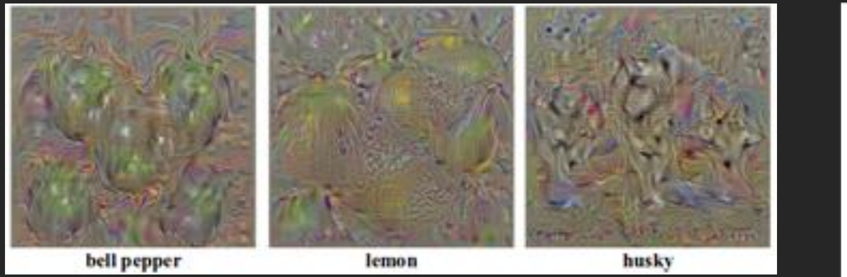
What is a unit doing? - Visualize the unit

Deconvolution



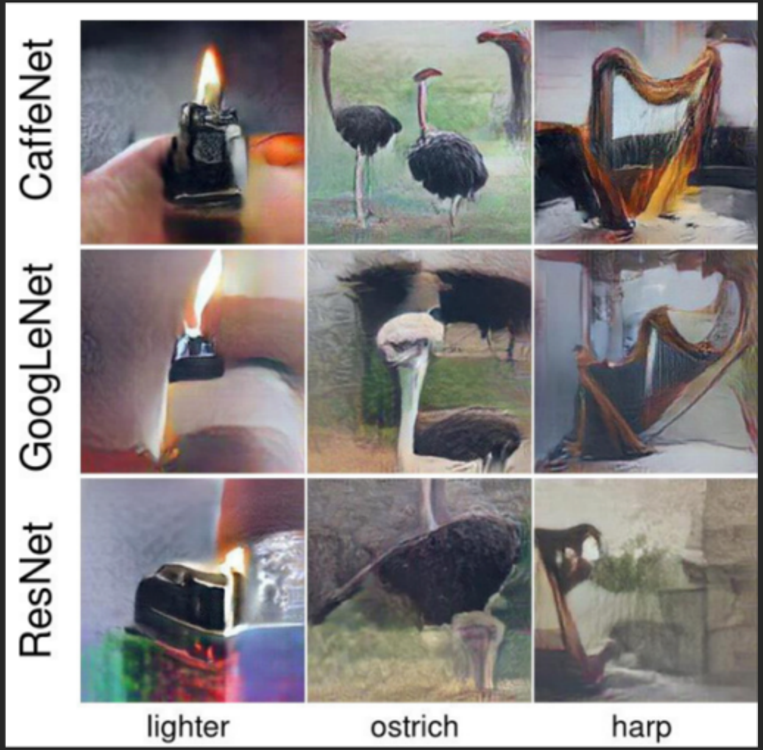
[Zeiler et al., ECCV’14]
[Girshick et al., CVPR’14]

Back-propagation



[Simonyan et al., ICLR’15]
[Springerberg et al., ICLR’15]
[Selvaraju, ICCV’17]

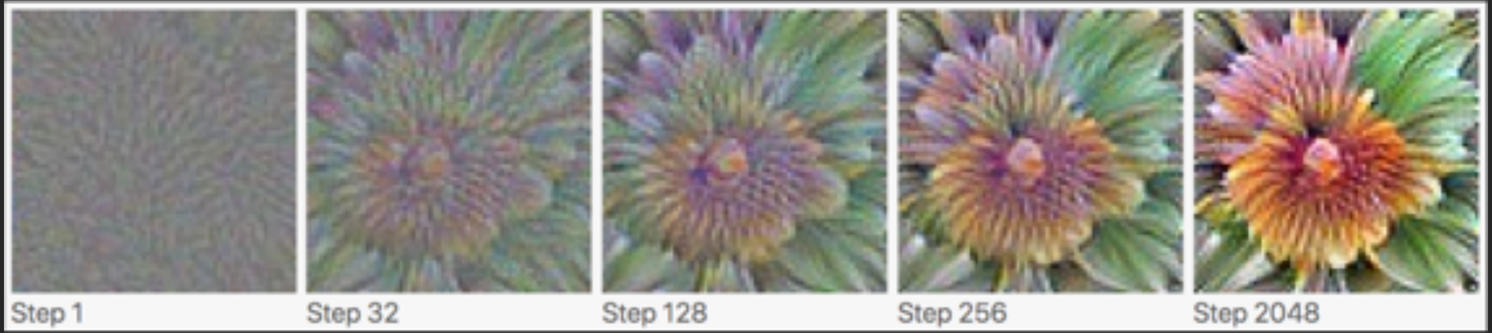
Image Synthesis

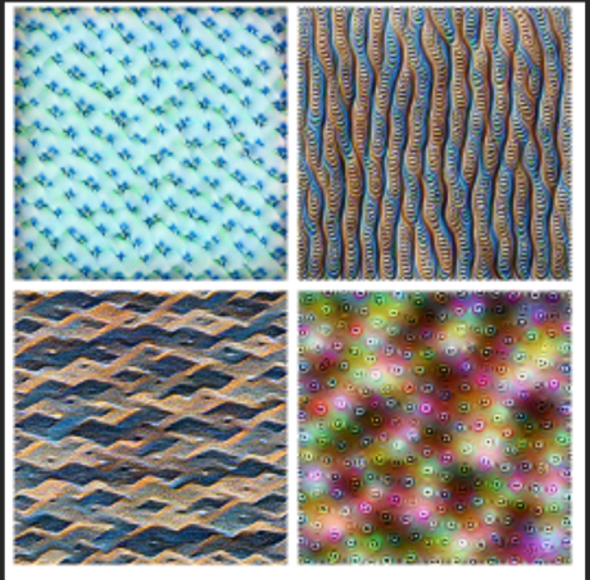


[Nguyen et al., NIPS’16]
[Dosovitskiy et al., CVPR’16]
[Mahendran, et al., CVPR’15]

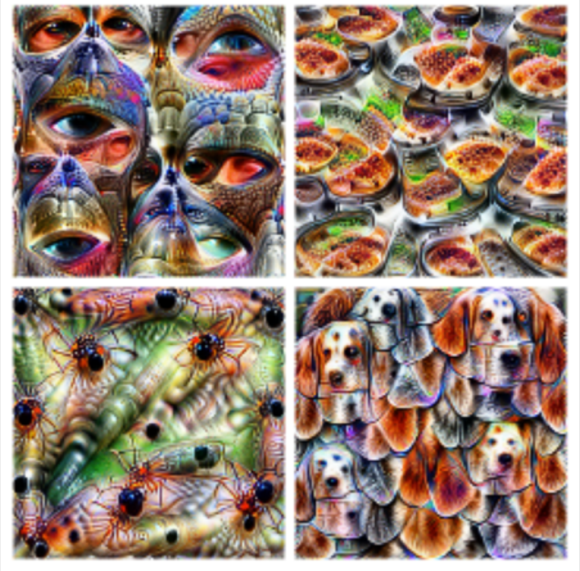
Gradient-based Visualization

Iteratively use gradient to optimize an image to activate a particular unit





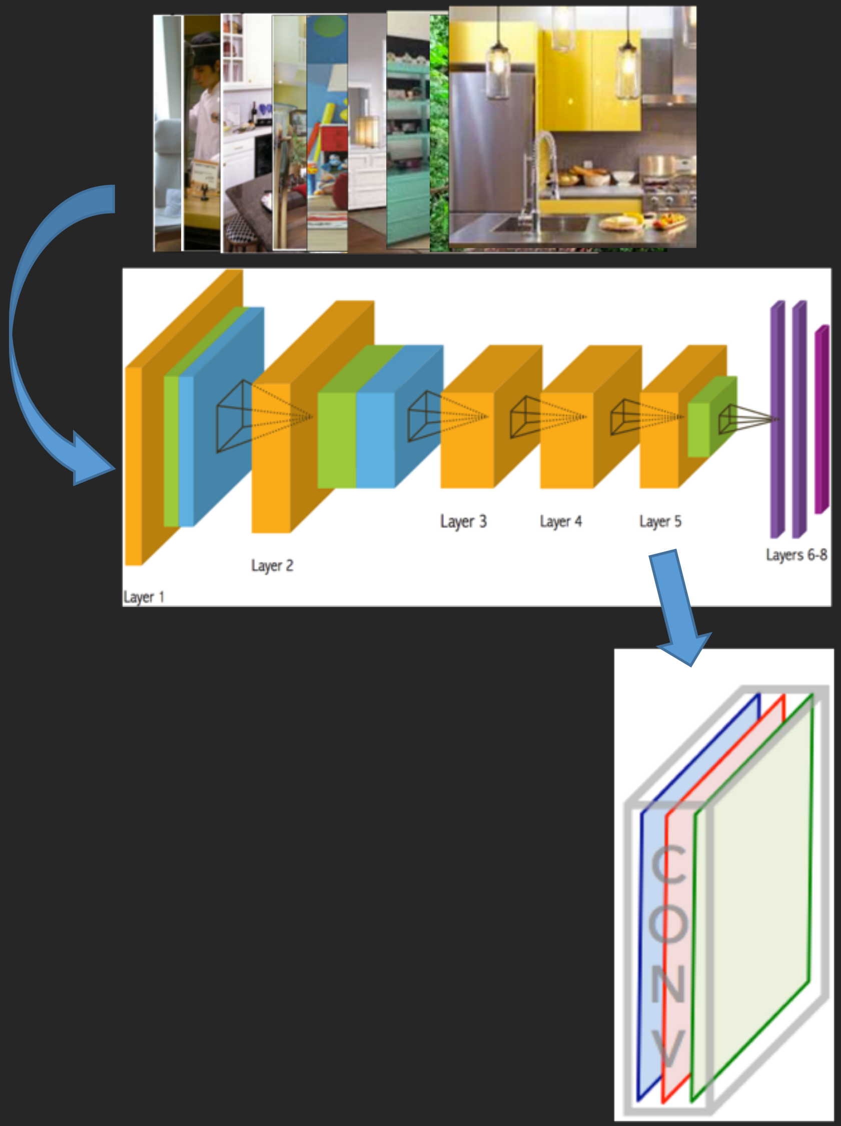
Textures (layer mixed3a)



Objects (layers mixed4d & mixed4e)

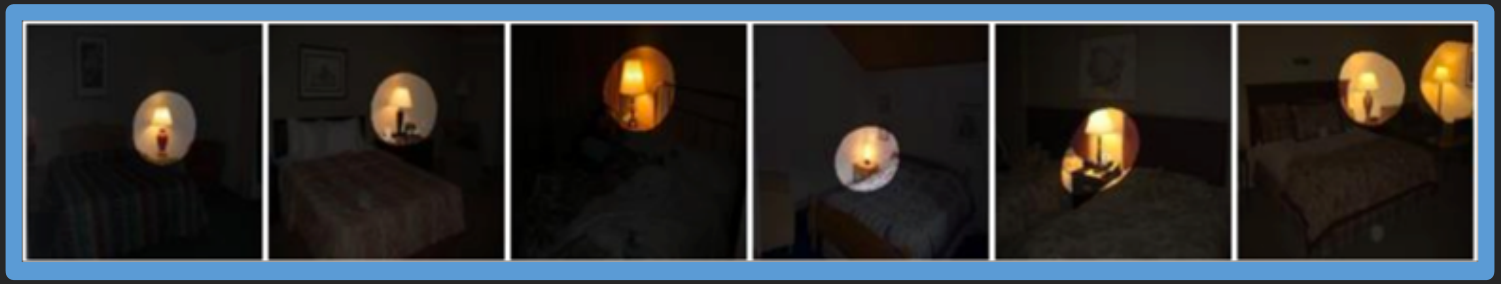
Chris Olah, et al. https://distill.pub/2017/feature-visualization/

Data Driven Visualization

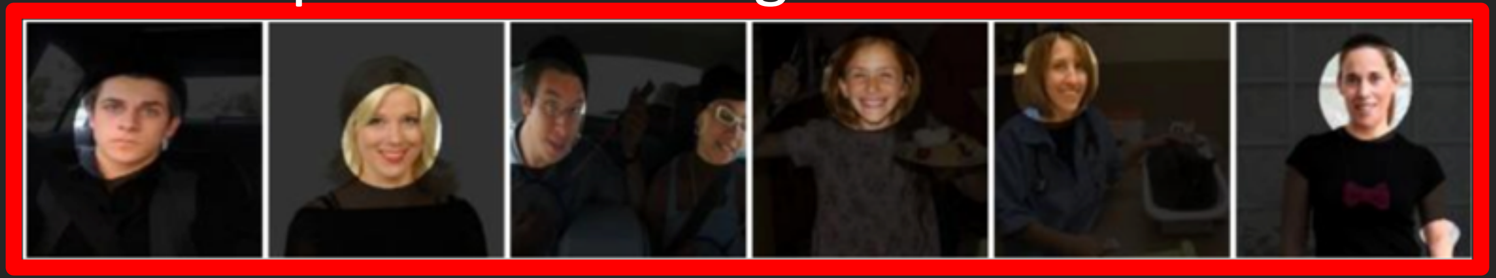


Layer 5

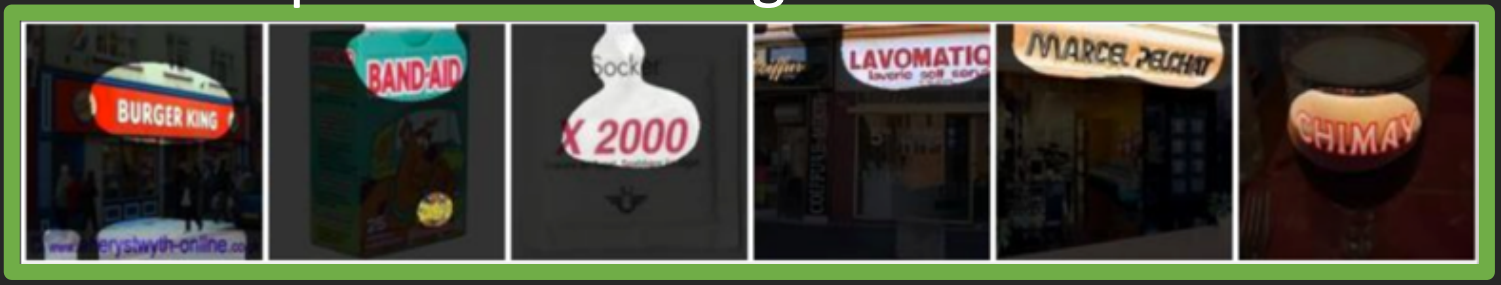
Unit1: Top activated images



Unit2: Top activated images



Unit3: Top activated images



https://github.com/metalbubble/cnnvisualizer

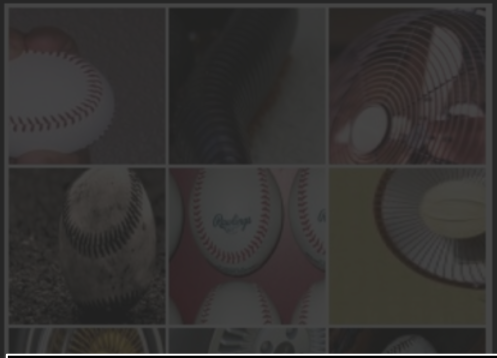
Comparison of Visualizations

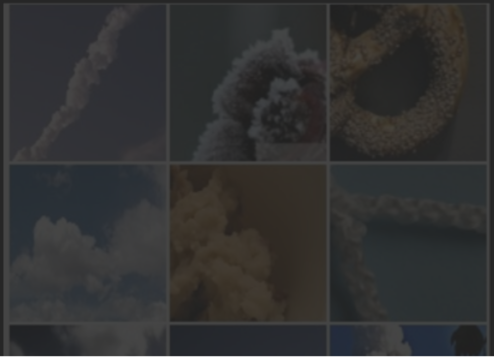
Mixed4a Unit 6

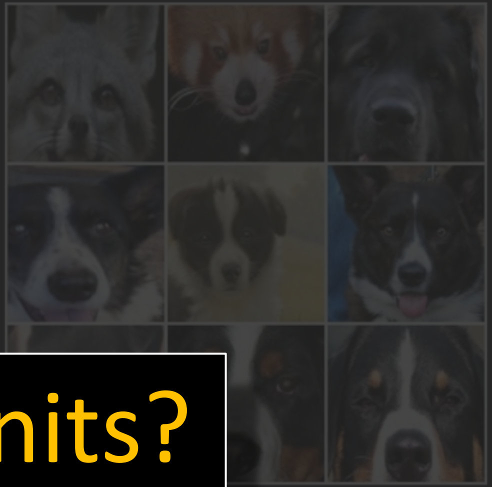
Mixed4a Unit 453

Mixed4a Unit 240

Data driven



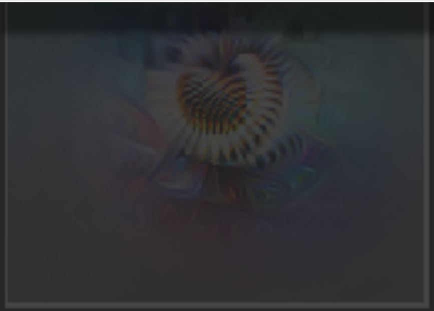


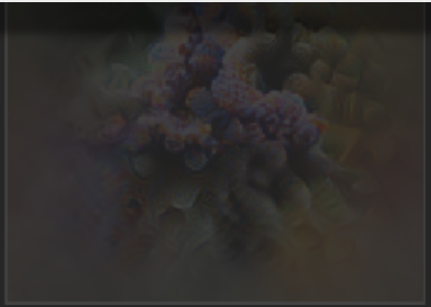


How to Compare Different Units?

