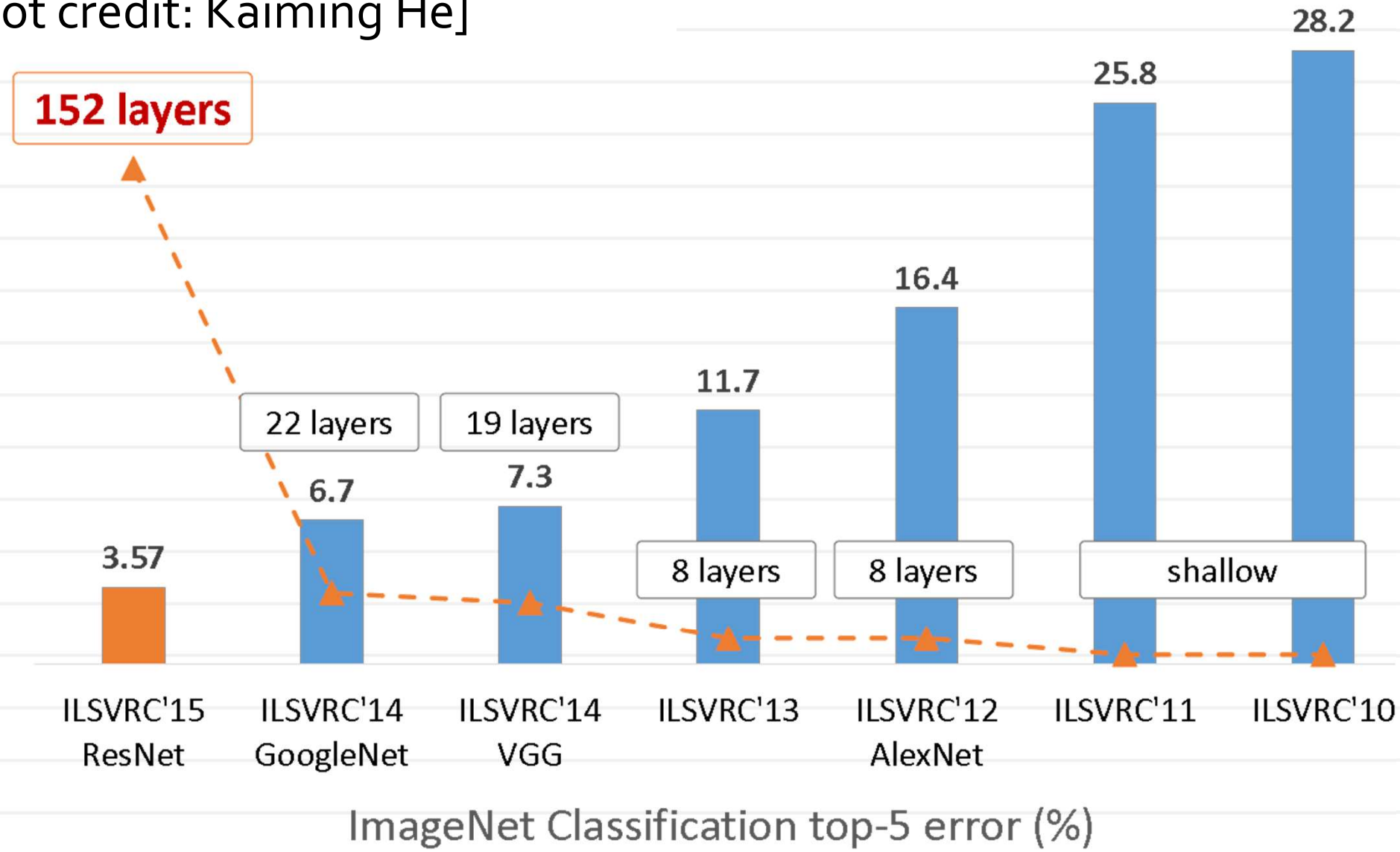
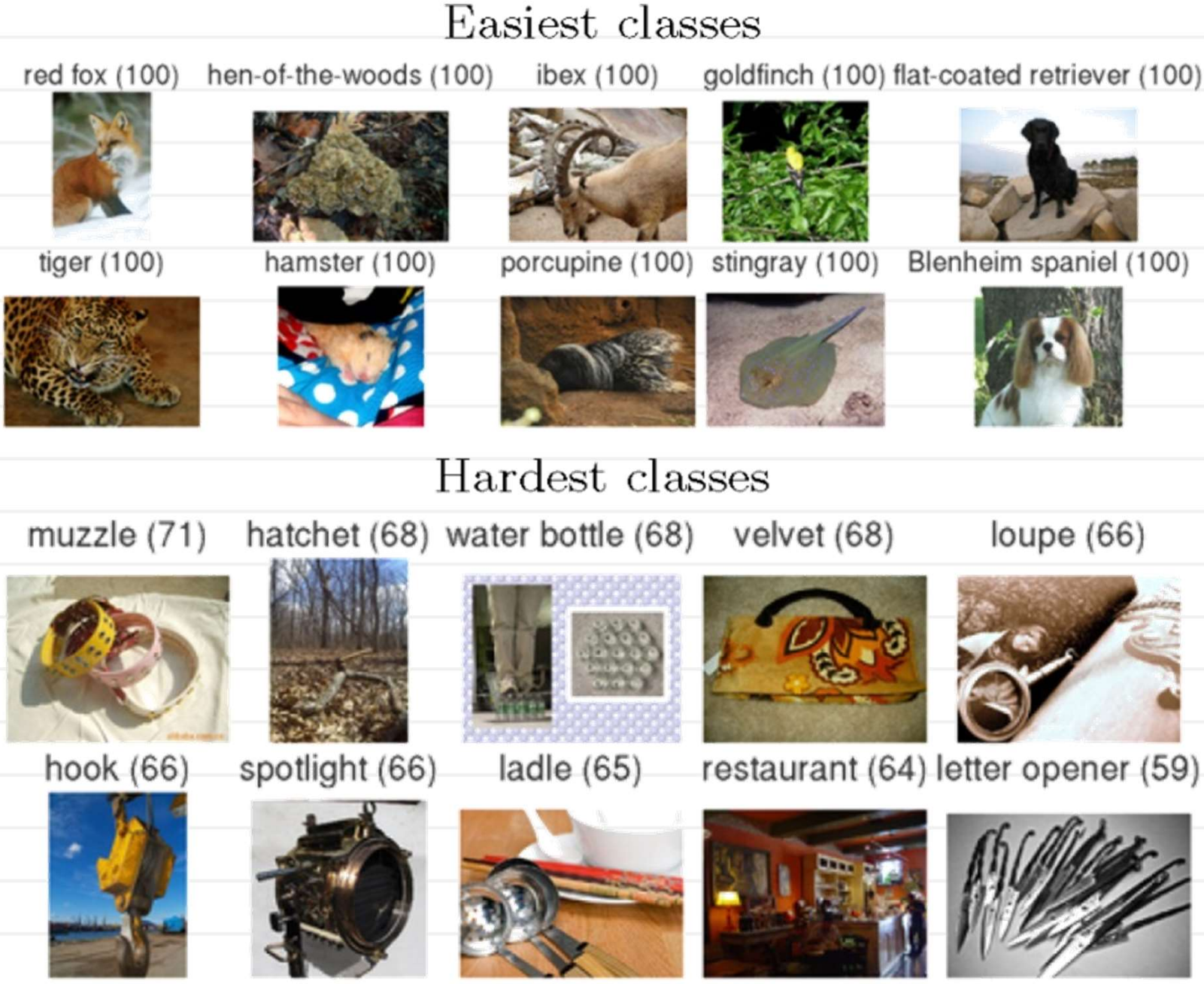
# Lecture 5: Representations within ConvNets

# Building the best network

[plot credit: Kaiming He]

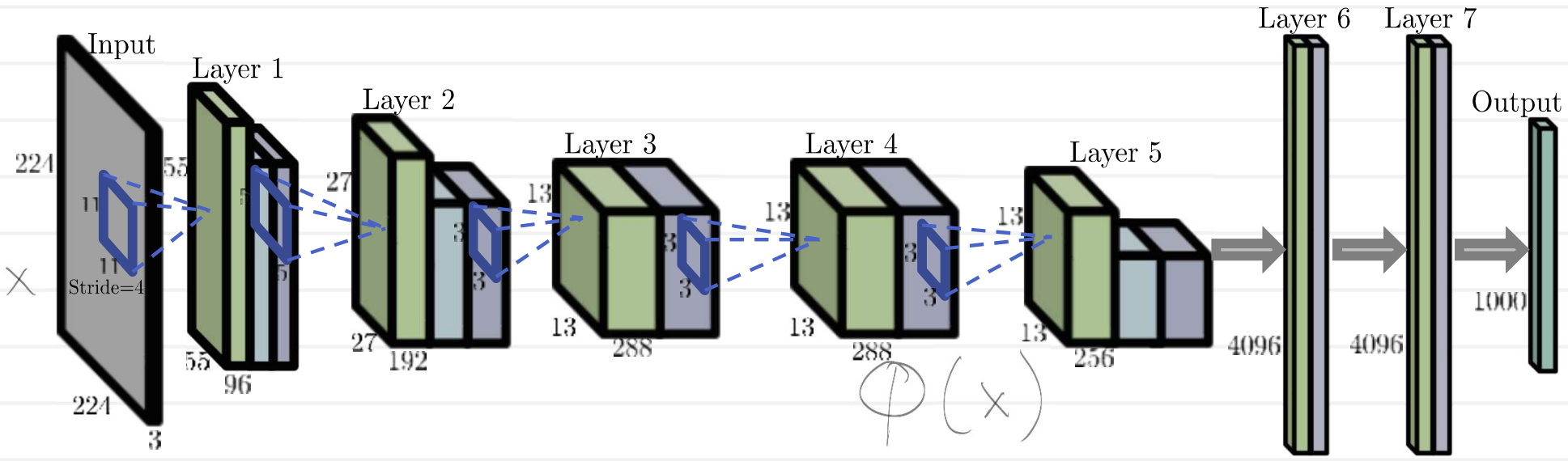


# Easy and hard cases




[Russakovsky et al. 2014]

# Representations inside the neural network

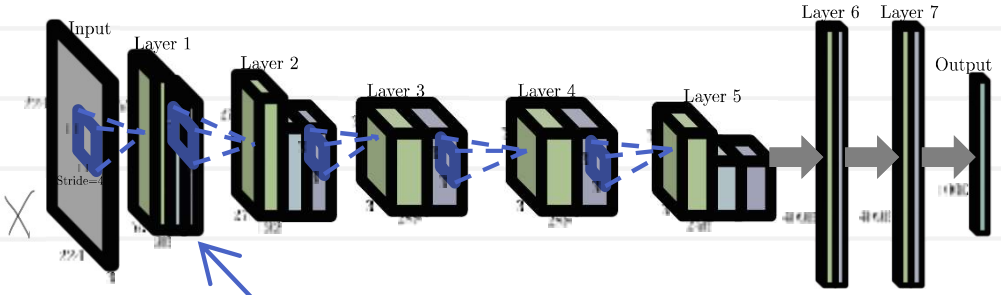



Lots of important questions:

- What are their properties?
- Are they redundant?
- Are they invertible?
- Are the intermediate representations useful?

# Pattern sensitivity

Types of patterns in each layer:




$\varphi_1(x)$

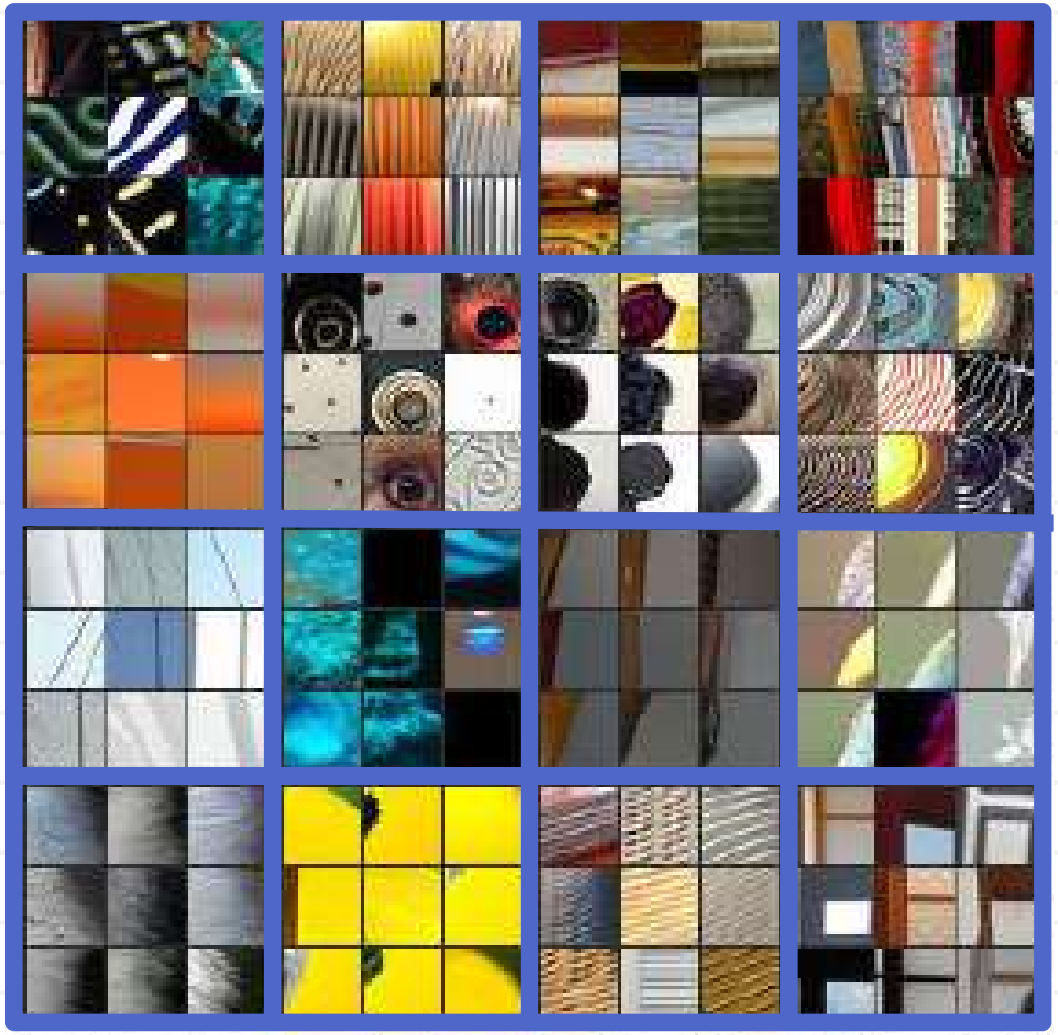
$\text{argmax}_{x \in S} \varphi_1(x)^{k,l} = ?$

[Zeiler Fergus 14]

# Pattern sensitivity

Types of patterns in each layer:

Layer 1



Layer 2

[Zeiler Fergus 14]

# Pattern sensitivity

Types of patterns in each layer:

[Zeiler Fergus 14]

Layer 3

# Pattern sensitivity

Types of patterns in each layer:

Layer 4

Layer 5

[Zeiler Fergus 14]

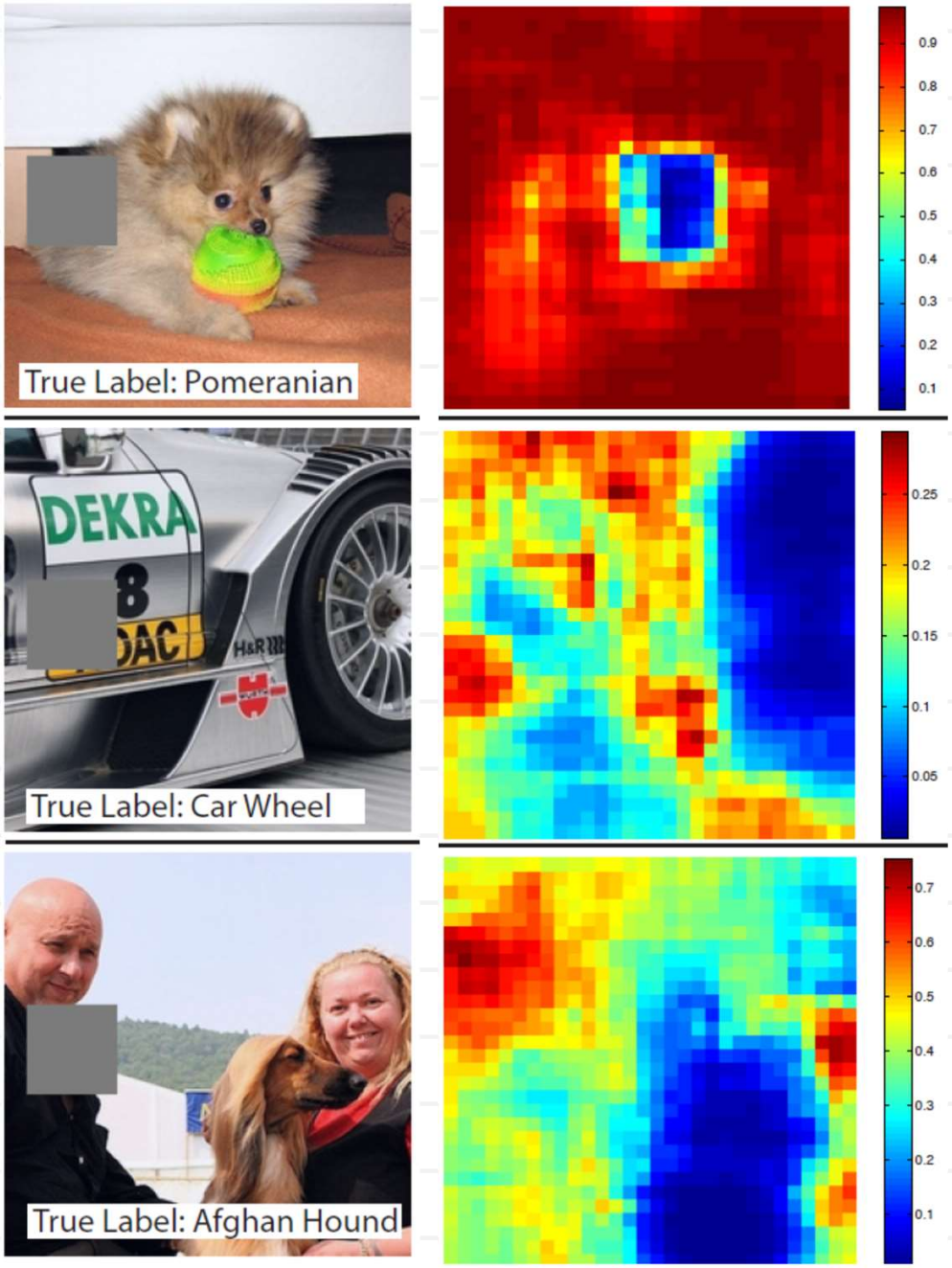
# Meaningful units or meaningful space?

Max activations for random units:

Max activations for random directions:

[Szegedy et al. 2014]

# Grounding CNN decisions

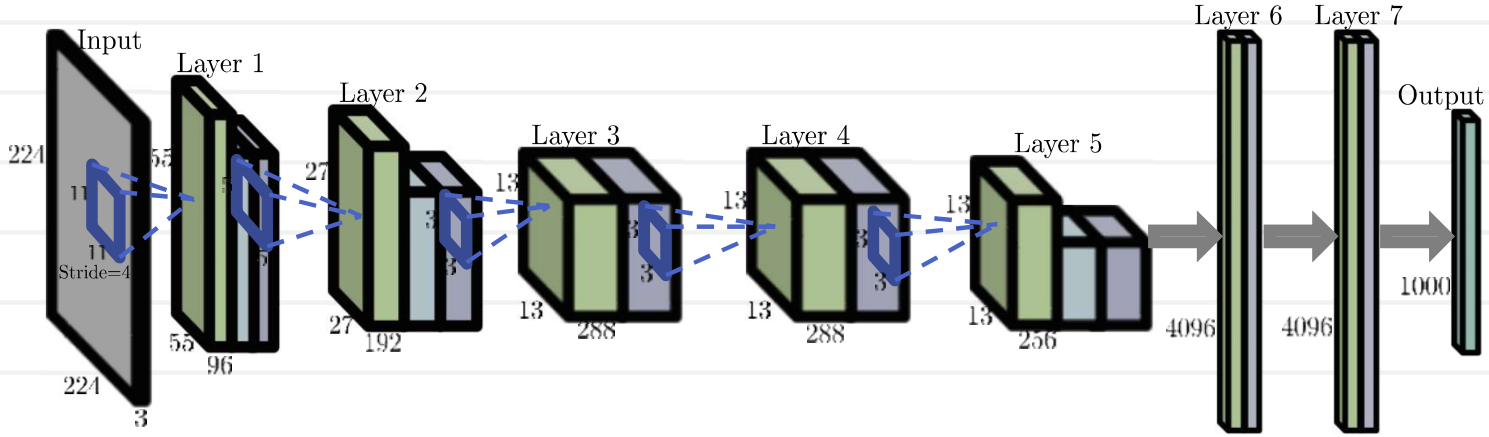


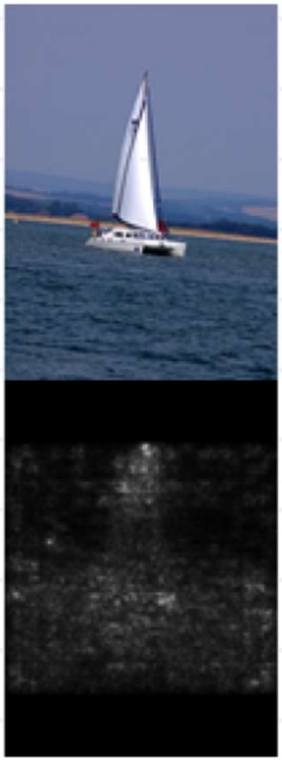

1. Occluding different regions
2. Visualizing the score
3. Good visualizations, but slow
4. Arbitrary choice of occluder

Can we do grounding faster?

# Grounding using gradient

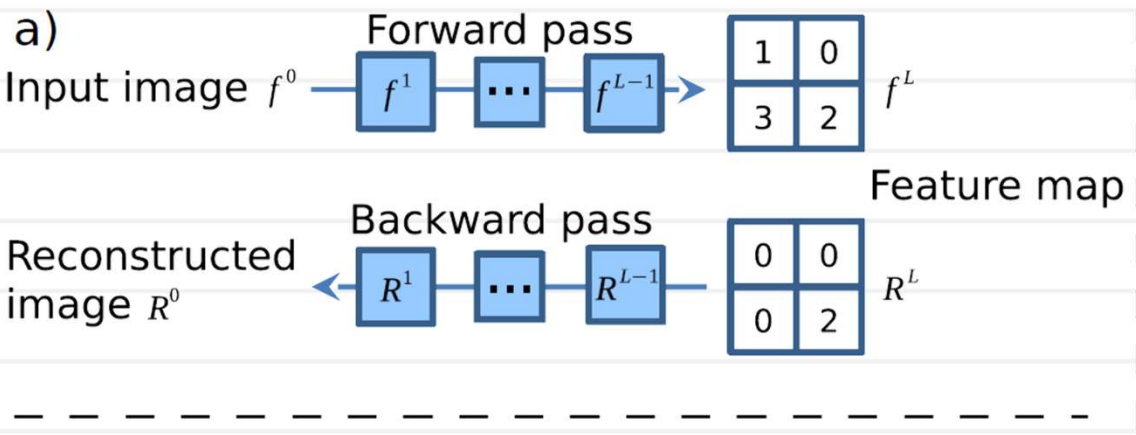
$$\left\| \frac{\partial \varphi_{out}}{\partial x_0} \right\|$$



[Simonyan et al. 2013]