How to Interpret All the Units?

Gradient-based





Baseball or Stripes?

Clouds or fluffiness?

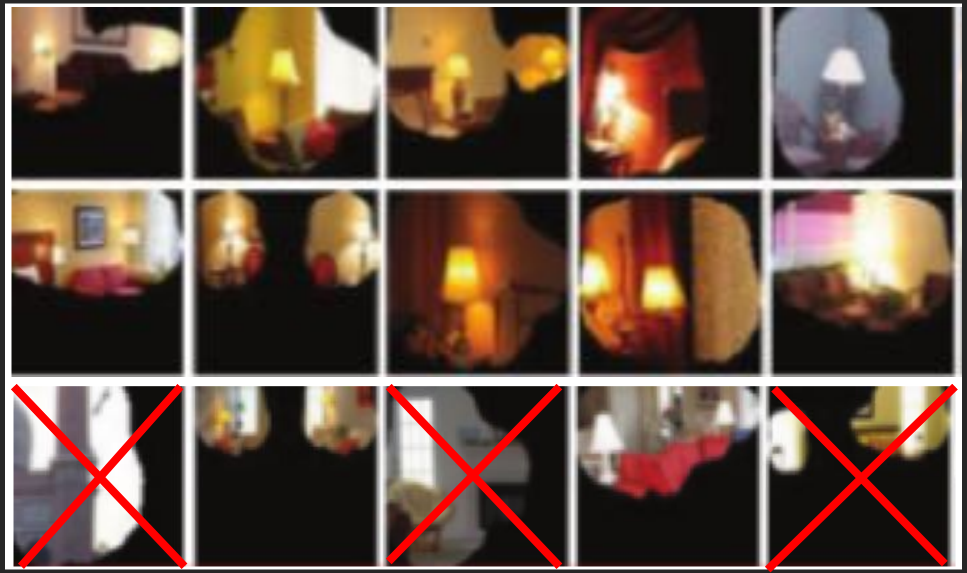
Dog face or snouts?

Annotating the Interpretation of Units

Amazon Mechanical Turk

Word/Description to summarize the images:

Lamp



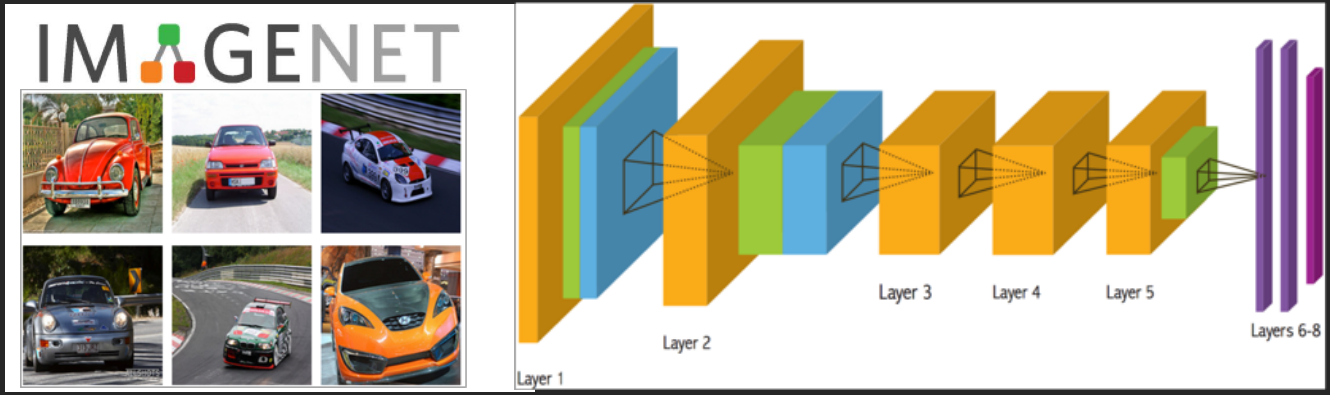
Which category the description belongs to:

- Scene
- Region or surface
- [✓ Object]
- Object part
- Texture or material
- Simple elements or colors

[Zhou, Khosla, Lapedriza, Oliva, Torralba. ICLR 2015]

Two Recognition Tasks and Two Networks

CNN for Object Classification

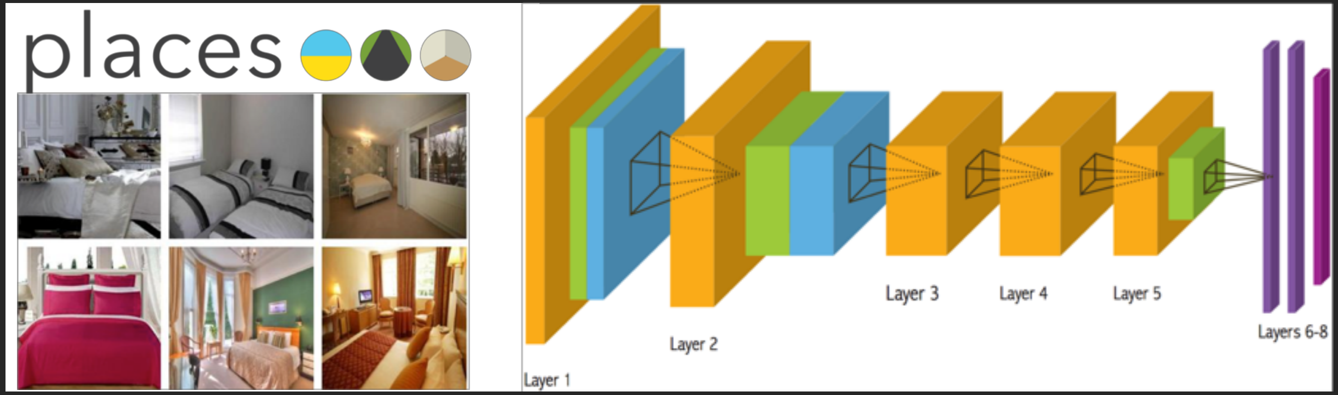


1000 classes

Race car

…

CNN for Scene Recognition



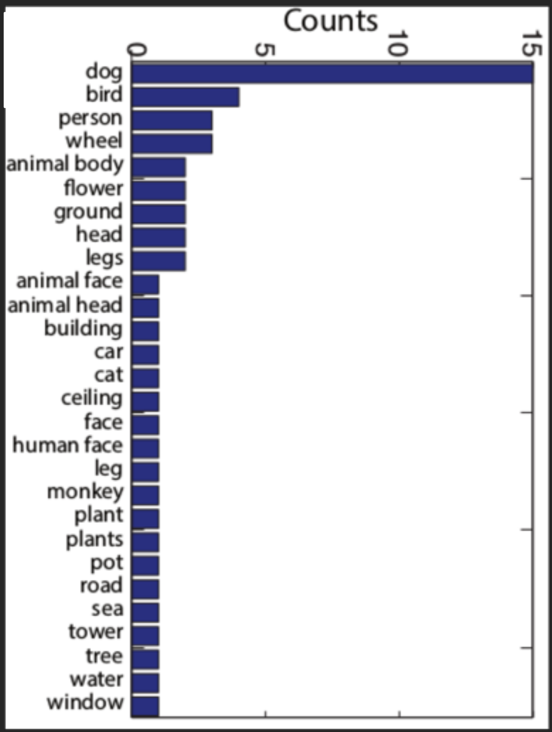
365 classes

Living room

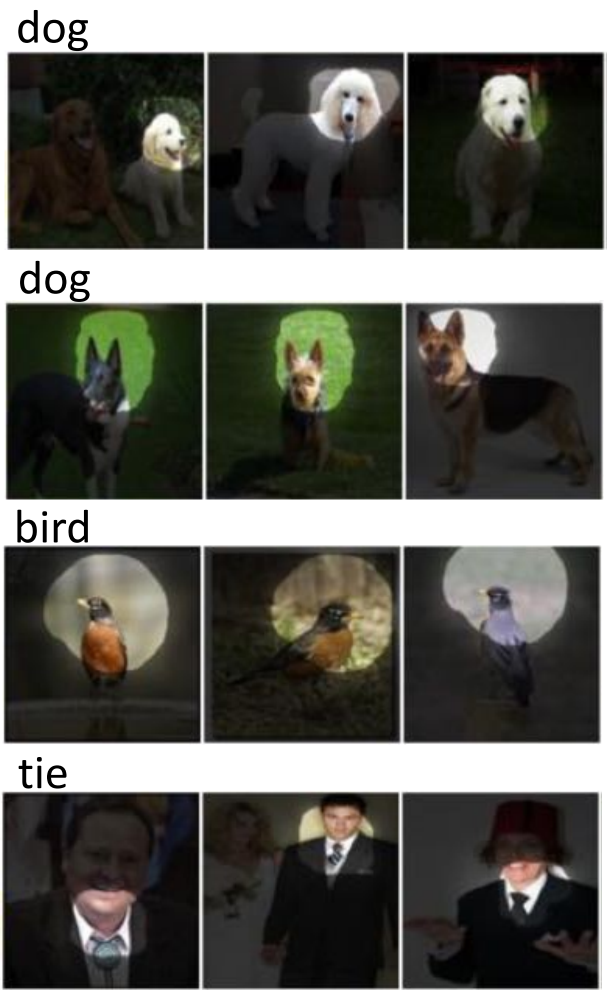
…

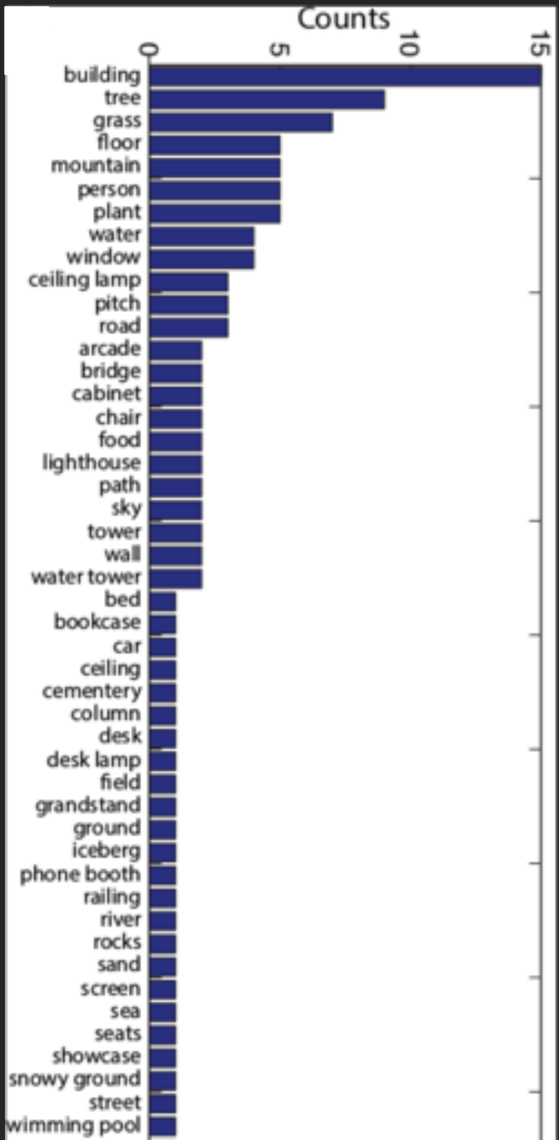
[**Zhou**, Khosla, Lapedriza, Oliva, Torralba. ICLR 2015]

Interpretable Representations for Objects and Scenes

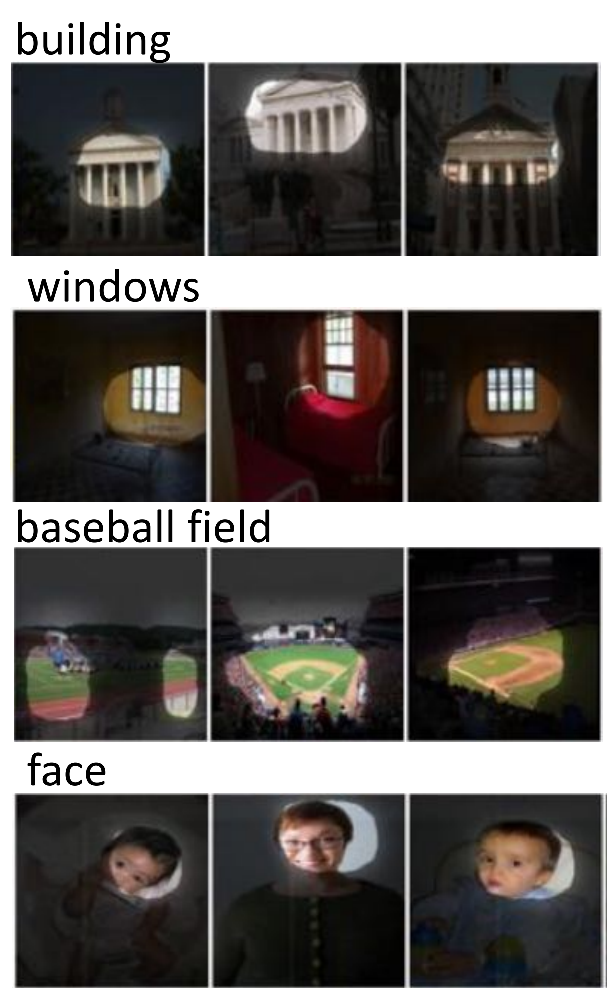


59 units as objects at conv5 of AlexNet on ImageNet

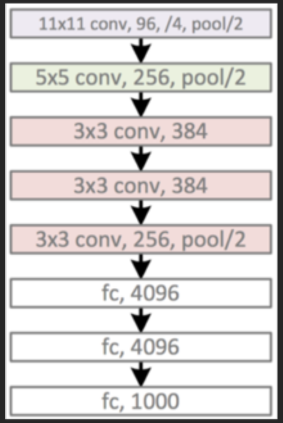




151 units as objects at conv5 of AlexNet on Places



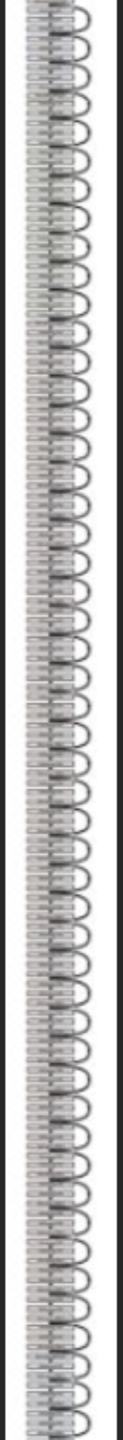
2012: AlexNet
5 layers
1,000 units



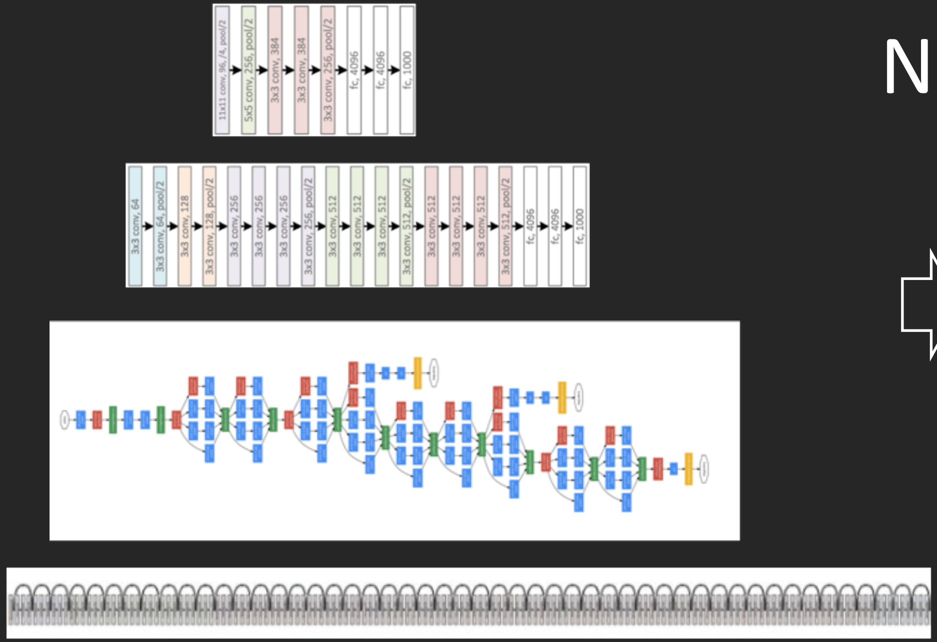
11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

Scale up Interpretation to Deep Networks

Now: ResNet, DenseNet
> 100 layers
> 100,000 units

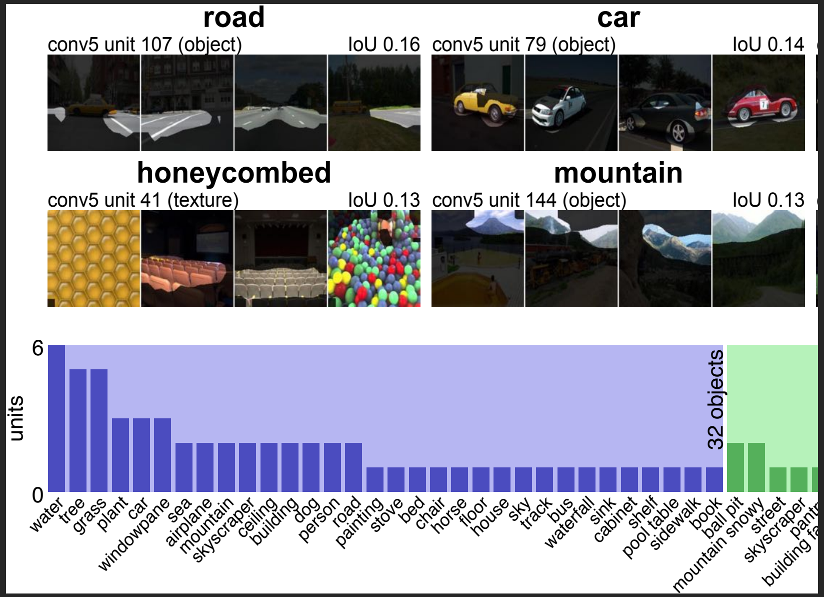


Quantify the Interpretability of Networks



Network Dissection

Interpretable Units

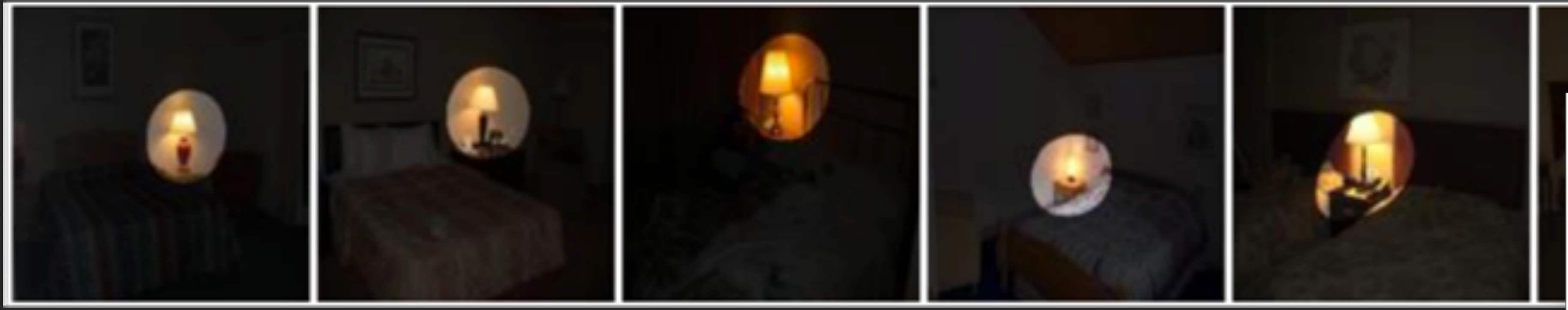


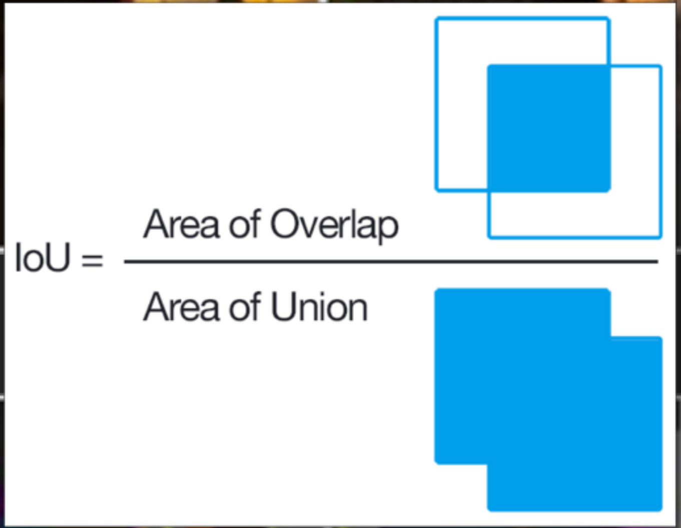
[Bau*, Zhou*, Khosla, Oliva, Torralba. CVPR 2017]

Evaluate Unit for Semantic Segmentation

Testing Dataset: 60,000 images annotated with 1,200 concepts

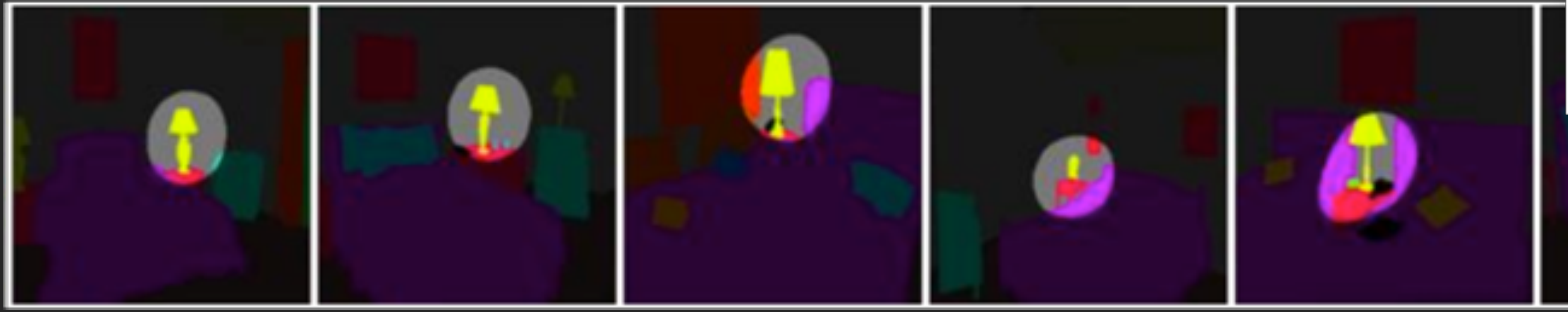
Unit 1: Top activated images from the Testing Dataset



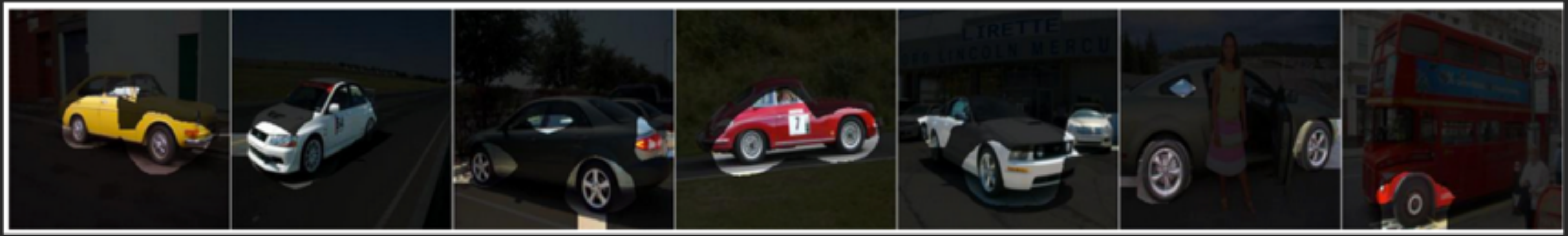


$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

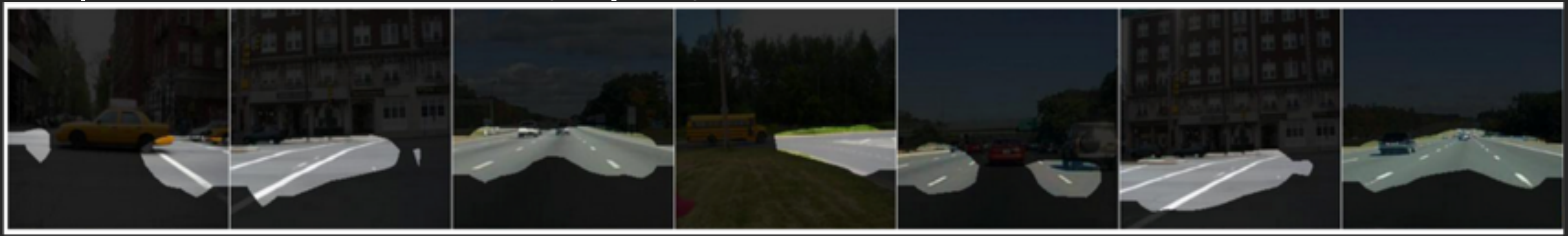
Top Concept: Lamp, Intersection over Union (IoU)= 0.23



Layer5 unit 79 car (object) IoU=0.13



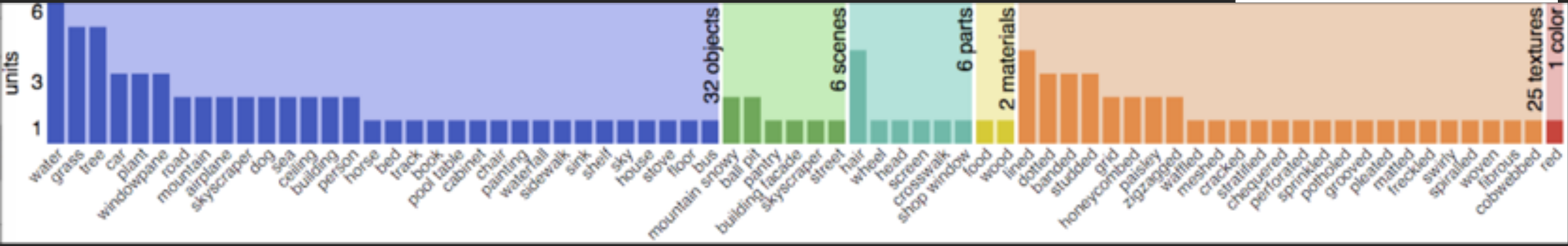
Layer5 unit 107 road (object) IoU=0.15



places

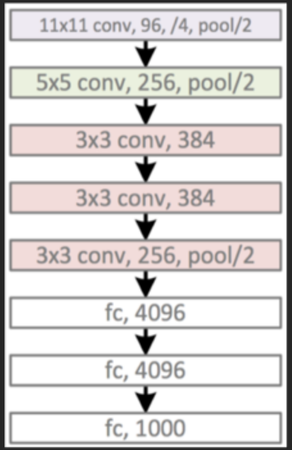
THE SCENE RECOGNITION DATABASE

118/256 units covering 72 unique concepts



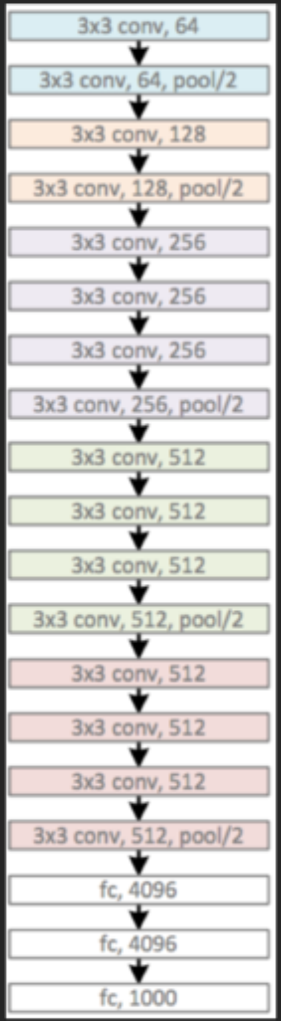
Compare Different Representations of Architectures

AlexNet



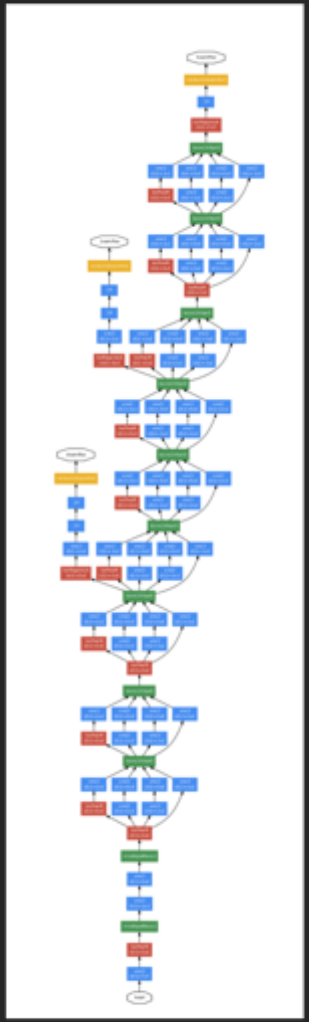
11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

VGG



3x3 conv, 64
3x3 conv, 64, pool/2
3x3 conv, 128
3x3 conv, 128, pool/2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
fc, 4096
fc, 4096
fc, 1000