# Better grounding using gradient



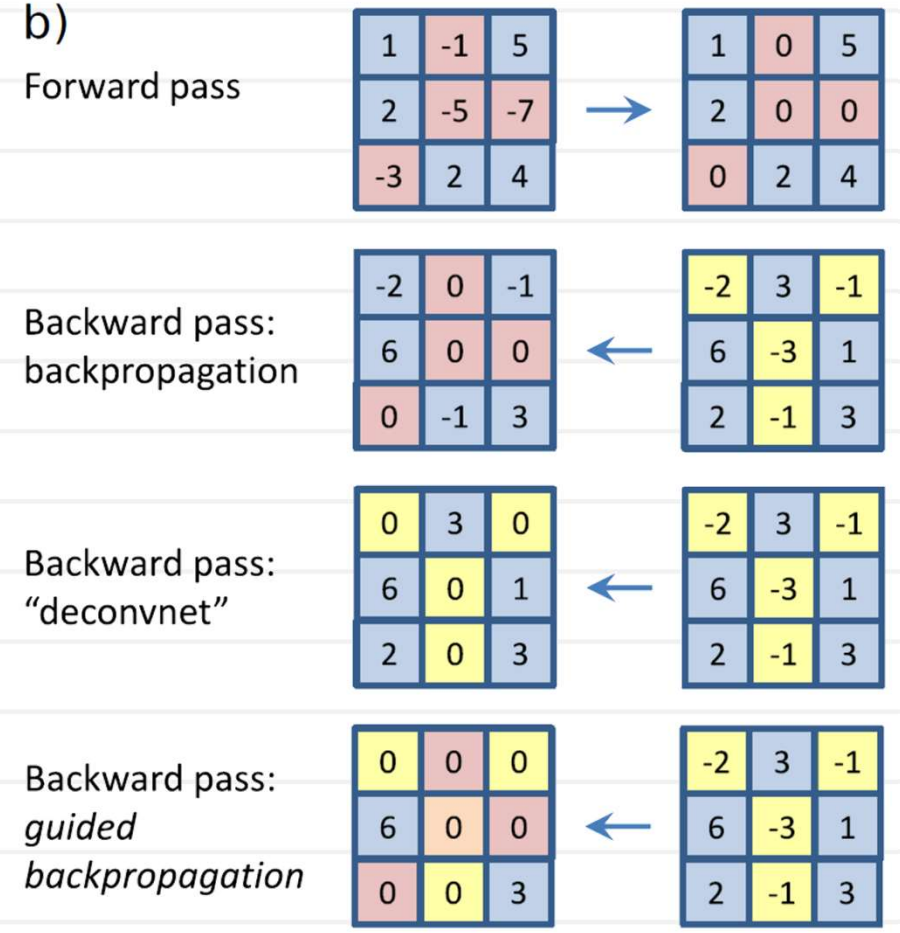

c)

activation: $f_i^{l+1} = relu(f_i^l) = \max(f_i^l, 0)$

backpropagation: $R_i^l = (f_i^l > 0) \cdot R_i^{l+1}$, where $R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$

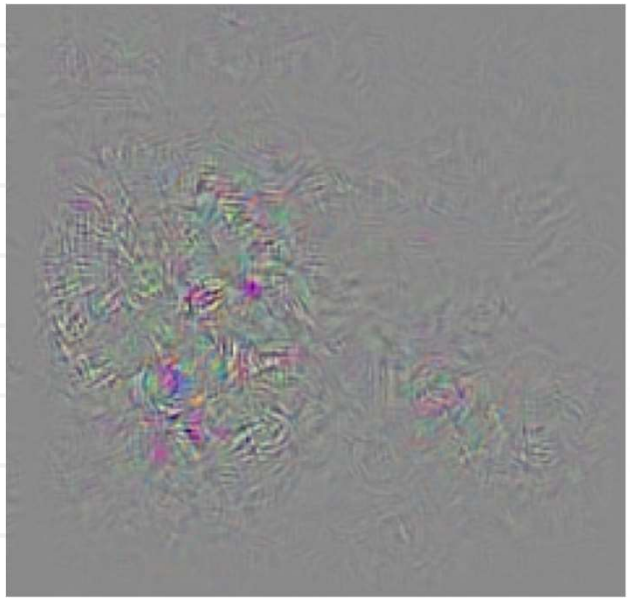
backward 'deconvnet': $R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1}$

guided backpropagation: $R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}$

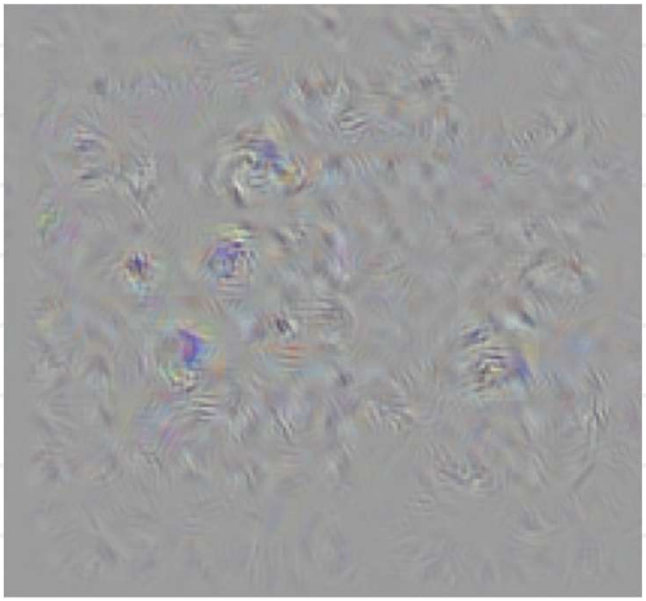



[Springenberg et al. 2015]

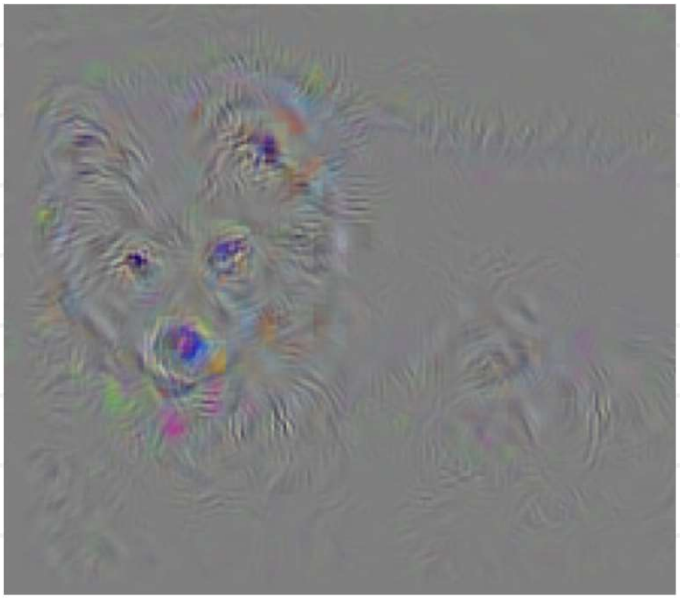
# Better grounding using gradient



gradient



DeConvNet

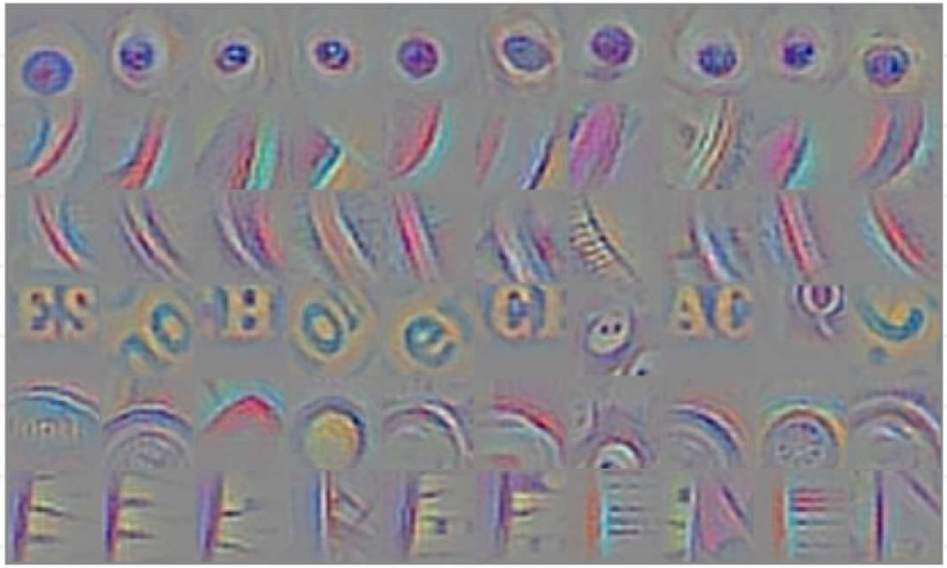


Guided backprop

[Springenberg et al. 2015]