GoogLeNet



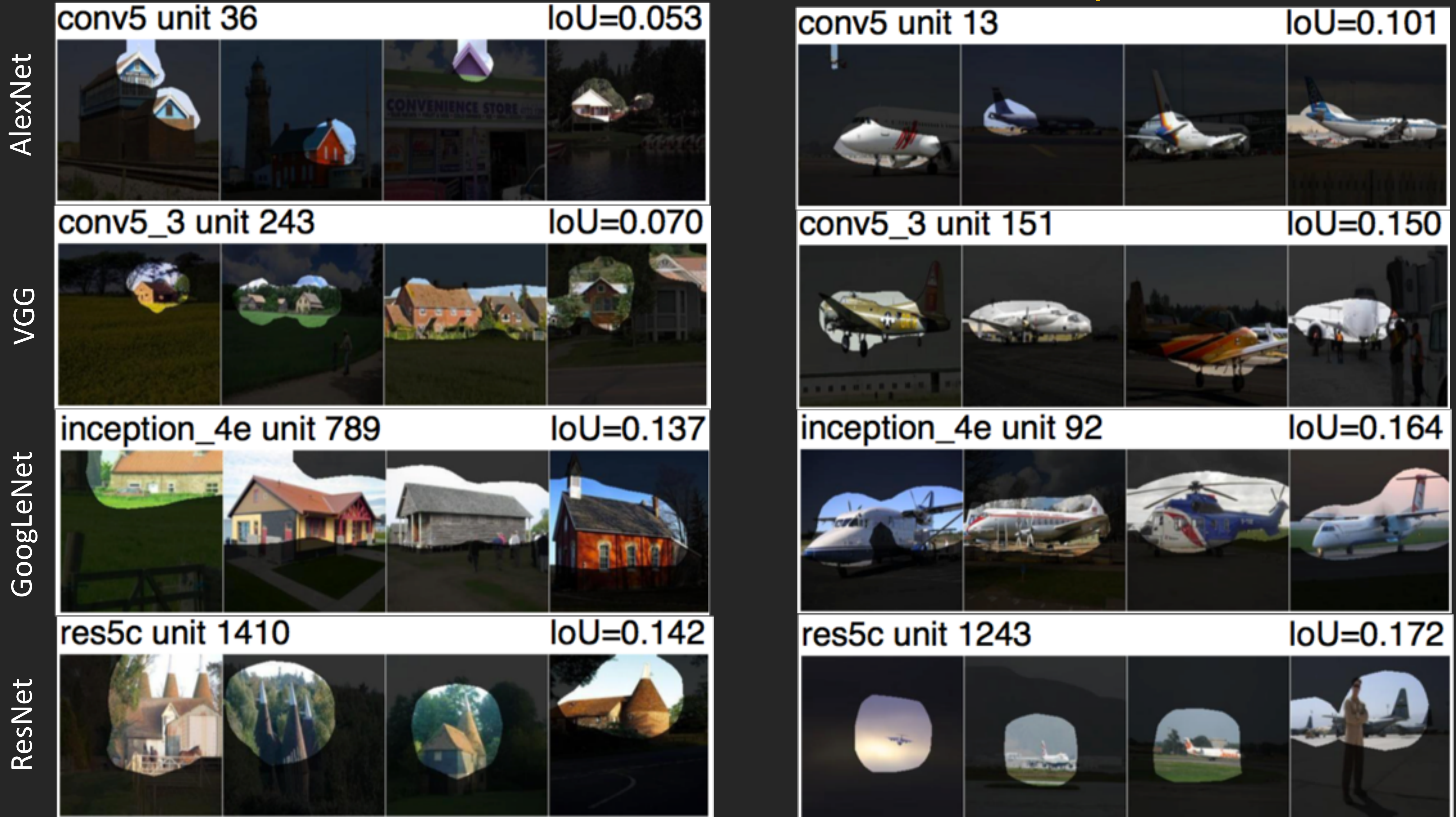
ResNet

Data sources

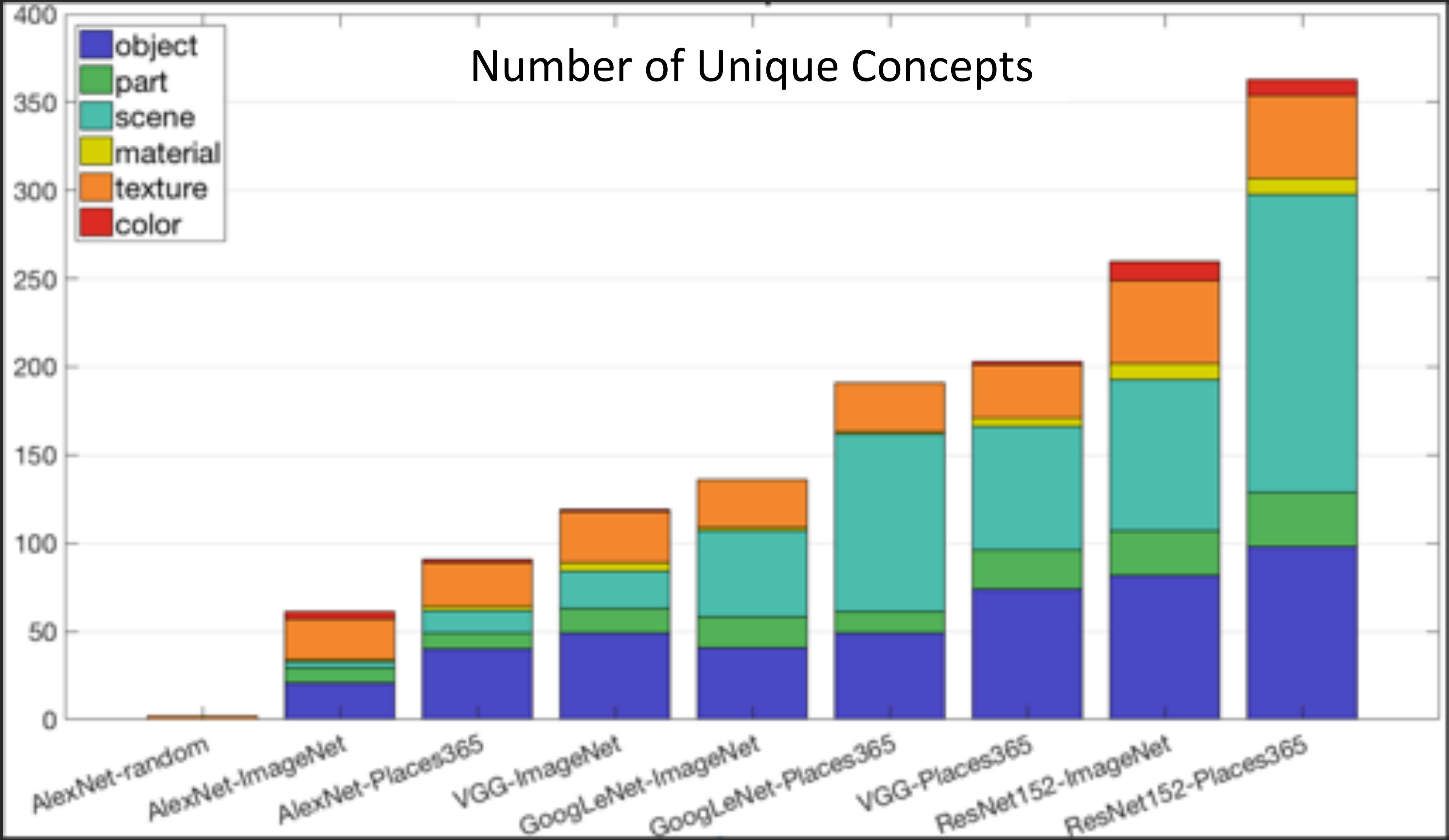
IMAGENET

places

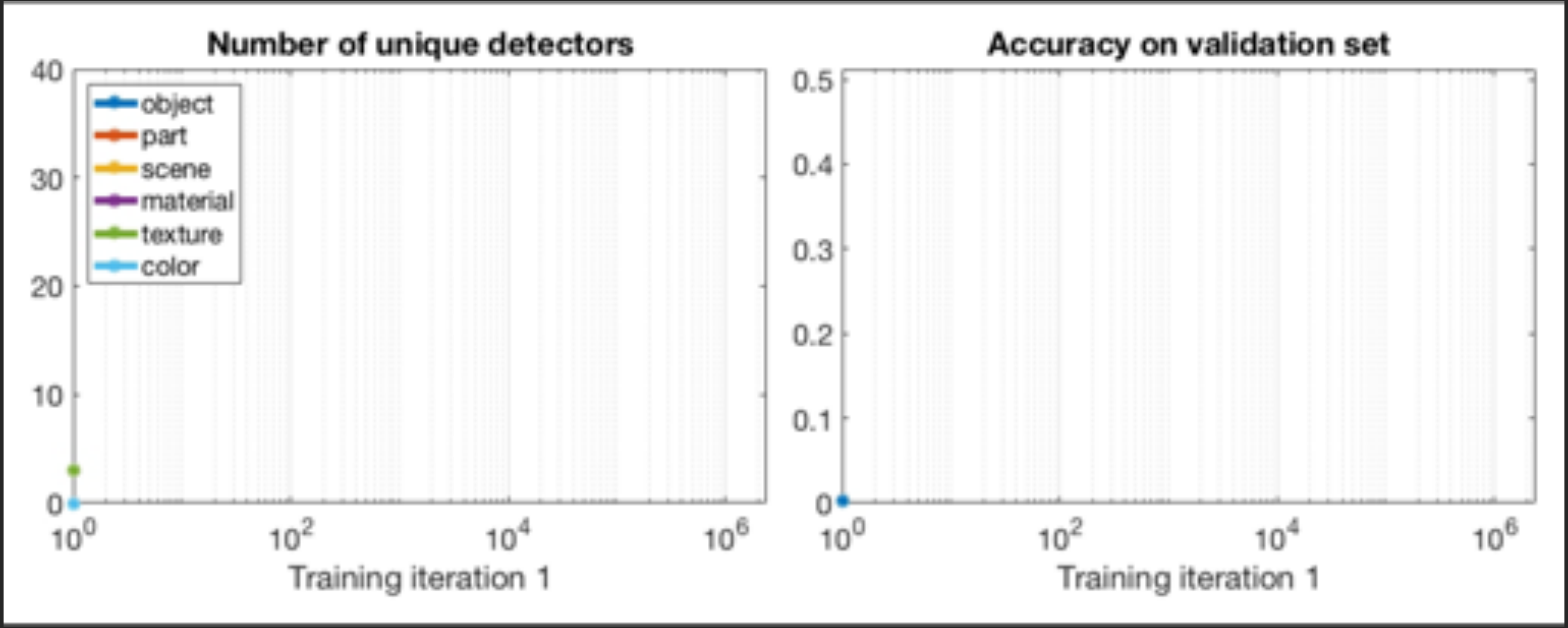
House Airplane


AlexNet
conv5 unit 36 IoU=0.053
conv5 unit 13 IoU=0.101
VGG
conv5_3 unit 243 IoU=0.070
conv5_3 unit 151 IoU=0.150
GoogLeNet
inception_4e unit 789 IoU=0.137
inception_4e unit 92 IoU=0.164
ResNet
res5c unit 1410 IoU=0.142
res5c unit 1243 IoU=0.172
CONVENIENCE STORE

Number of Unique Concepts

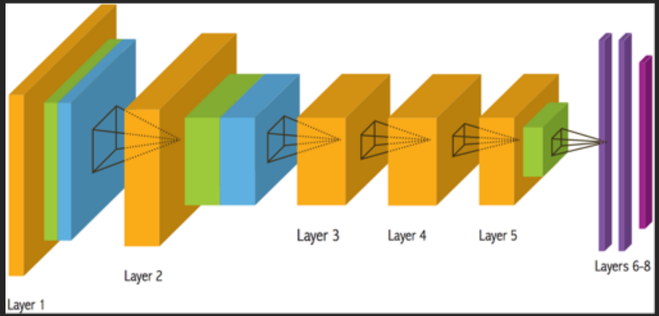


What Happens During the Training?



Transfer Learning across Datasets

Pretrained Network



Fine-Tuning

Target Dataset

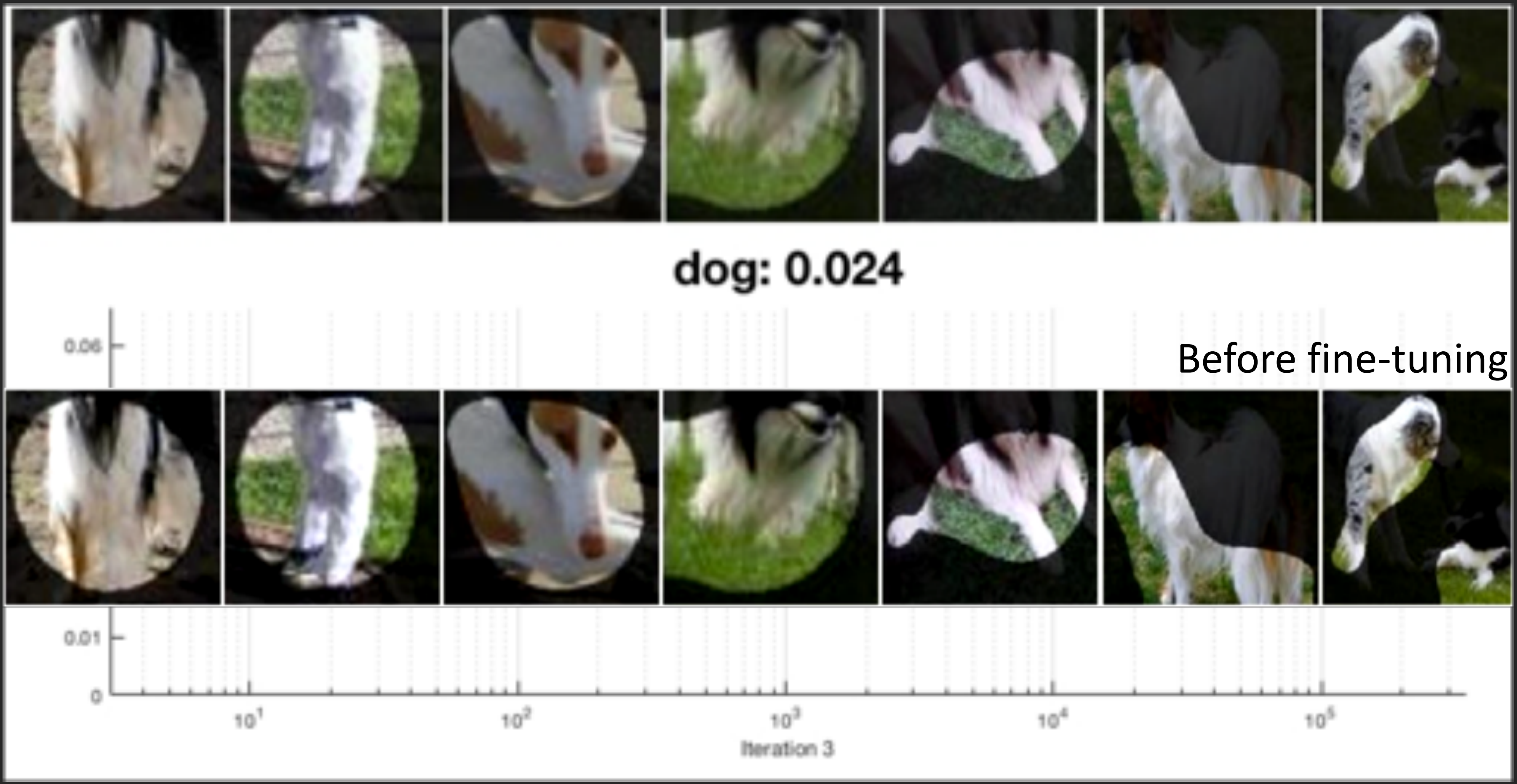
IMAGENET

Pretrained Network

Fine-Tuning

places

Unit 8 at Layer 5 layer

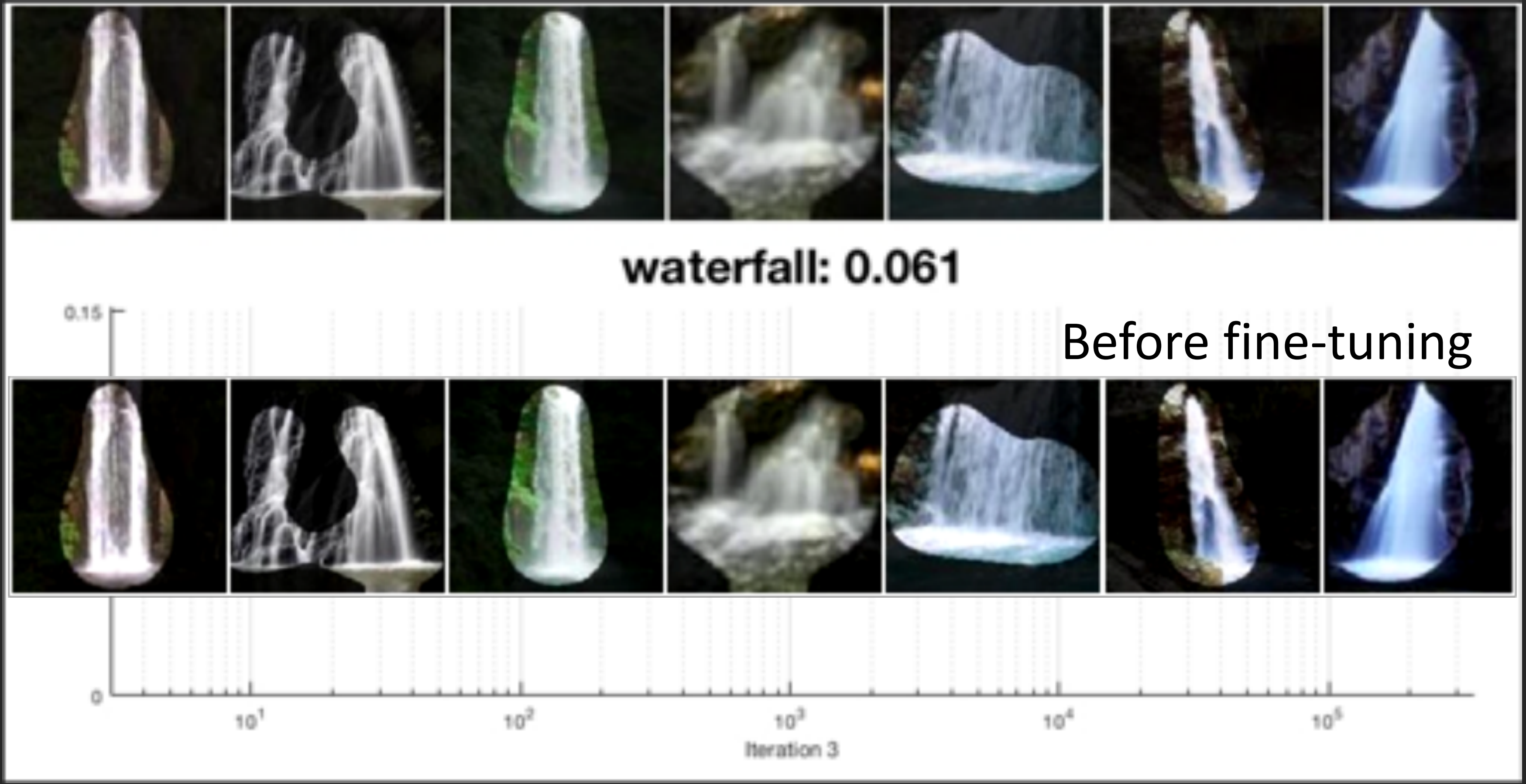


Before fine-tuning

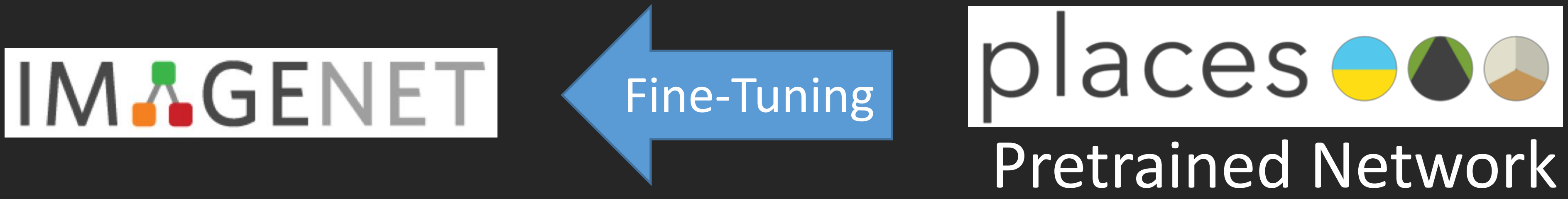
IMAGENET ← Fine-Tuning ← places

Pretrained Network

Unit 35 at Layer 5 layer



Before fine-tuning



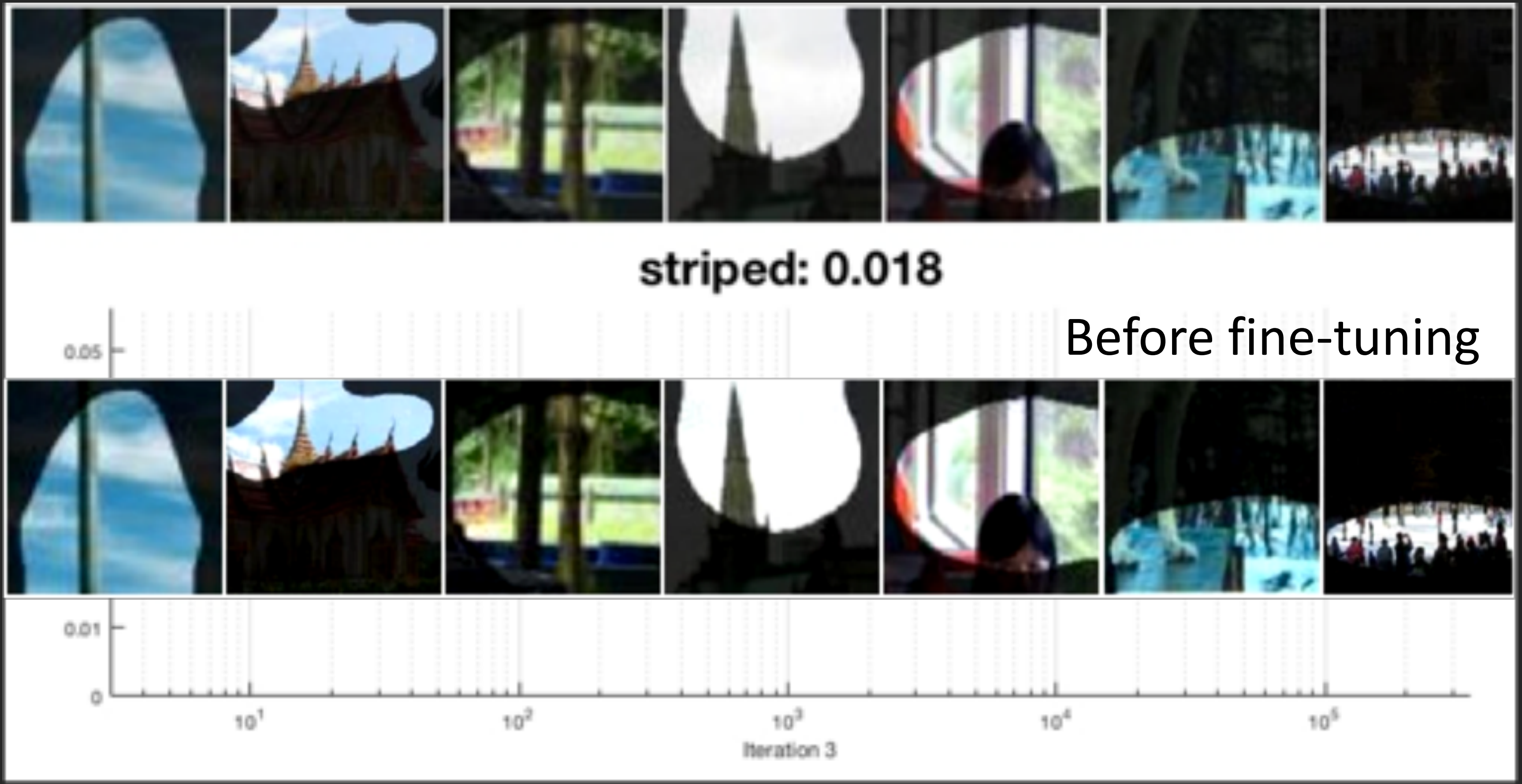
IMAGENET

Fine-Tuning

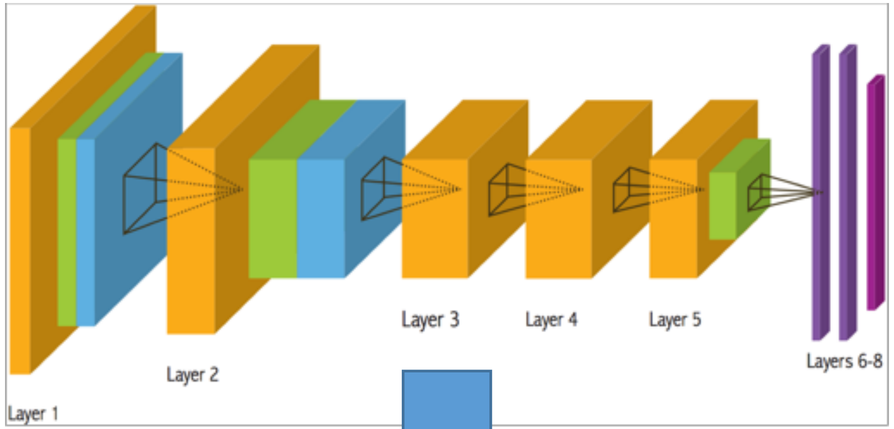
places

Pretrained Network

Unit 103 at Layer 5 layer



Internal Units and Final Prediction



Cafeteria (0.9)

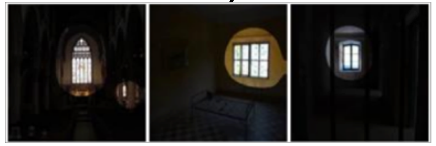
Interpretable units as concept detectors

Unit2 at Layer4: Lamp

Unit 22 at Layer 5: Face

Unit42 at Layer3 : Trademark

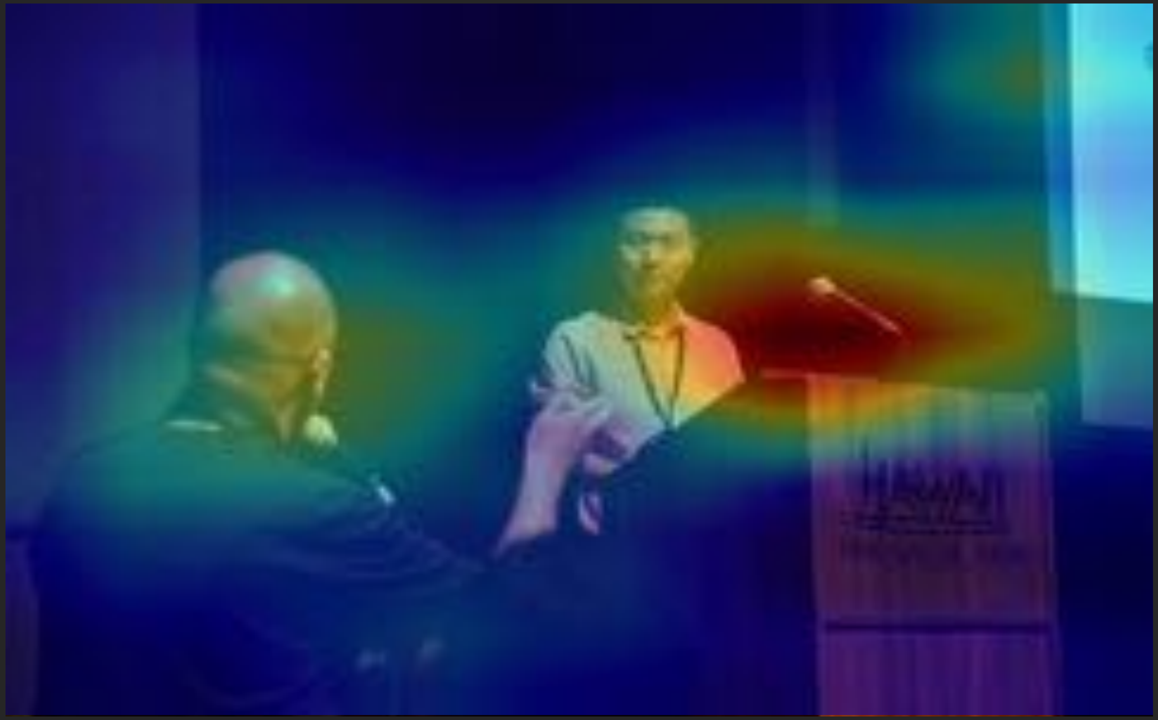
Unit 57 at Layer4: Windows



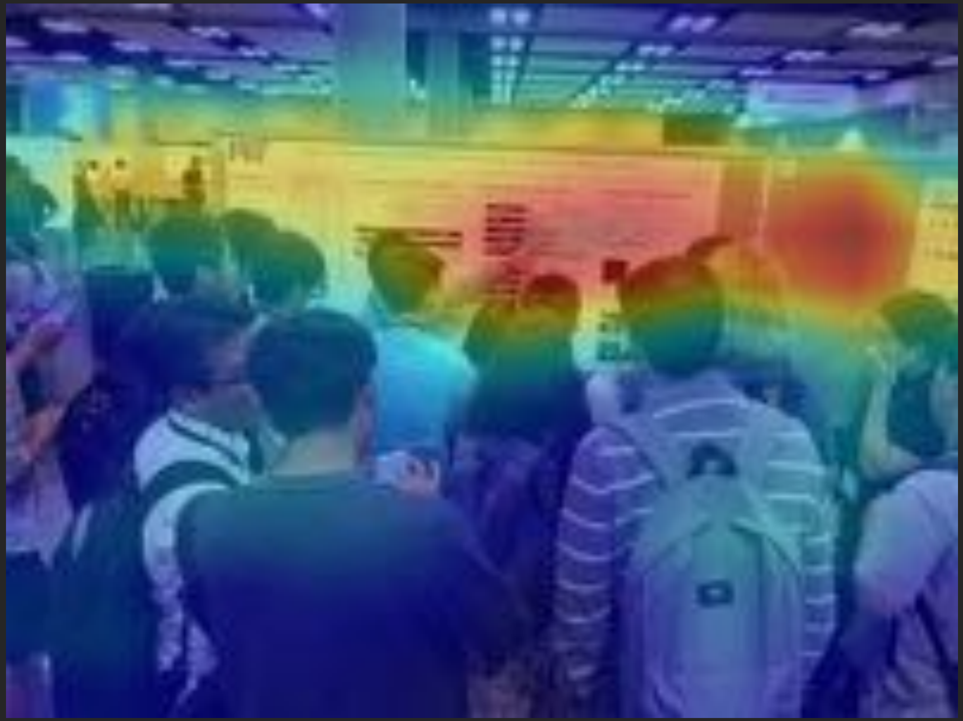
Why this prediction?

Class Activation Mapping:
Explain Prediction of Deep Neural Network

Prediction: Conference Center



Prediction: Indoor Booth



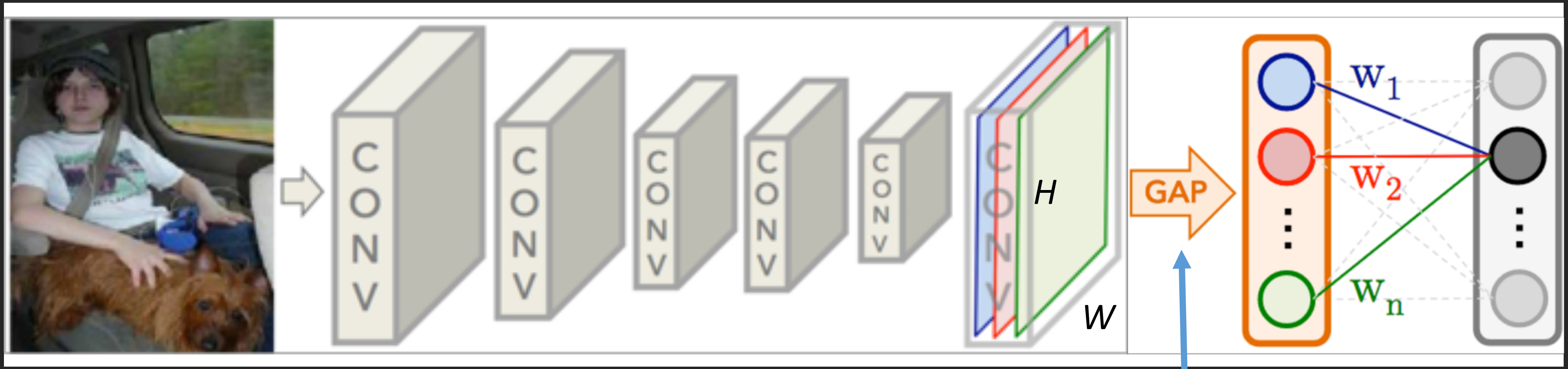
[**Zhou**, Khosla, Lapedriza, Oliva, Torralba. CVPR 2016]

Unit Activation Maps

$f_k(h,w)$

Class prob.

y_c



Dog: 0.8

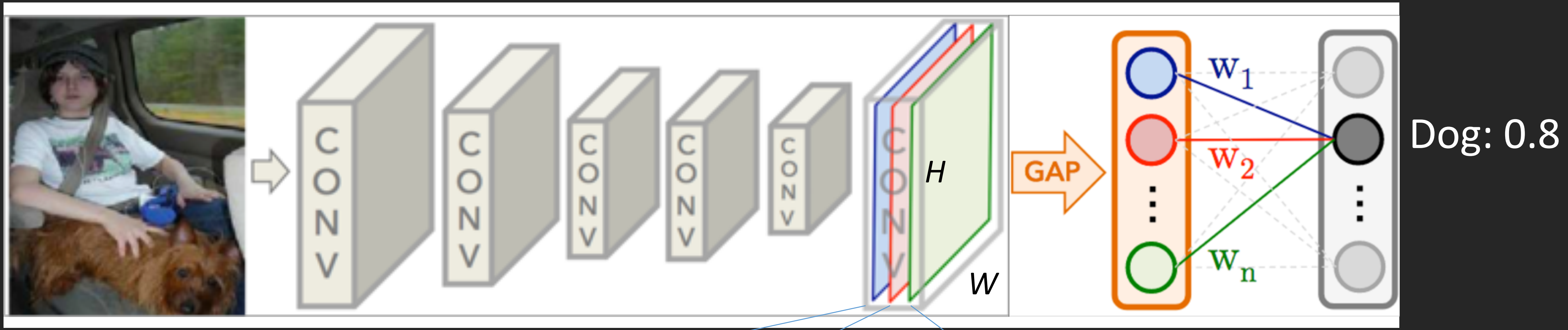
Global Average Pooling (GAP)

$$\frac{1}{HW}\sum_{h,w} f_k(h,w)$$

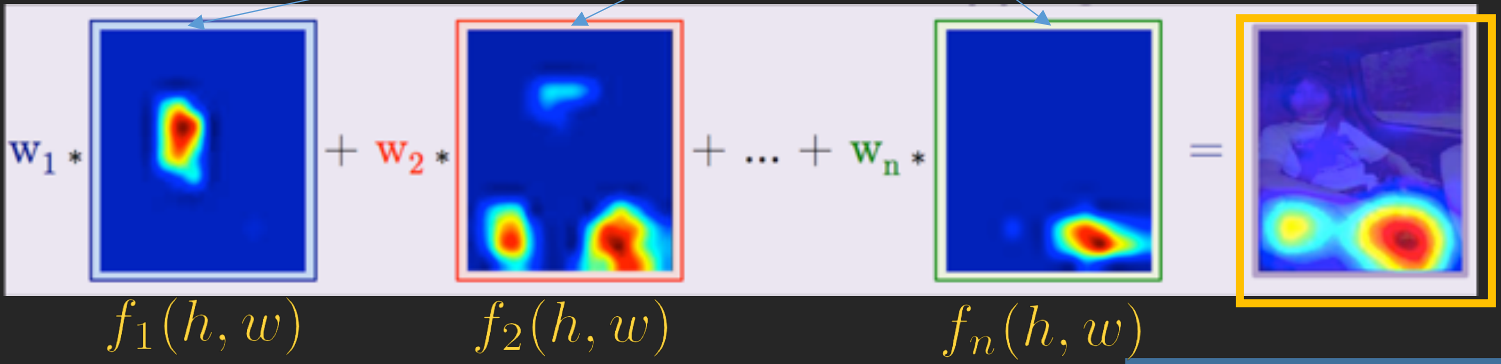
$$y_c \propto \sigma\Big(\sum_k w_k^c \sum_{h,w} f_k(h,w)\Big)$$