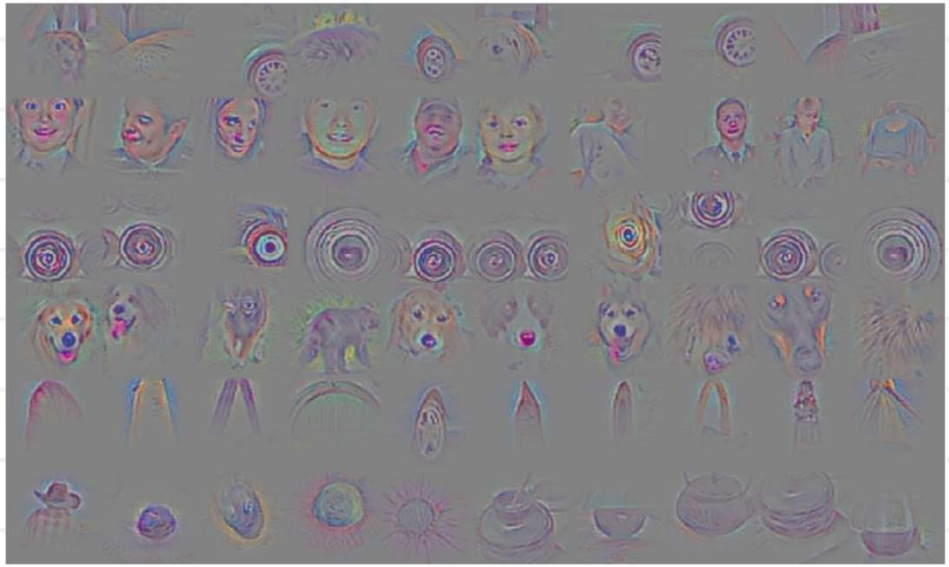
# Better grounding using gradient

guided backpropagation

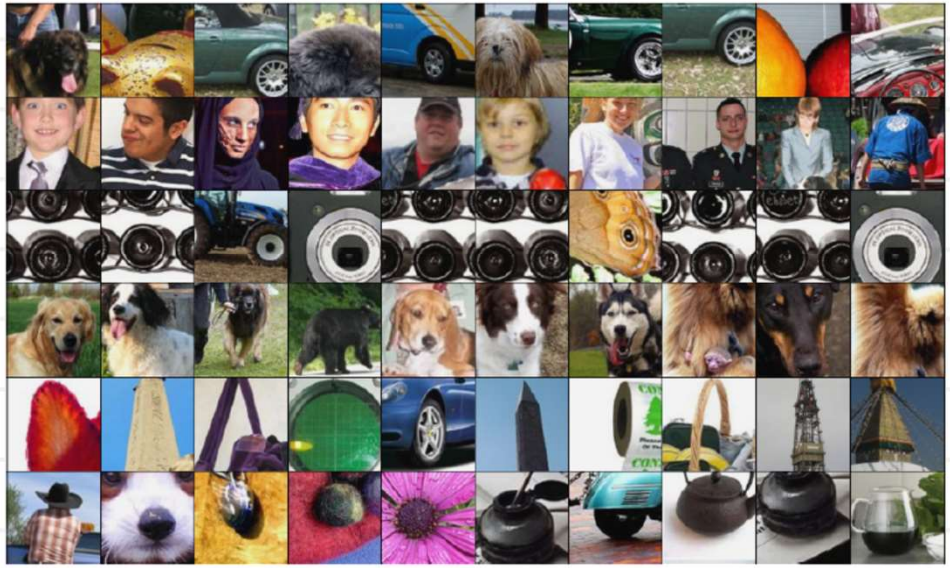


corresponding image crops

guided backpropagation



corresponding image crops



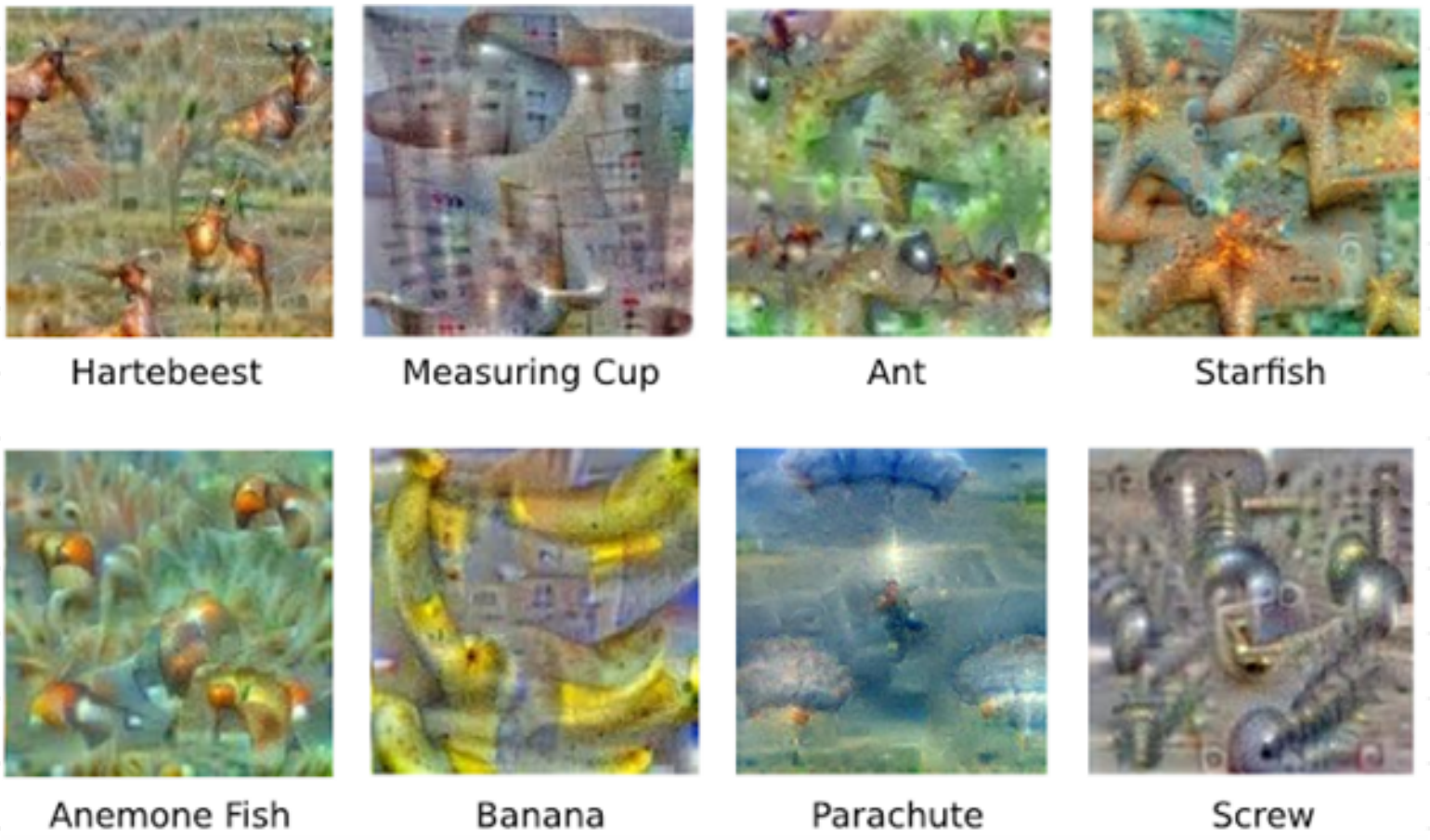
[Springenberg et al. 2015]

# Maximum impulses

$$\hat{x} = \arg\max_{x} \left( \Phi_{out}(x)[i] \right) - \lambda R(x)$$

[Simonyan et al. 2013]

# Inceptionism: maximum impulses



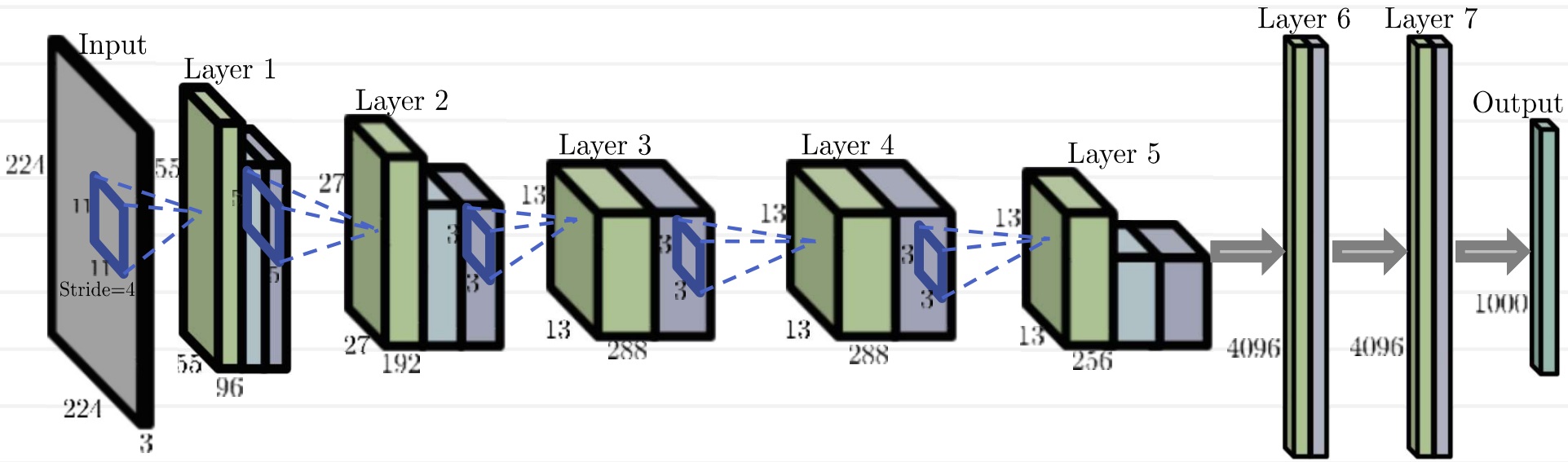

Deeper network + "Smart" regularizer (jitter)

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

# Inceptionism: maximum impulses

Deeper network + "Smart" regularizer (jitter)

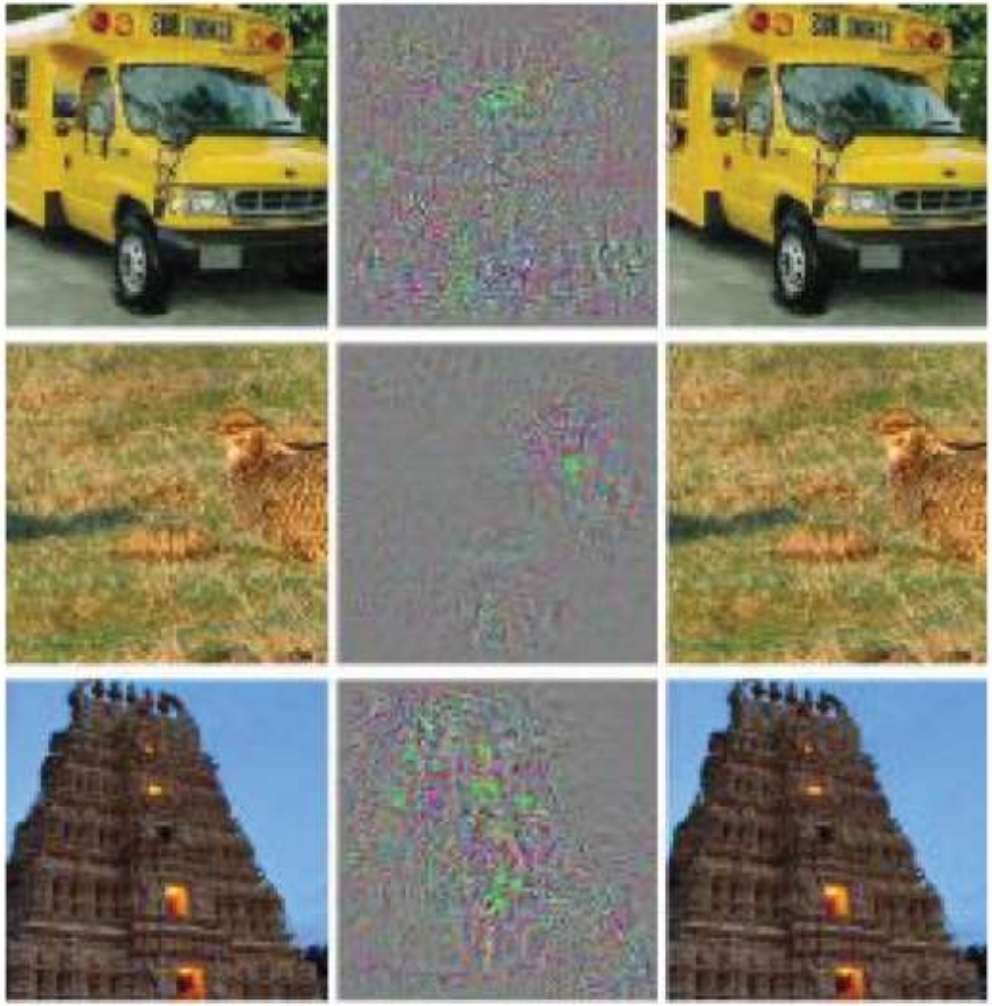
# Generating adversarial perturbations

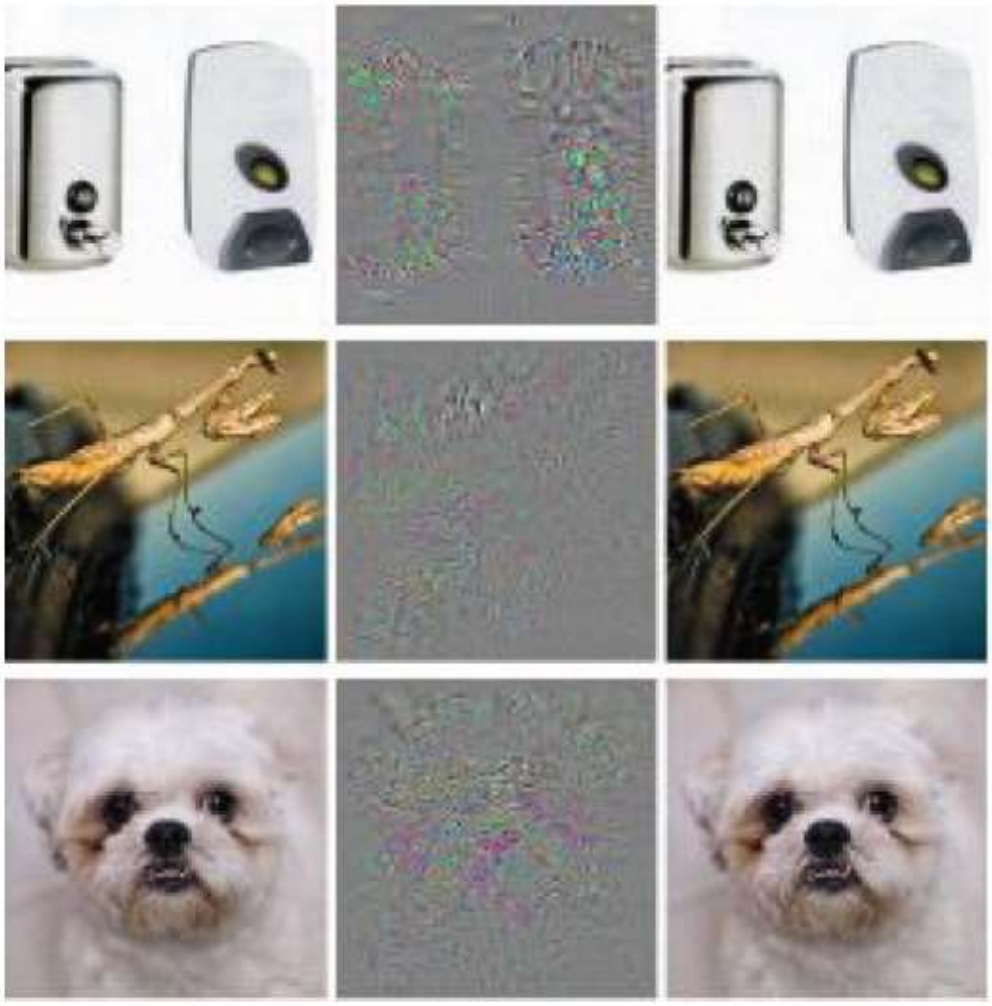



$$\min \; c|r| + L(x+r; k)$$

[Szegedy et al 2014]