Unit Activation Maps $f_k(h, w)$

Class prob. y_c



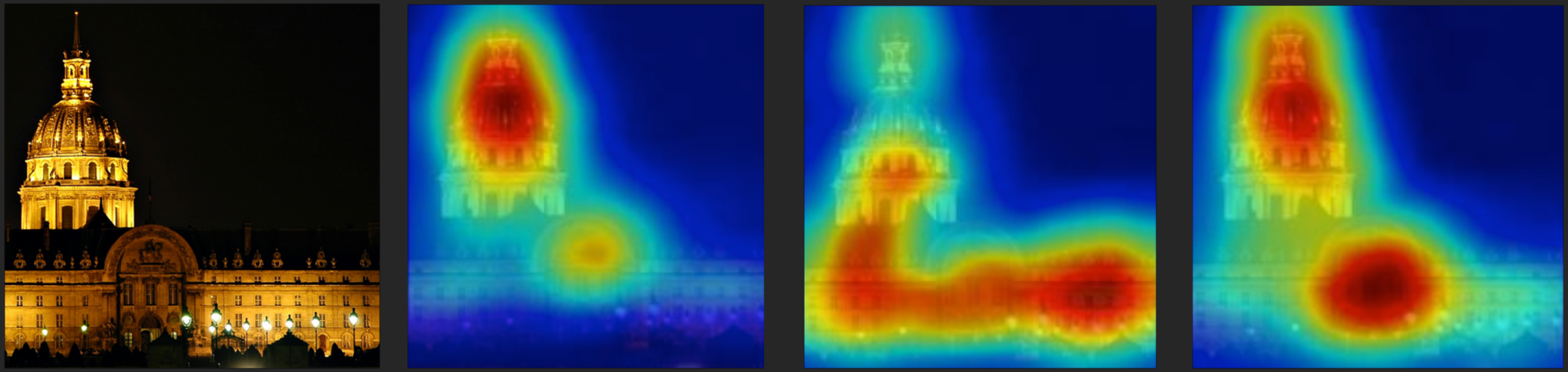
Class Activation Map



$$y_c \propto \sigma\Big(\sum_k w_k^c \sum_{h,w} f_k(h, w)\Big) = \sigma\Big(\sum_{h,w} \boxed{\sum_k w_k^c f_k(h, w)}\Big)$$

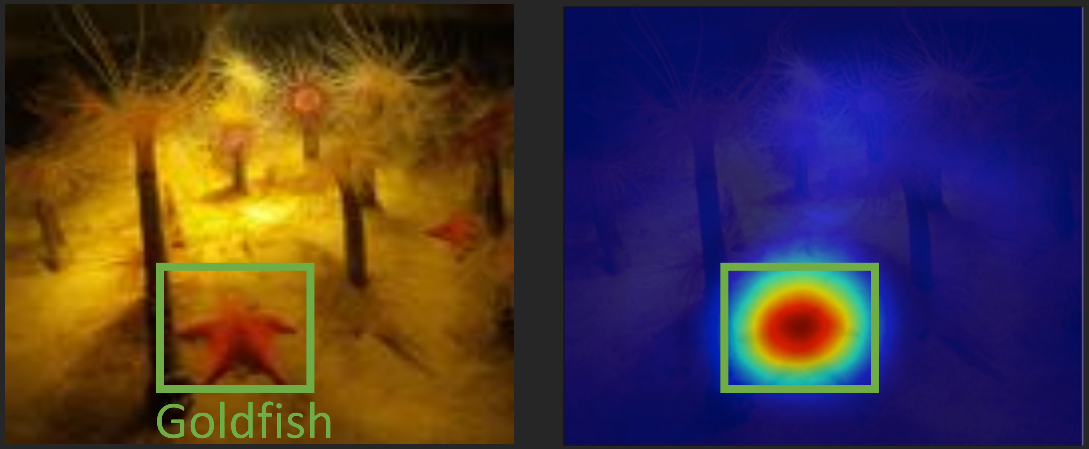
Class Activation Mapping:
Explain Prediction of Deep Neural Network

Top3 Predictions: Dome (0.45) Palace (0.21) Church (0.10)

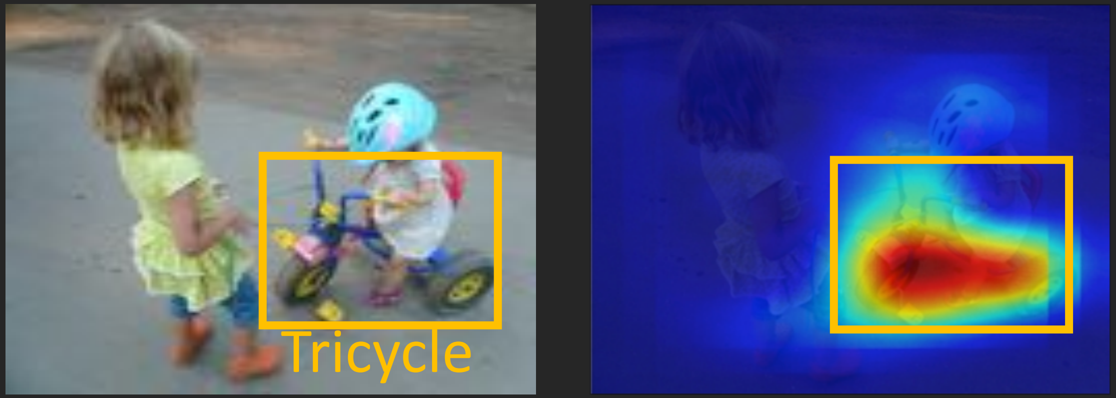


Evaluation on Weakly-Supervised Localization

Prediction: Starfish (0.83)



Prediction: Tricycle (0.92)

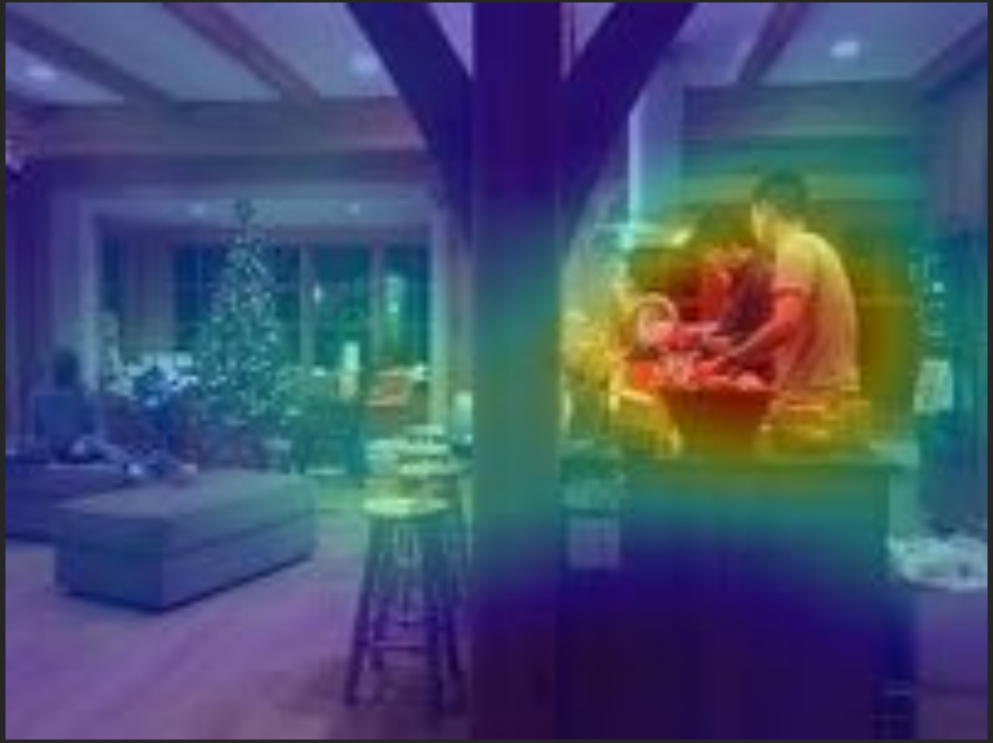


Method	Supervision	Localization Accuracy(%)
Backpropagation	weakly	53.6
Our method	weakly	**62.9**
AlexNet	full	65.8

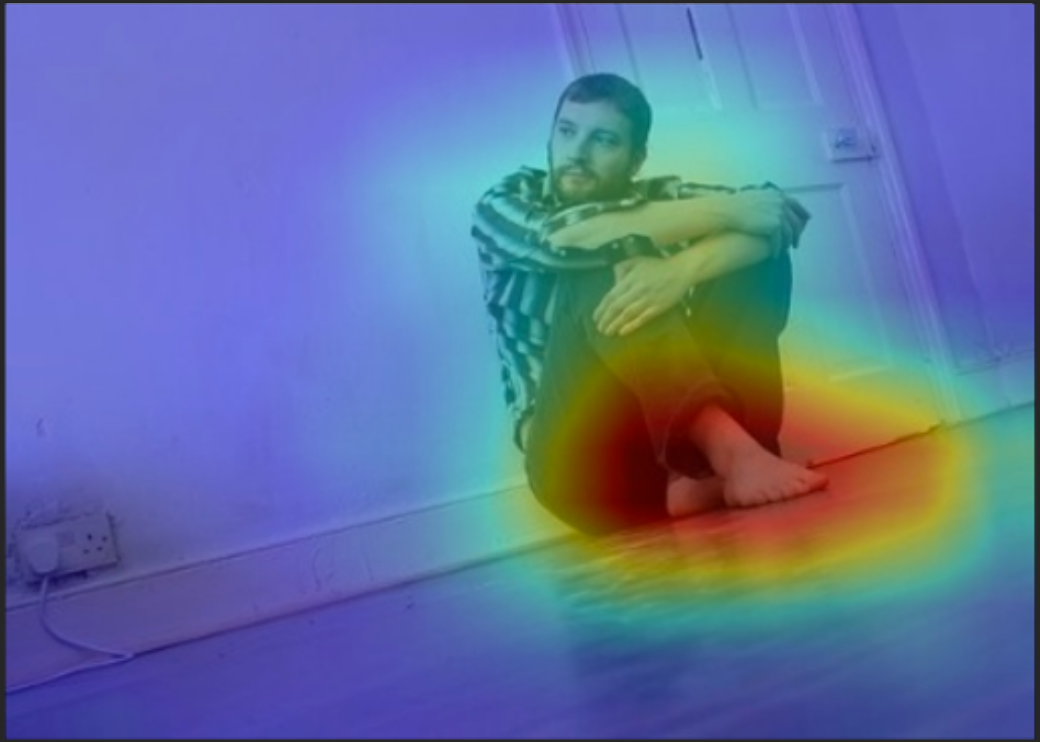
Result on ImageNet Localization Benchmark

Explaining the Failure Cases

Prediction: Sushi Bar (0.63)

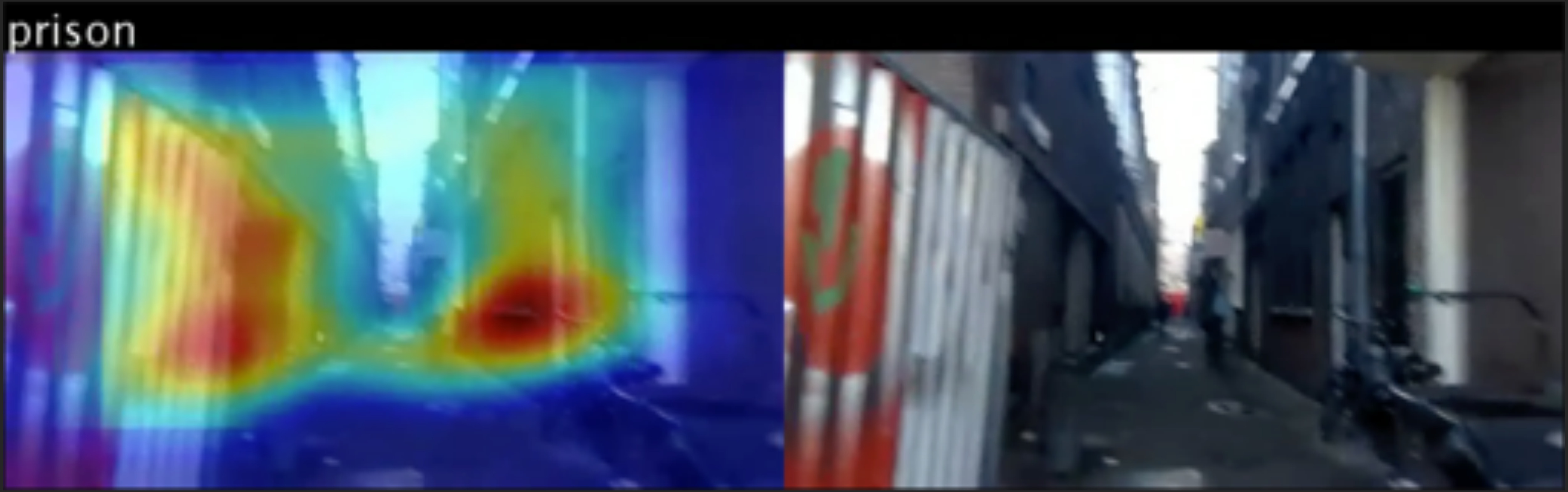


Prediction: Martial Arts Gym (0.21)



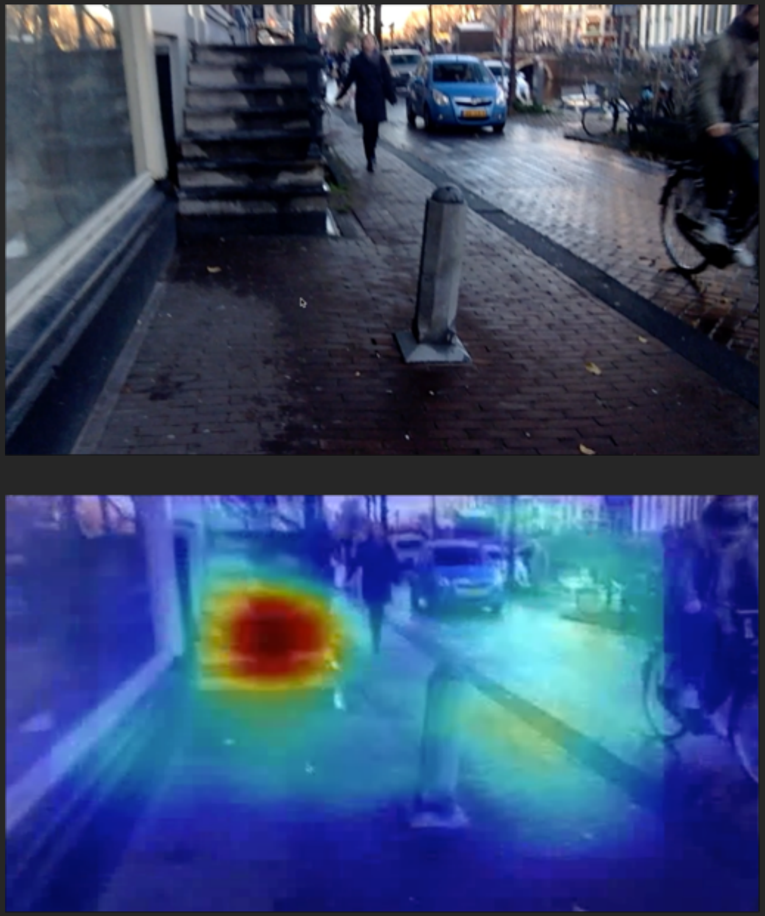
Explaining the Failure Cases in Video

Predictions from a model pretrained on ImageNet

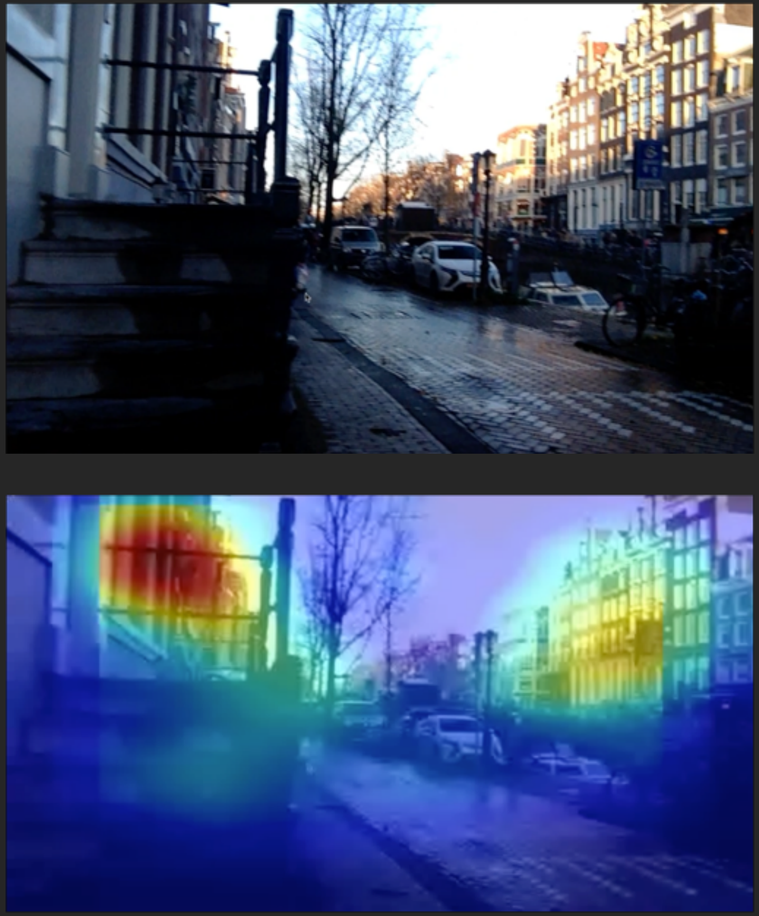


Explaining the Failure Cases

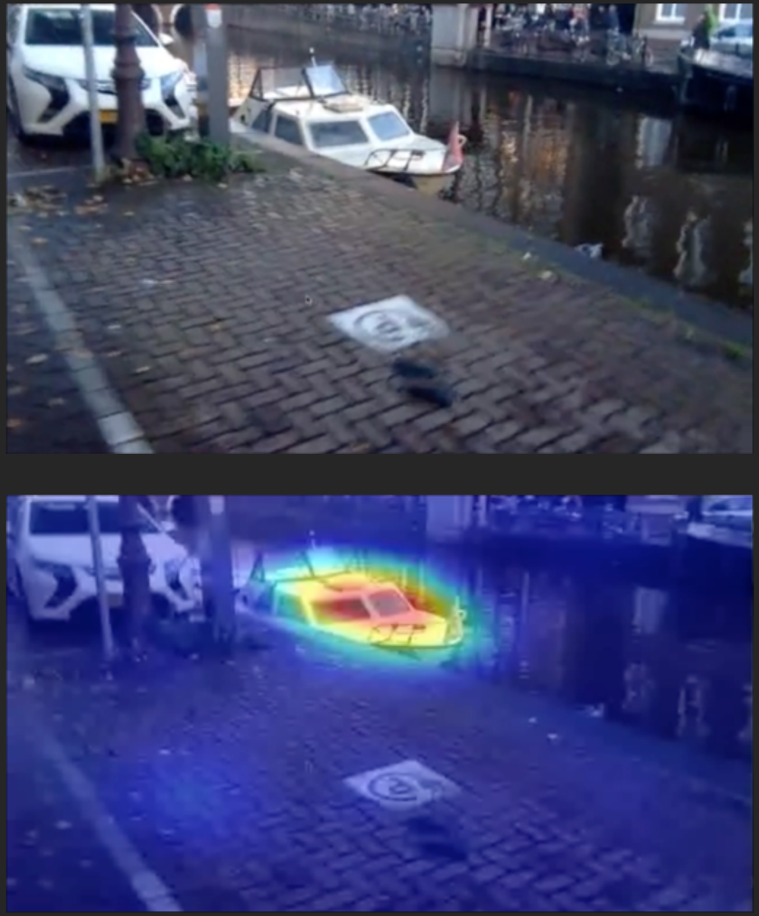
Prediction: Park bench



Prediction: Prison

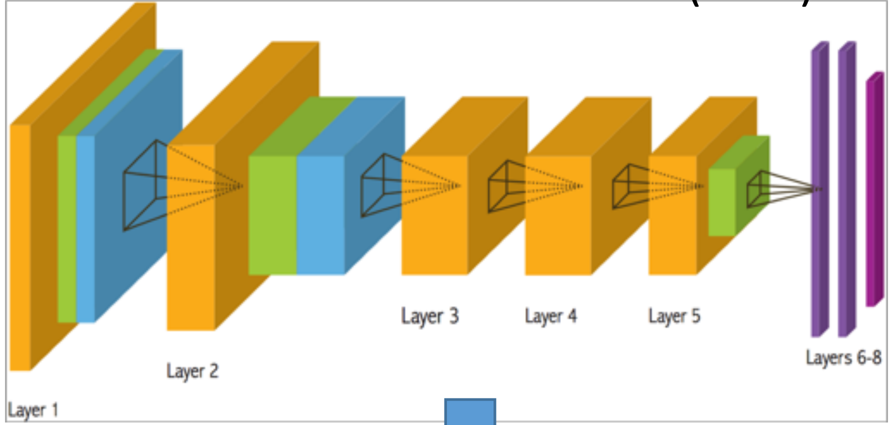


Prediction: Aircraft carrier



Interpretable Representation for Classifying Scenes

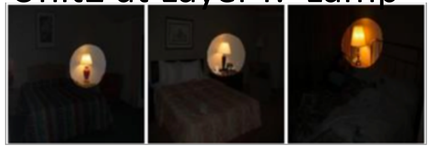
Convolutional Neural Network (CNN)



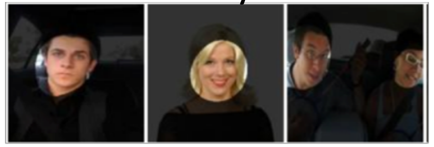
Cafeteria (0.9)

Units as object detectors

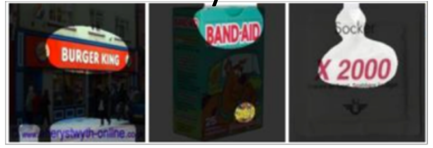
Unit2 at Layer4: Lamp



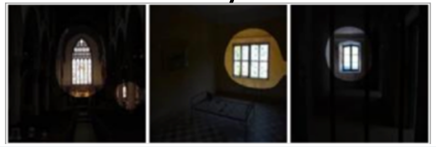
Unit 22 at Layer 5: Face



Unit42 at Layer3 : Trademark

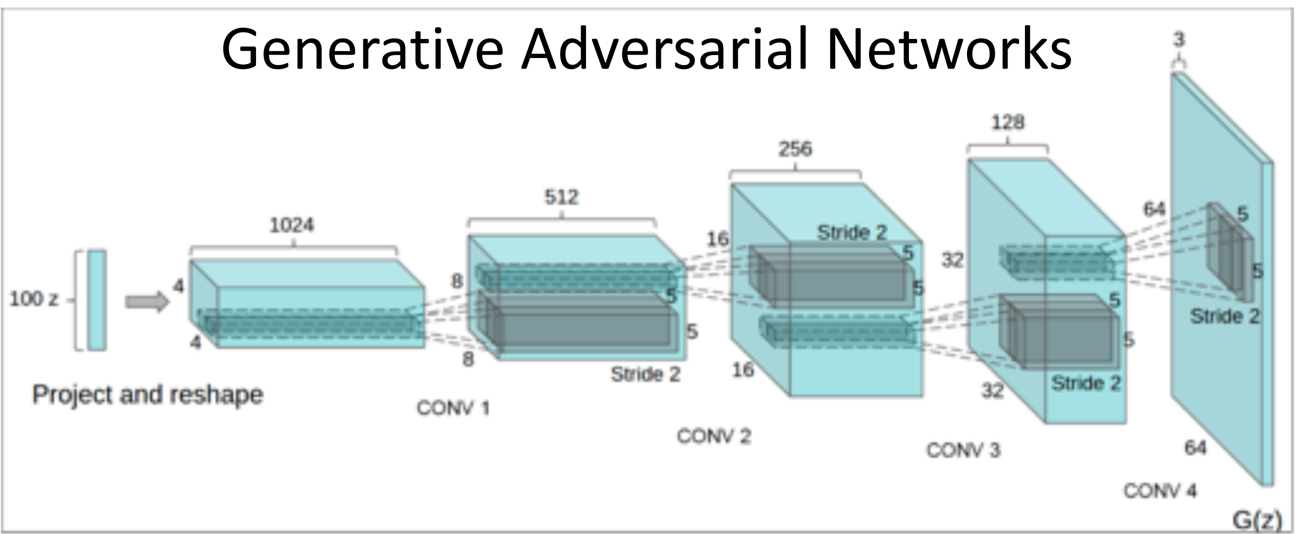


Unit 57 at Layer4: Windows



Zhou et al, ICLR’15, CVPR’17
TPAMI’18, etc.

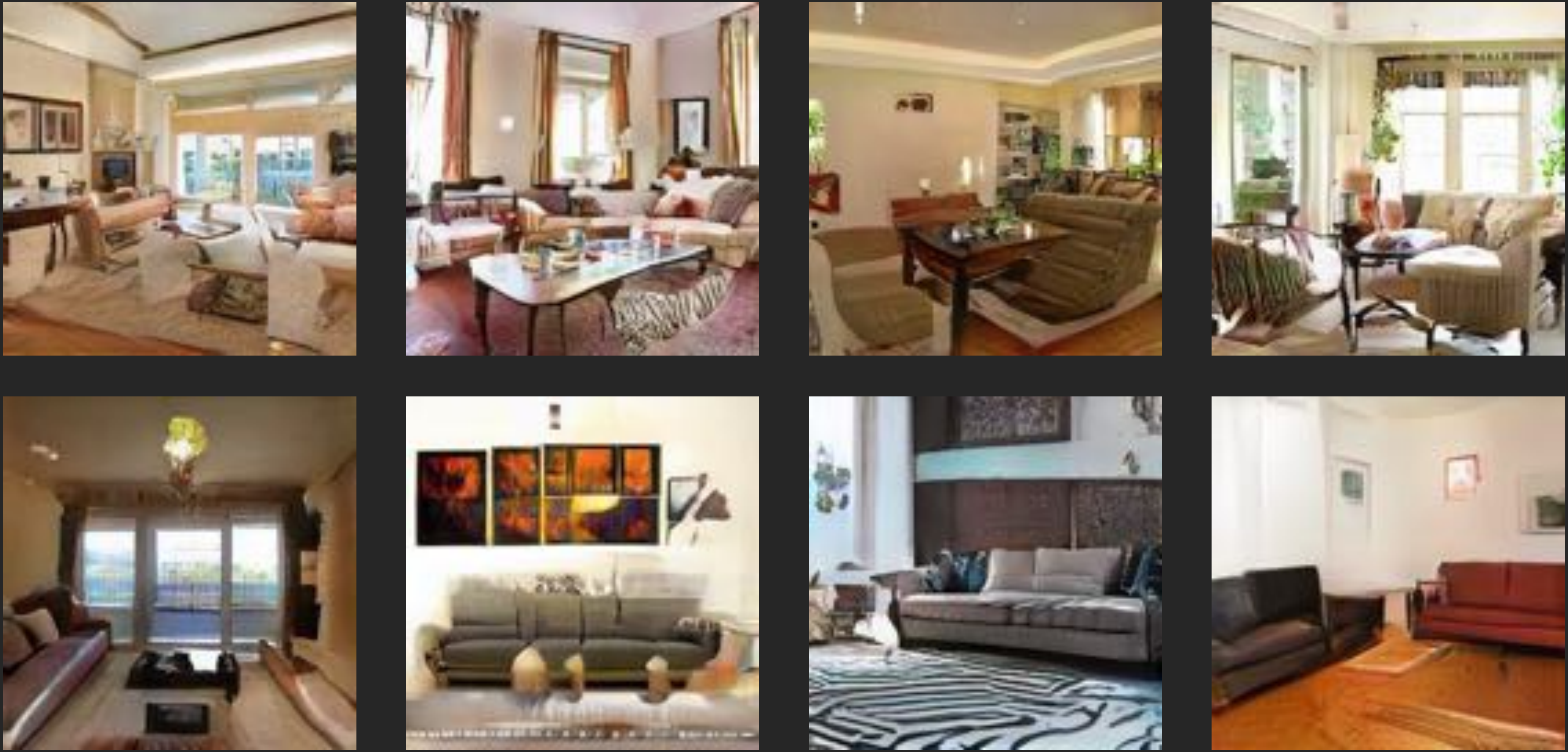
What’s inside the deep generative model?



Generative Adversarial Networks

Goodfellow, et al. NIPS’14
Radford, et al. ICLR’15
T Karras et al. 2017
A. Brock, et al. 2018

They are all synthesized living rooms

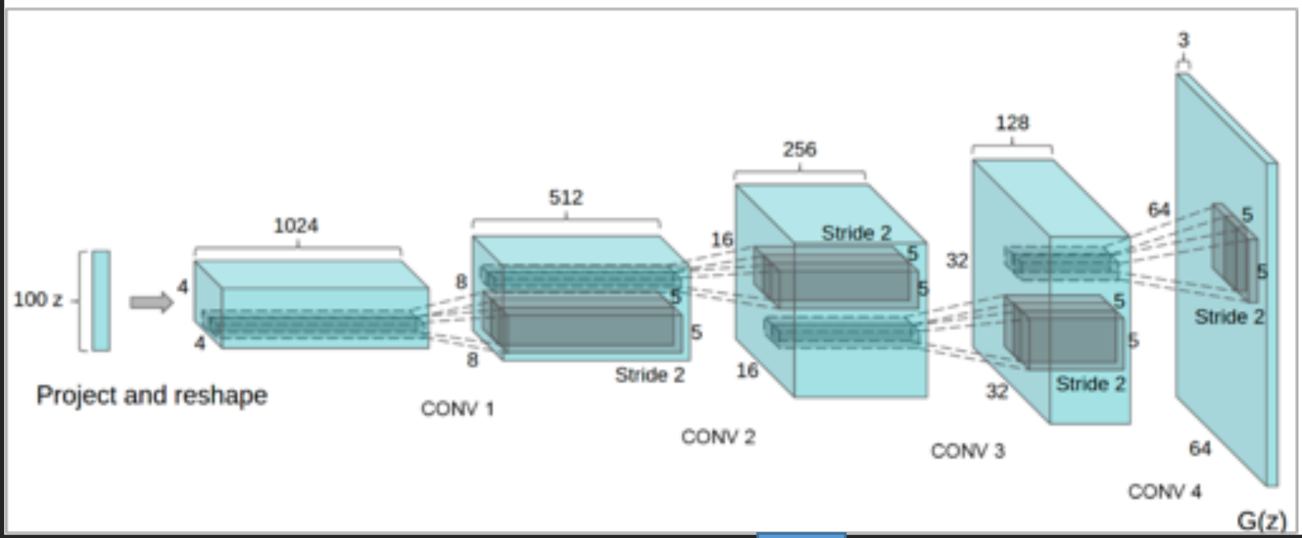


T Karras et al. 2017

Understanding the Internal Units in GANs

Input:
Random noise

Output:
Synthesized image

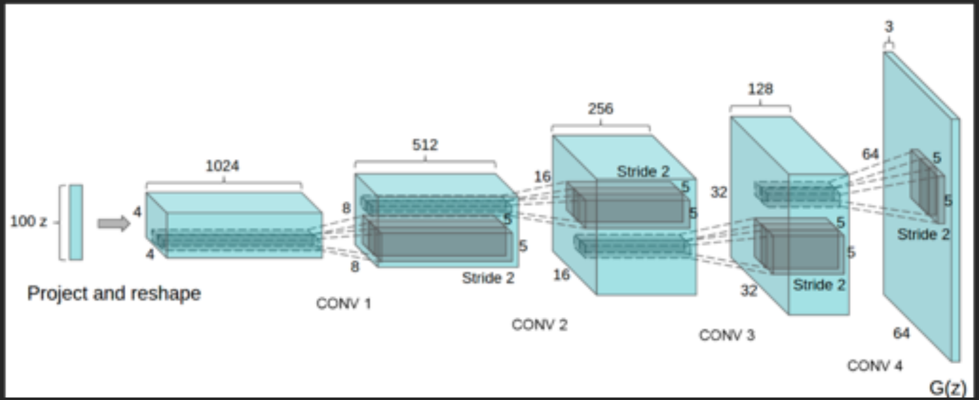


What are they doing?