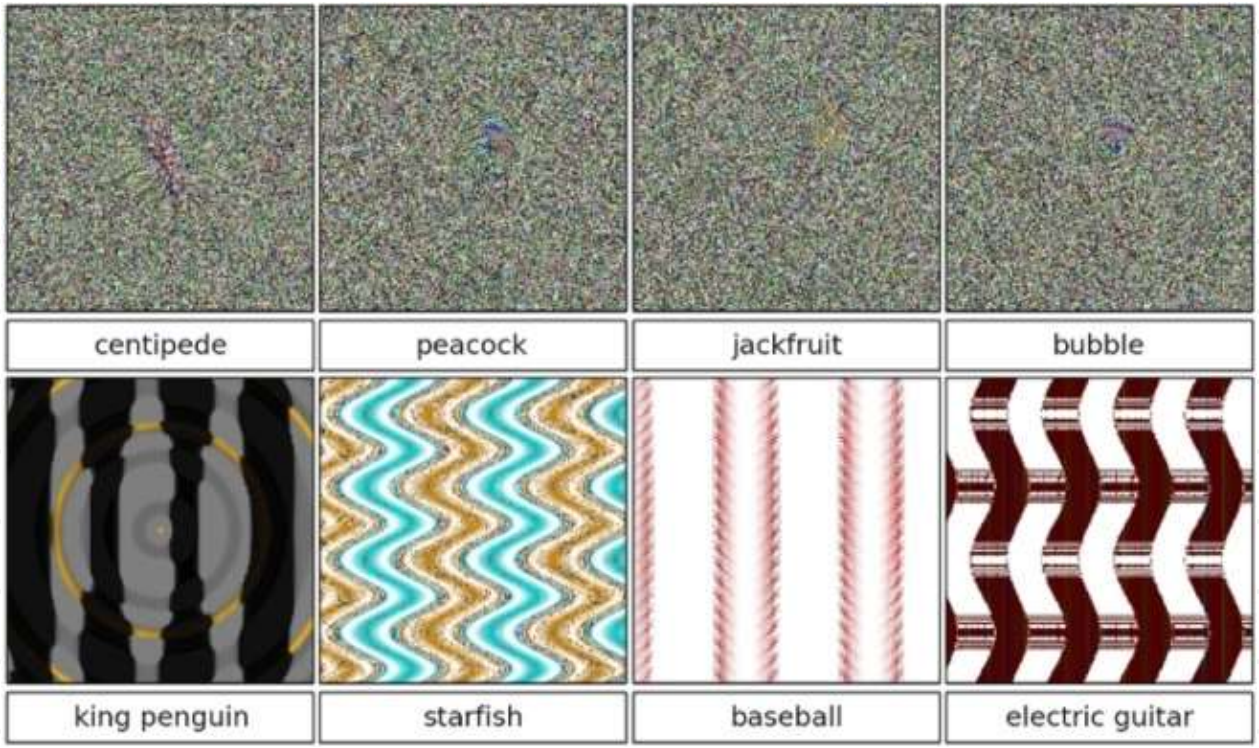
# Generating adversarial perturbations





[Szegedy et al 2014]

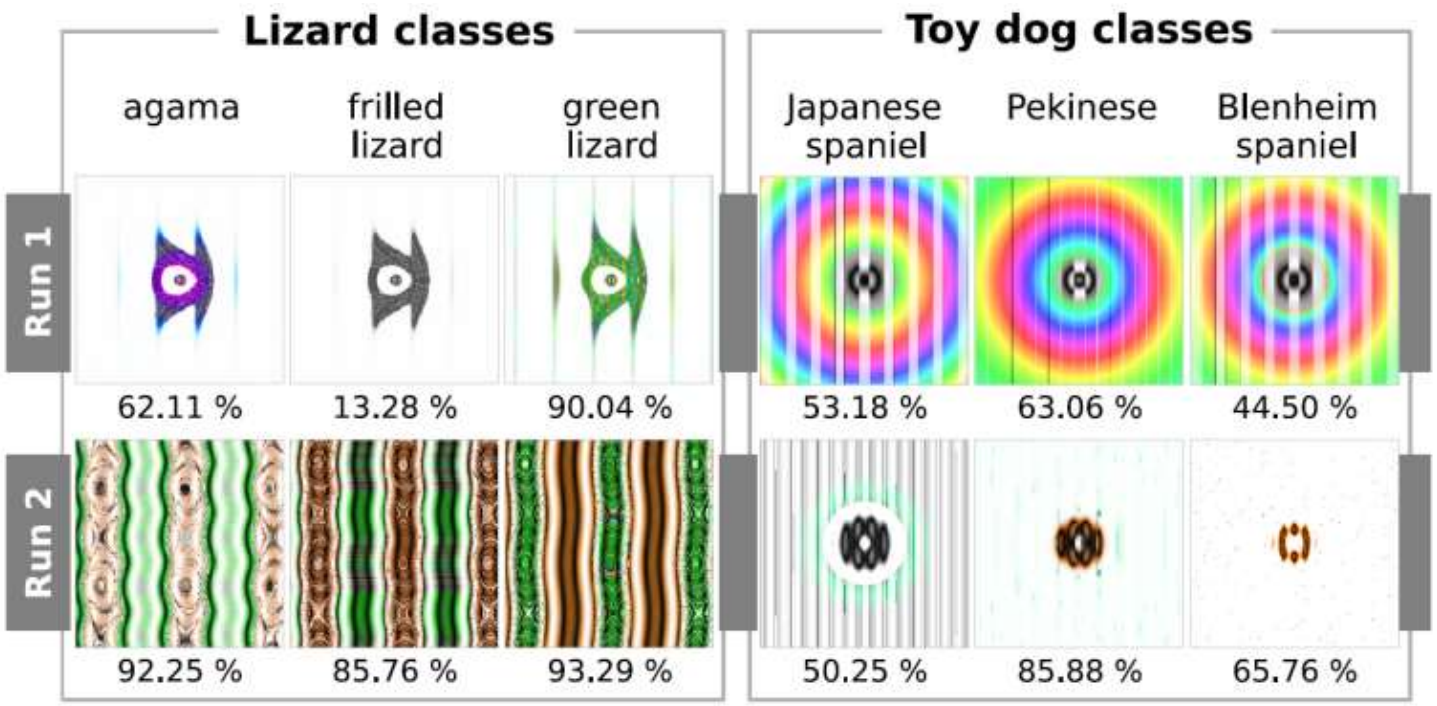
# High-confidense patterns

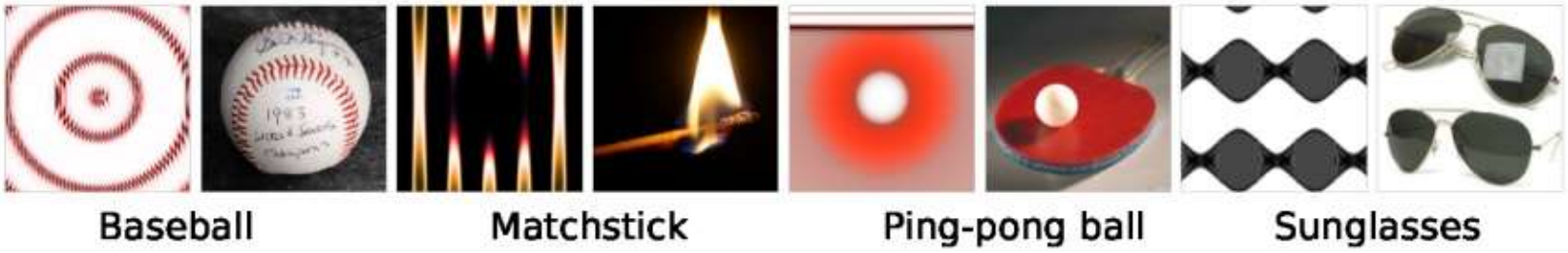



[Nguyen et al. CVPR15]:

- Evolutionary optimization
- Compositional pattern-producing networks

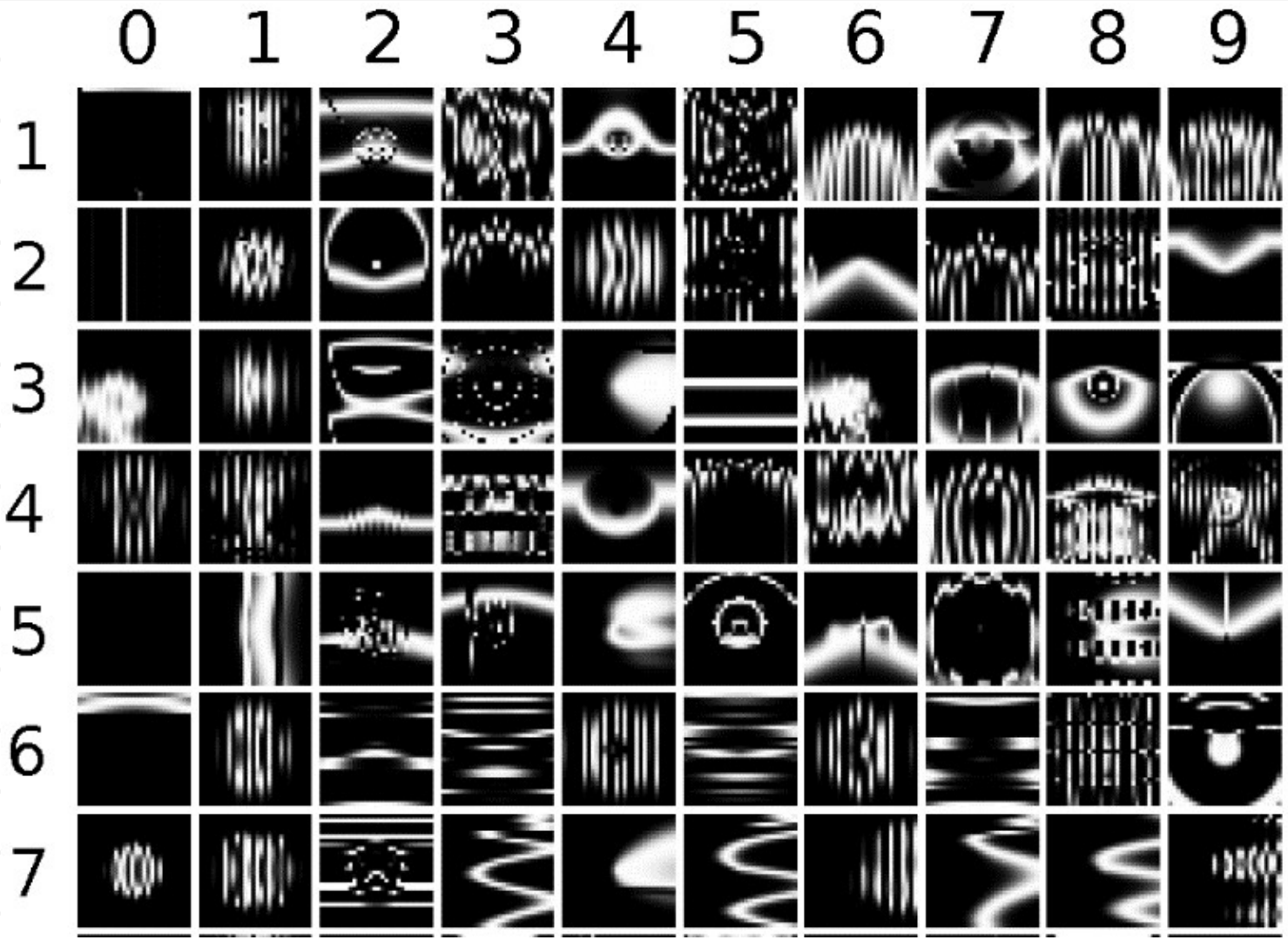
# High-confidense patterns




[Nguyen et al. CVPR15]

# Adversarial examples do not help




[Nguyen et al. CVPR15]

# ConvNets can learn anything!

[Zhang et al. ICLR17]: **typical ConvNet architectures can fit random labels on ImageNet!**

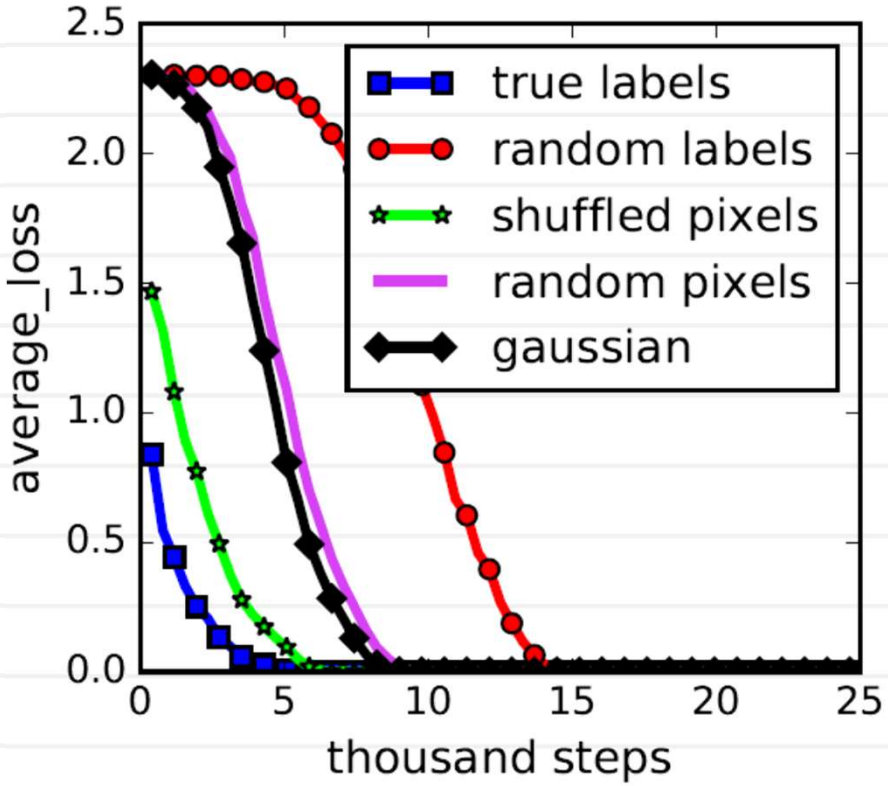
(and regularization tricks do not help much!)

| data aug | dropout | weight decay | top-1 train | top-5 train | top-1 test | top-5 test |
|---|---|---|---|---|---|---|
| ImageNet 1000 classes with the original labels | | | | | | |
| yes | yes | yes | 92.18 | 99.21 | 77.84 | 93.92 |
| yes | no | no | 92.33 | 99.17 | 72.95 | 90.43 |
| no | no | yes | 90.60 | 100.0 | 67.18 (72.57) | 86.44 (91.31) |
| no | no | no | 99.53 | 100.0 | 59.80 (63.16) | 80.38 (84.49) |
| Alexnet (Krizhevsky et al., 2012) | | | - | - | - | 83.6 |
| ImageNet 1000 classes with random labels | | | | | | |
| no | yes | yes | 91.18 | 97.95 | 0.09 | 0.49 |
| no | no | yes | 87.81 | 96.15 | 0.12 | 0.50 |
| no | no | no | 95.20 | 99.14 | 0.11 | 0.56 |

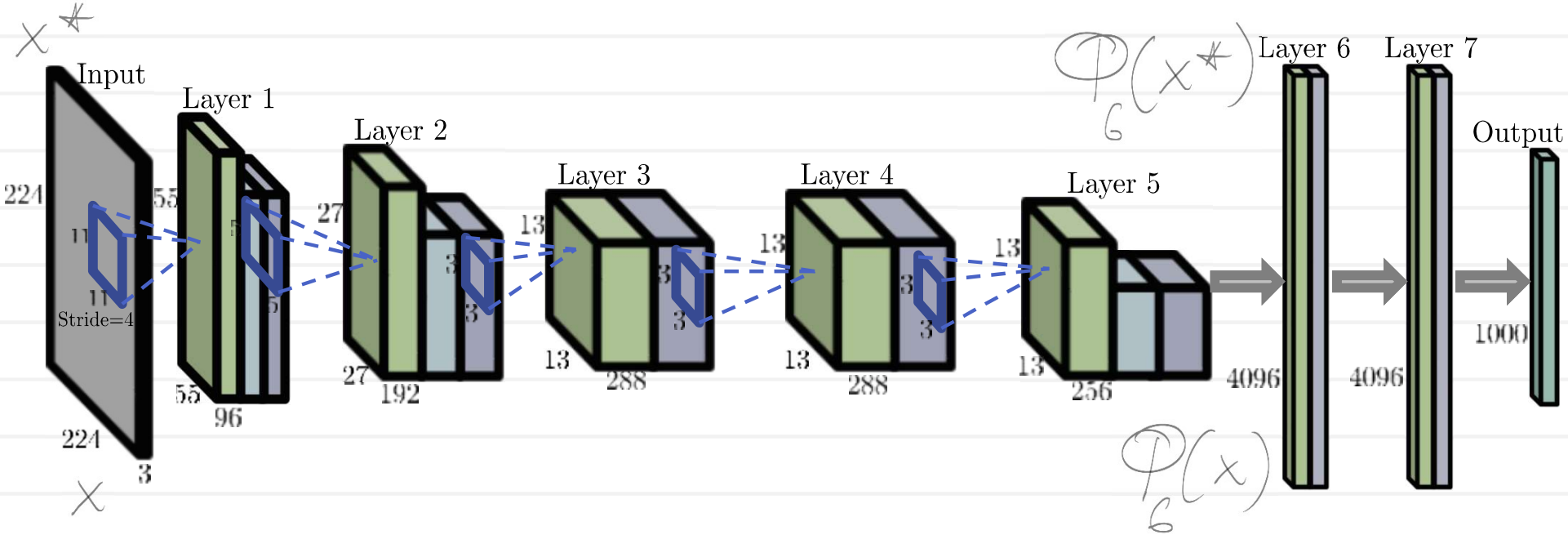
# ConvNets can learn anything!

[Zhang et al. ICLR17]: **typical ConvNet architectures can fit random labels on CIFAR perfectly!**

(and regularization tricks do not help much!)



# Generating images by Inverting CNNs




$$x = \arg\min \|\Phi_6(x^*) - \Phi_6(x)\|^2 + \lambda R(x)$$

Standard regularizer:

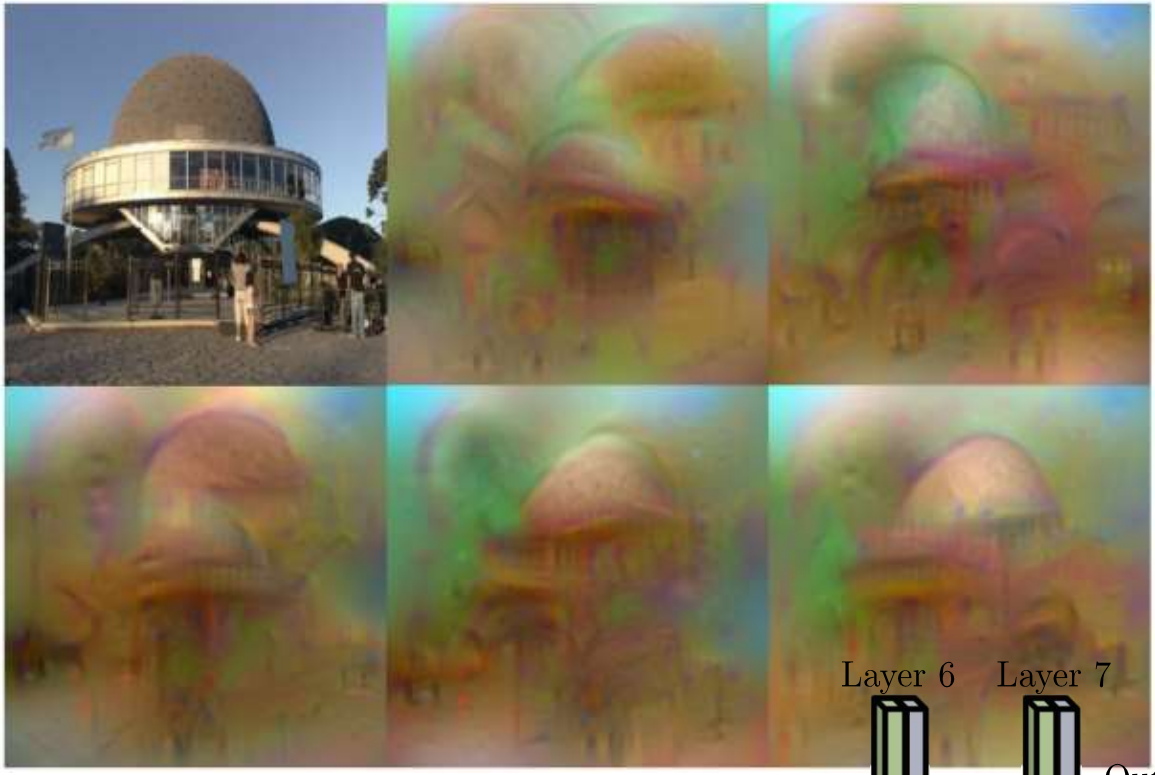
$$\sum \left( (x_{i,j+1} - x_{ij})^2 + (x_{i+1,j} - x_{ij})^2 \right)^{\beta/2}$$

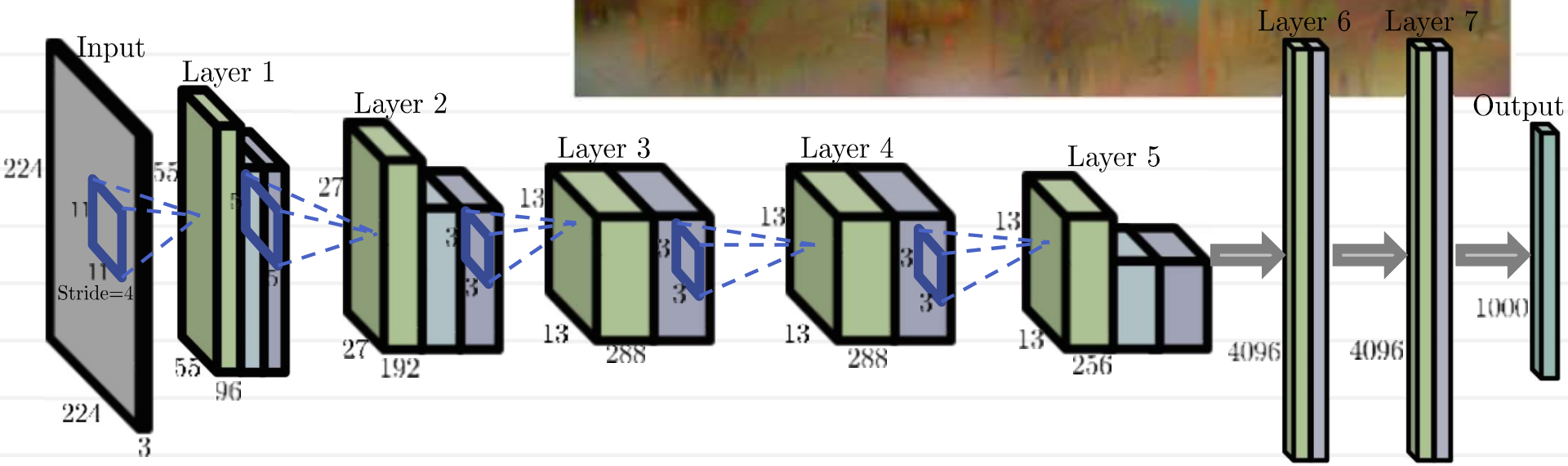
[Mahendrahan & Vedaldi CVPR15]