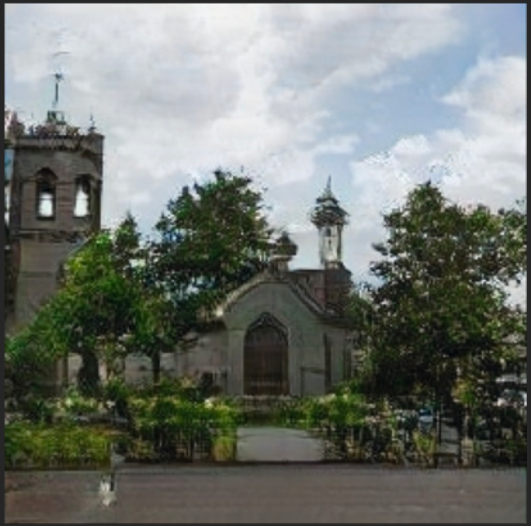
David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, J. Tenenbaum, W. Freeman, A. Torralba.
GAN Dissection: Visualizing and Understanding GANs. ICLR’19. https://arxiv.org/pdf/1811.10597.pdf

More Practical Issue: How to Modify Contents?

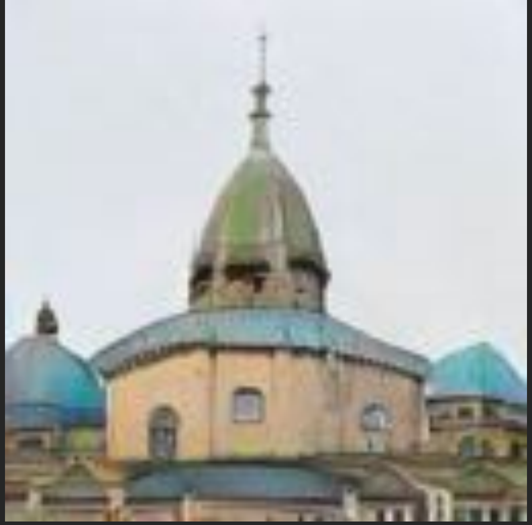
Input:
Random noise



Output:
Synthesized image

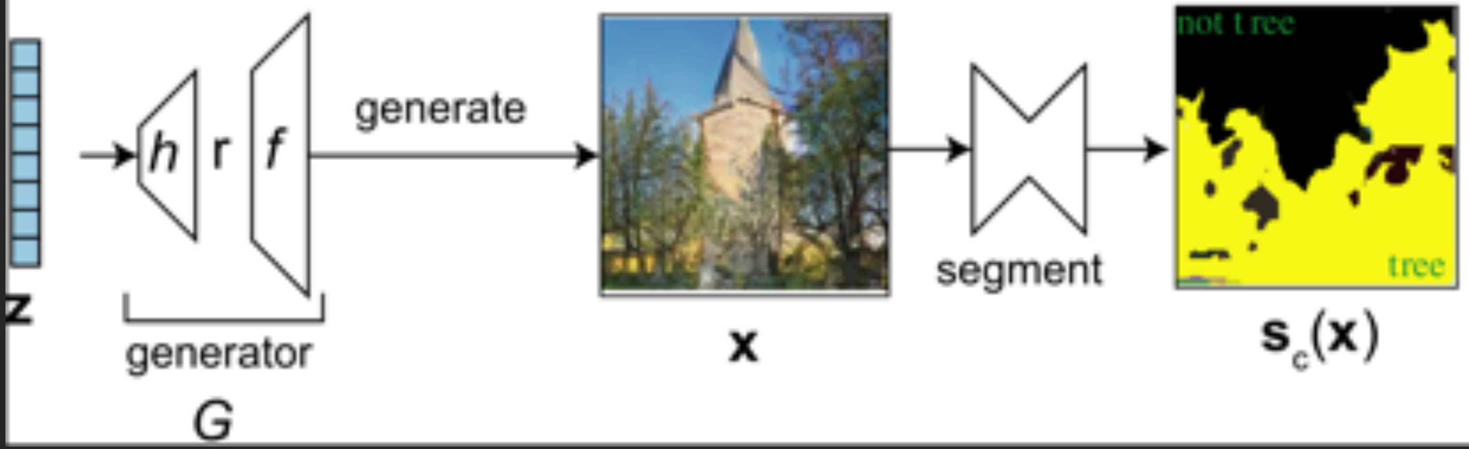


Add trees



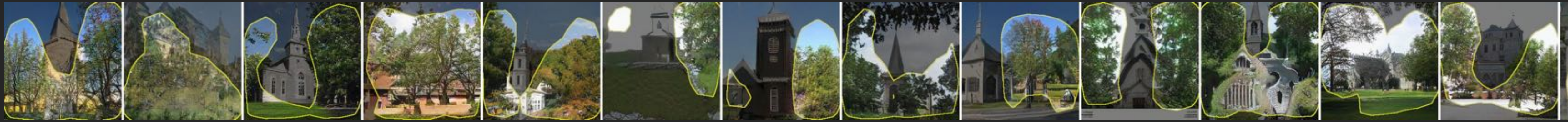
Change dome

Framework of GAN Dissection

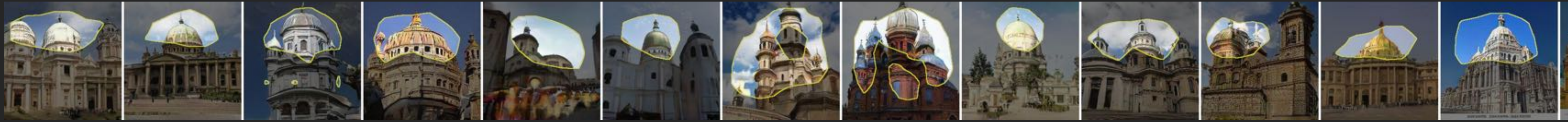


Units Emerge as Drawing Objects

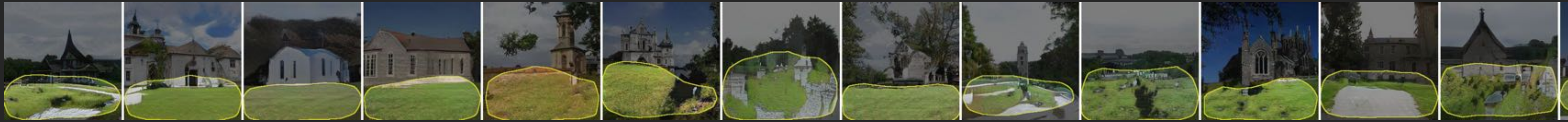
Unit 365 draws trees.



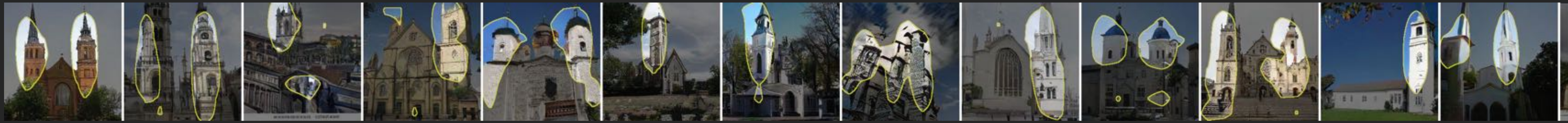
Unit 43 draws domes.



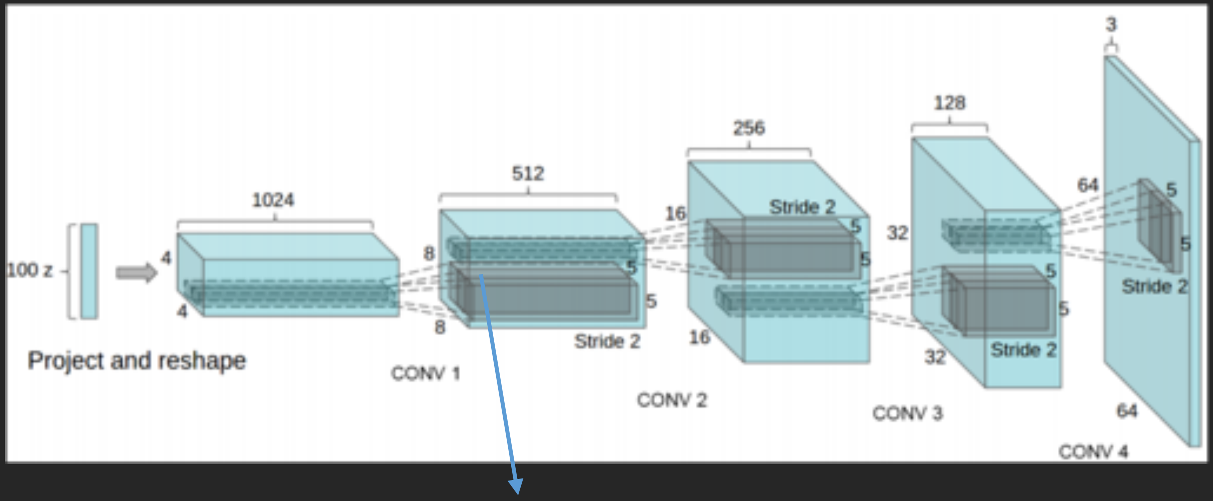
Unit 14 draws grass.



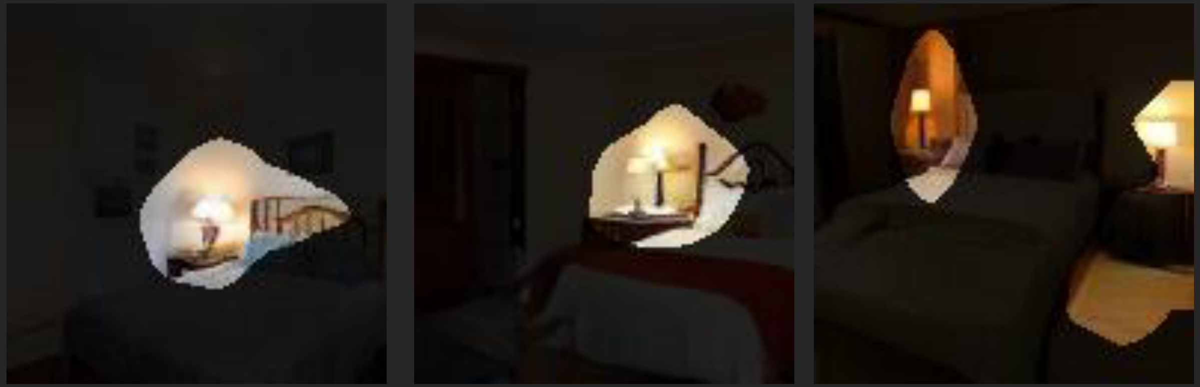
Unit 276 draws towers.



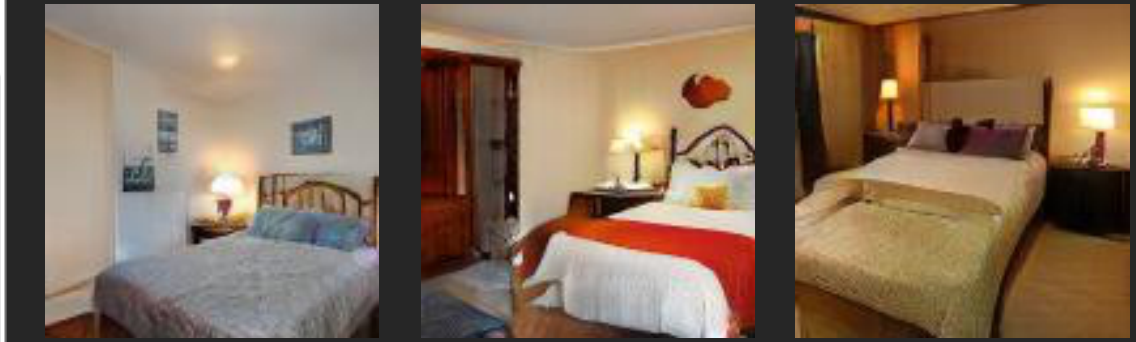
Manipulating the Synthesized Images



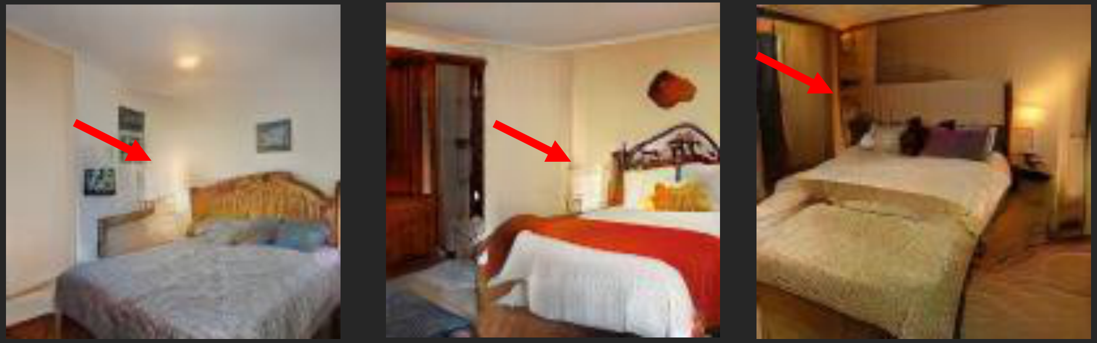
Unit 4 for drawing Lamp



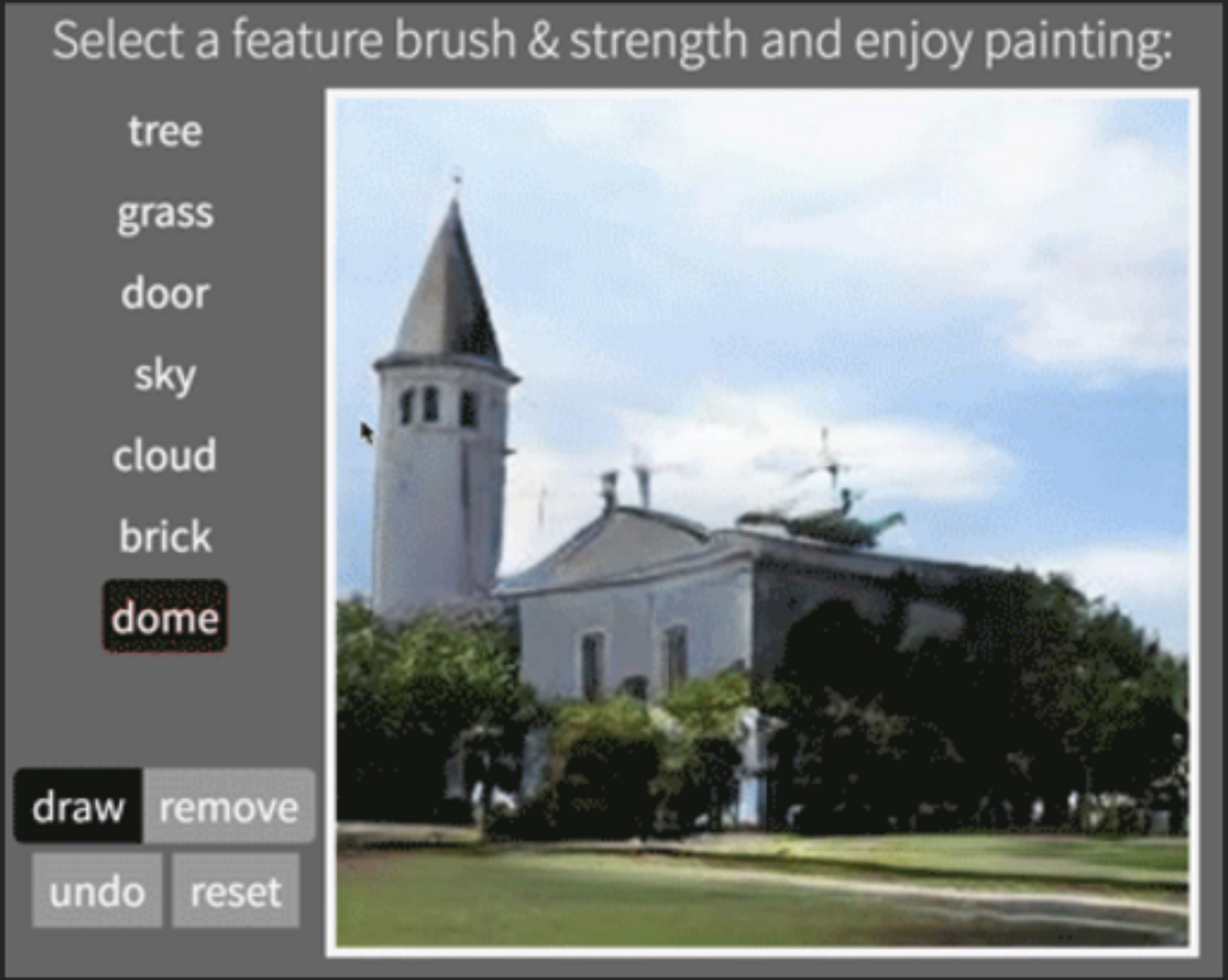
Synthesized Images



Synthesized Images with Unit 4 removed



Interactive Image Manipulation



Select a feature brush & strength and enjoy painting:

Code and paper are at
http://gandissect.csail.mit.edu

Why Care About Interpretability?

‘Alchemy’ of Deep Learning

Scientific Understanding

‘Chemistry’ of Deep Learning