# Generating images by Inverting CNNs

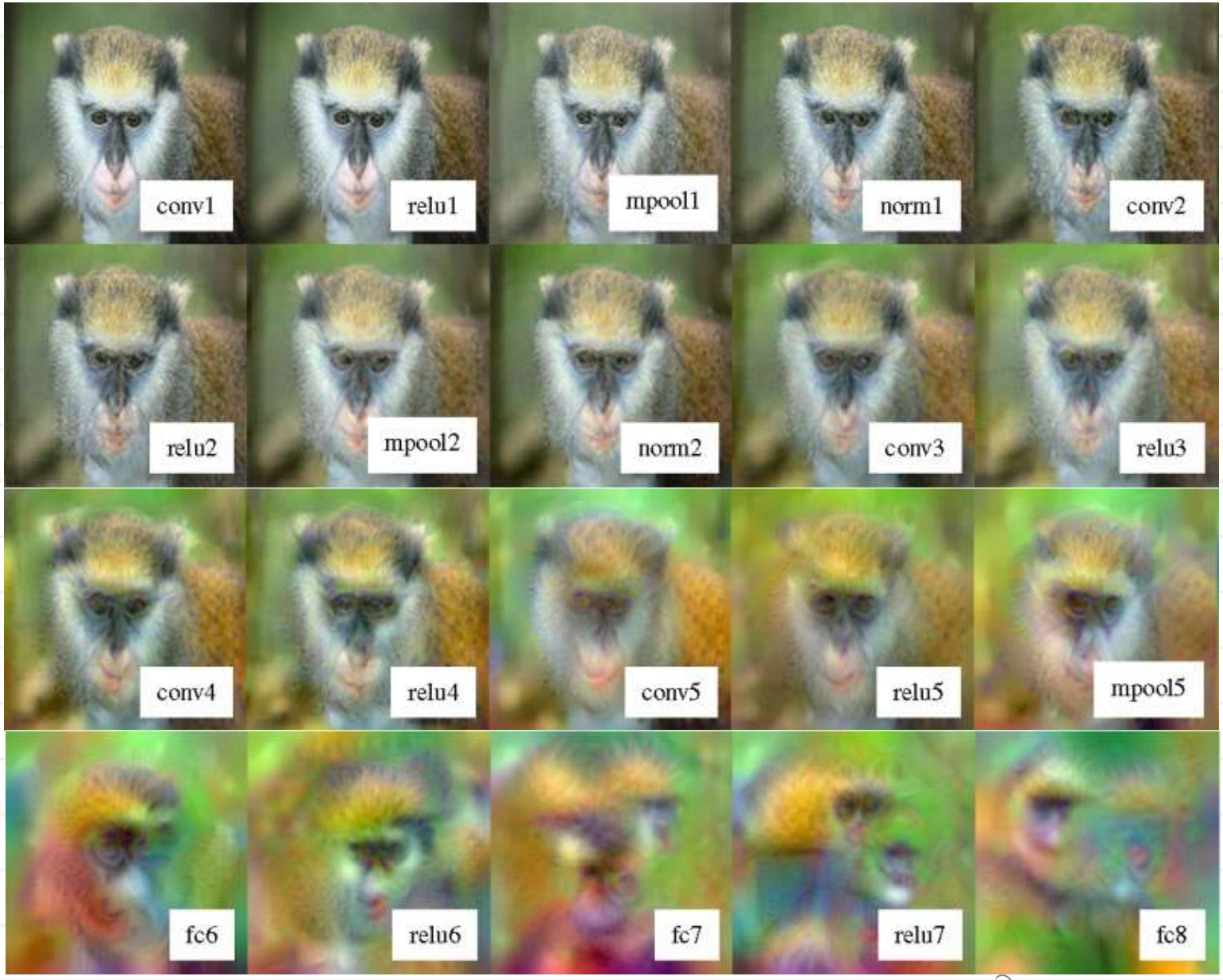
$$x^* = \arg\min \|\Phi(x^*) - \Phi(x)\|^2 + \lambda R(x)$$

[Mahendrahan & Vedaldi CVPR15]





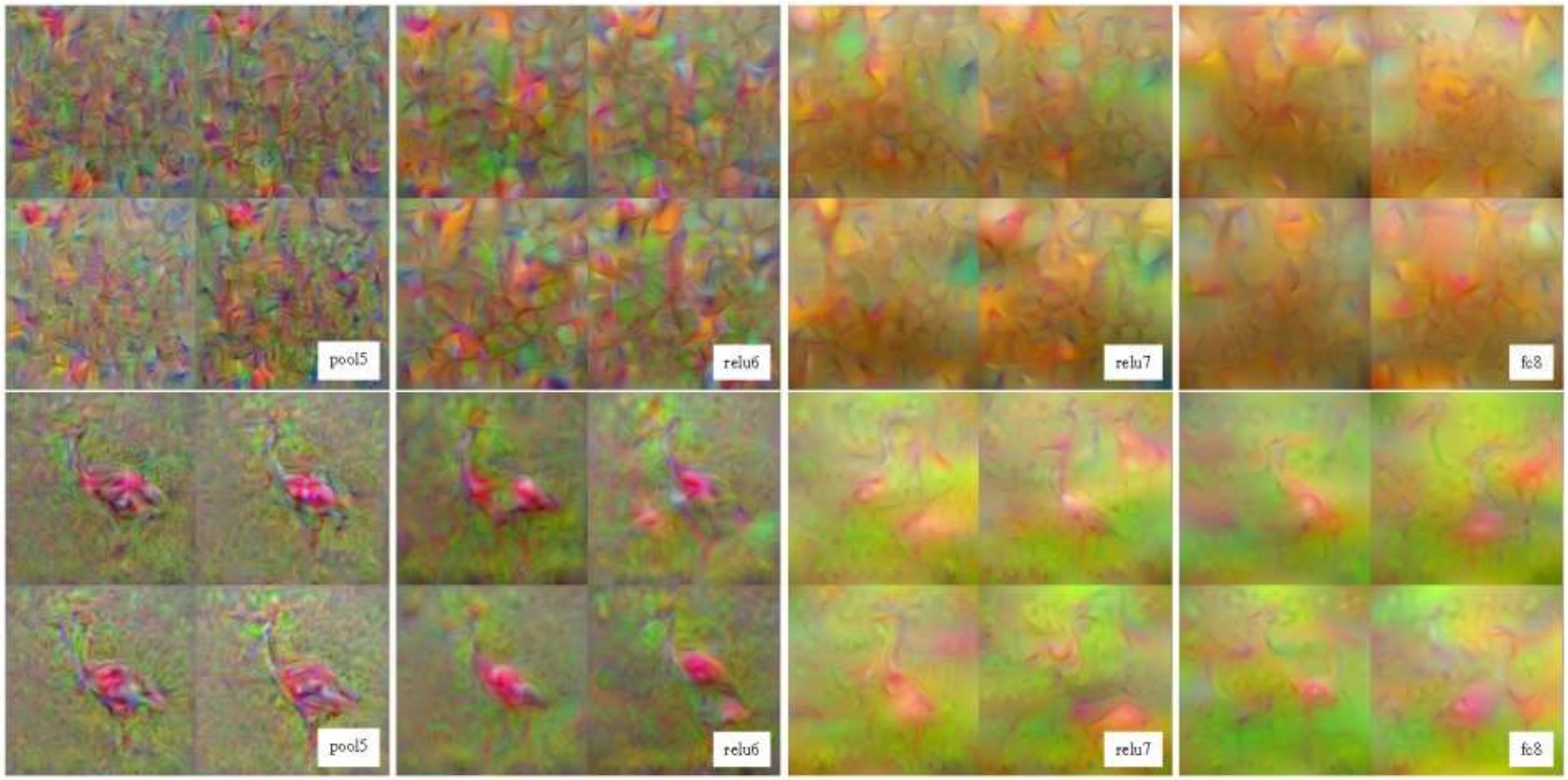
# Generating images by Inverting CNNs




[Mahendrahan & Vedaldi CVPR15]

$$x^* = \arg\min \|\Phi(x^*) - \Phi(x)\|^2 + \lambda R(x)$$

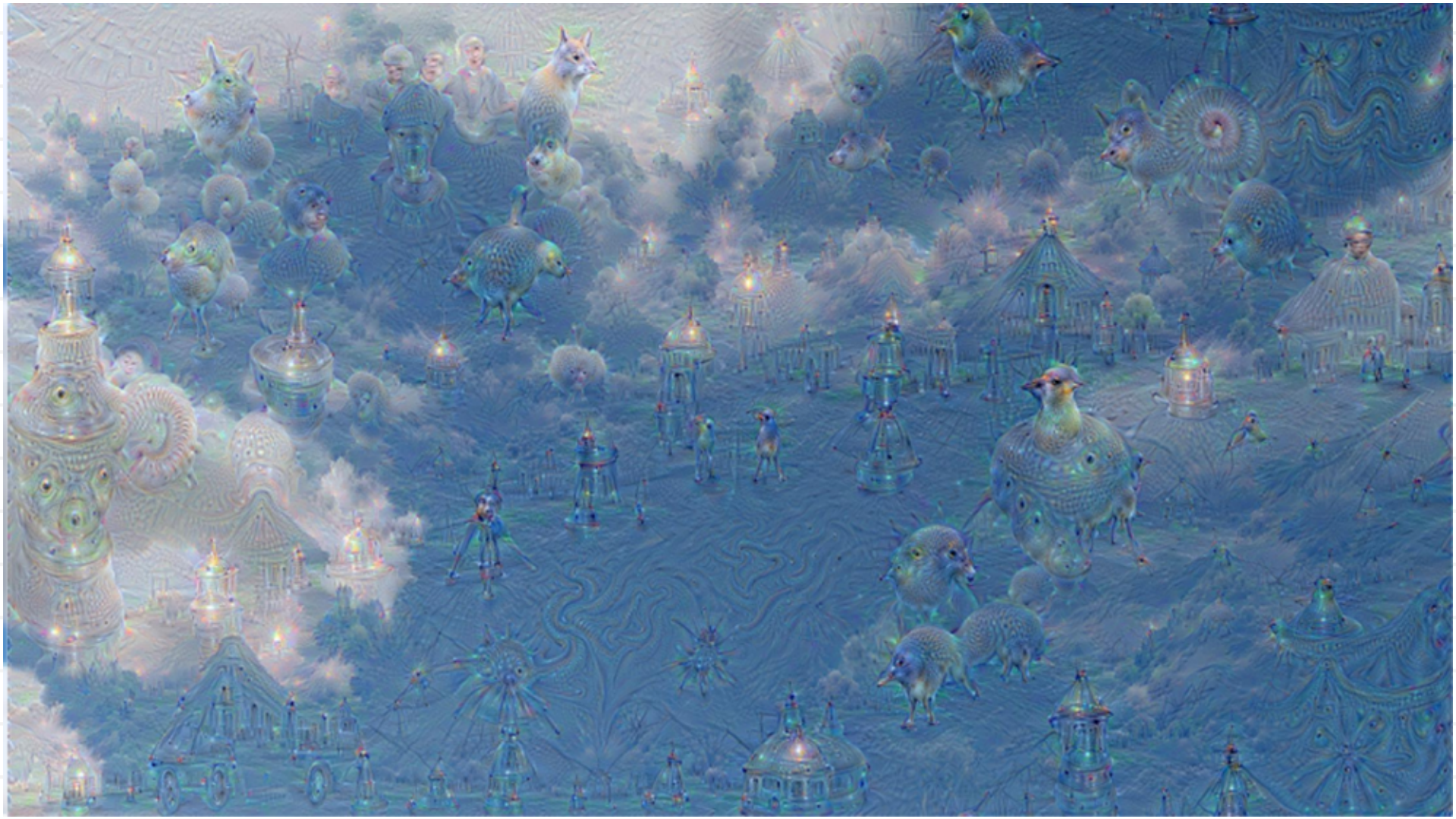
# Multiple restarts




[Mahendrahan & Vedaldi CVPR15]

# Inceptionism: DeepDream

ConvNet on drugs:

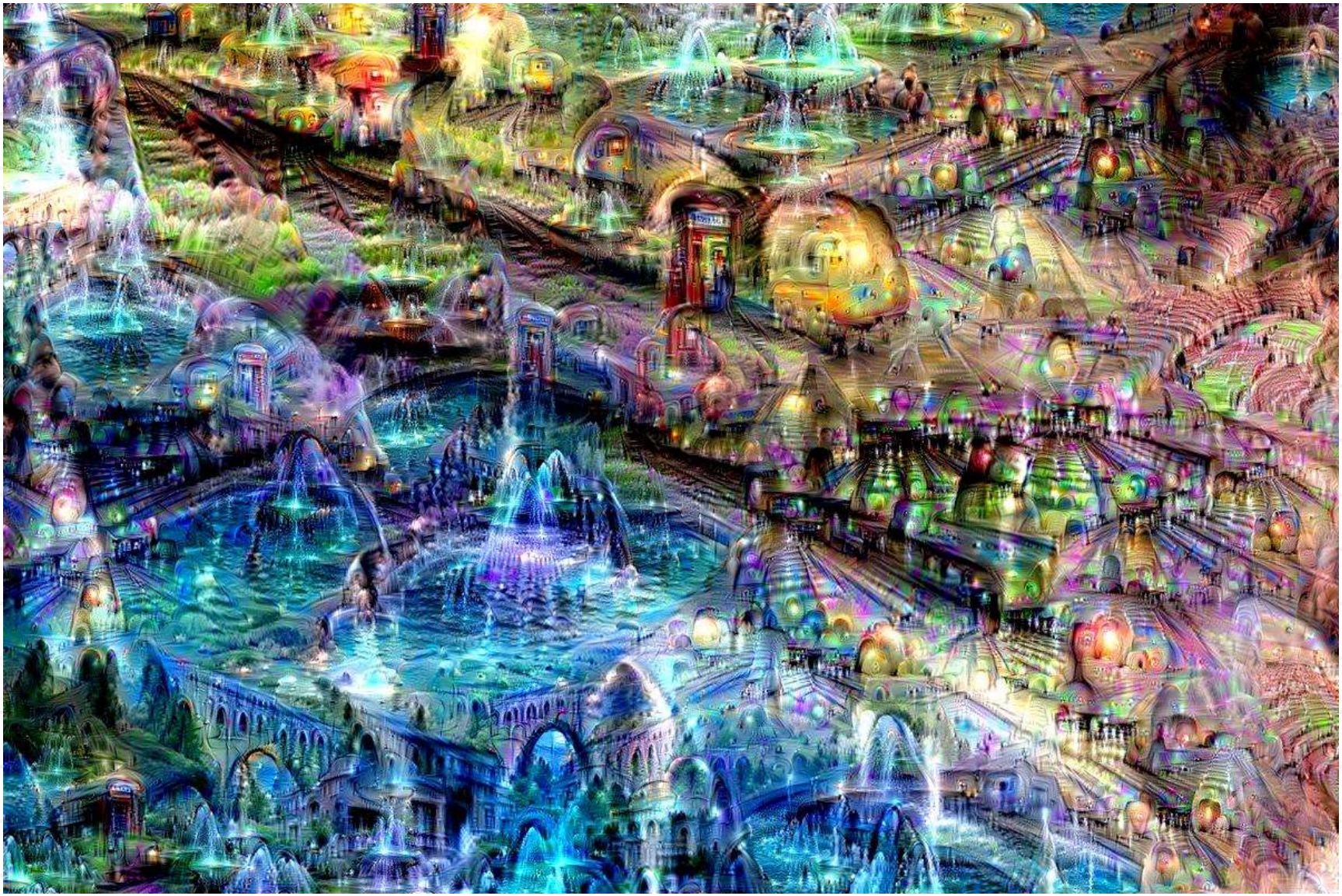


# Inceptionism: DeepDream

ConvNet on drugs:

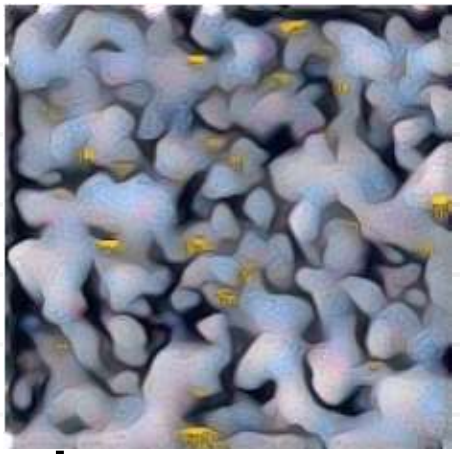
# Inceptionism: DeepDream

ConvNet on drugs:



# Texture synthesis

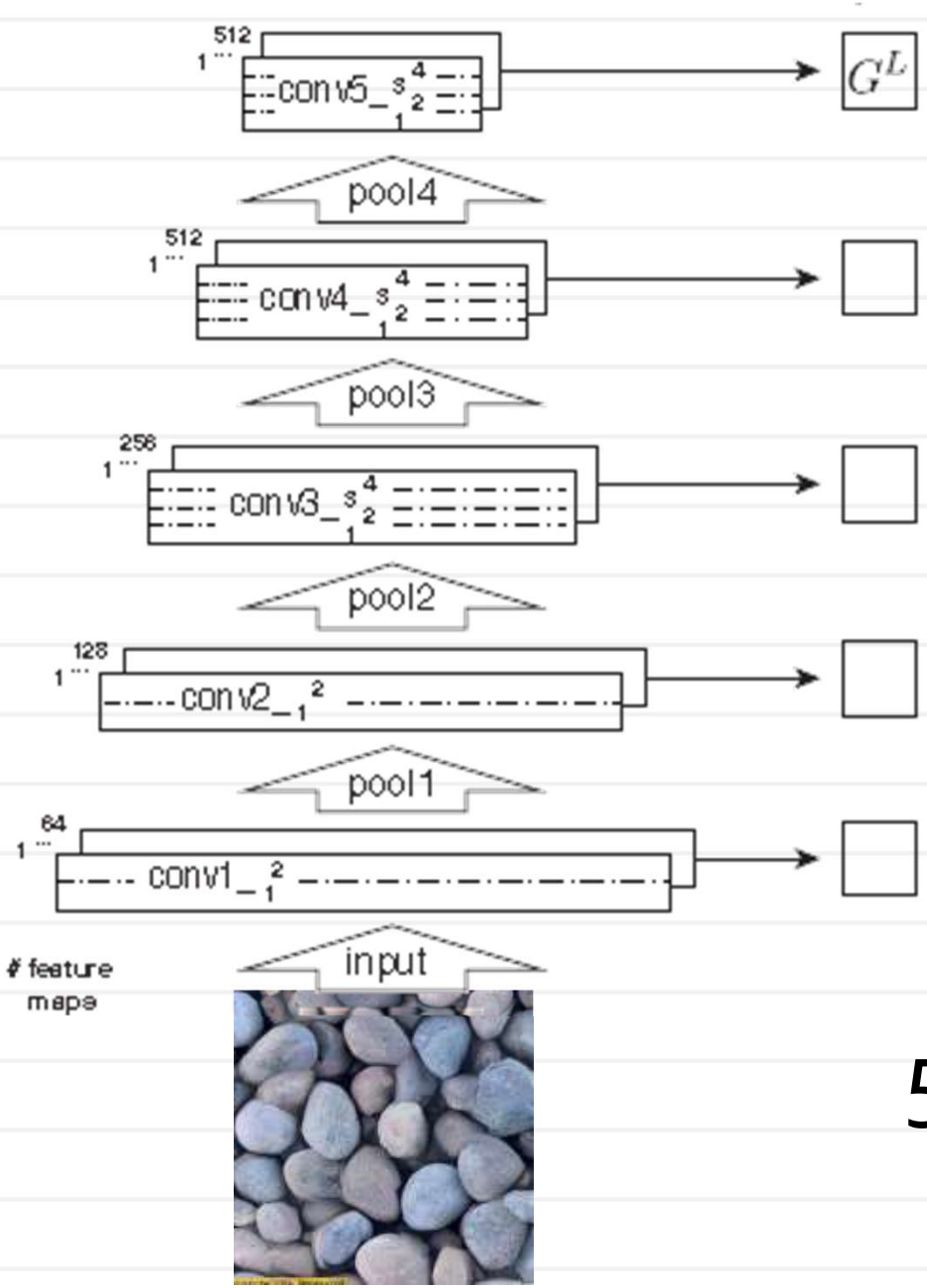
How can we measure the similarity of *textures (i.e. spatially-stationary natural visual patterns)*?

similar different

- Pixel-wise measures are useless
- Histograms are not too useful
- Long history of research

# What describes a texture?




Gram matrix:

$$\left[ G^l \right]$$

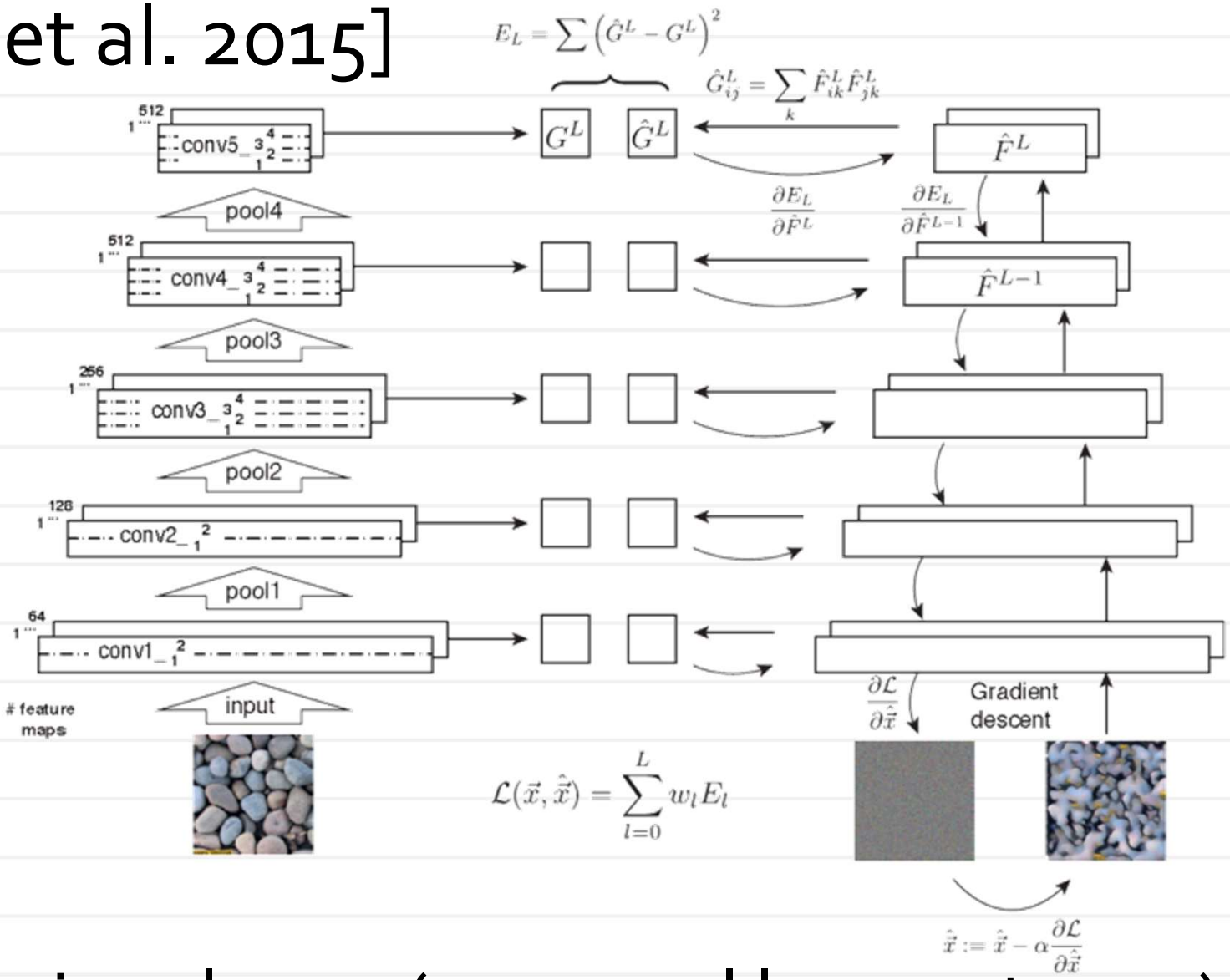
$$G^l_{ij} = \sum_k F^l_{i,k} \cdot F^l_{j,k}$$

512x512-dim descriptor

[Gatys et al. 2015]

# Synthesis using pre-image method

[Gatys et al. 2015]




- Various layers (or several layers at once) can be used for different textures

# Results for Texture synthesis

Synthesised

Source

Synthesised

Source

[Gatys, Ecker, Bethge 2015]

# Results for Texture synthesis

Synthesised

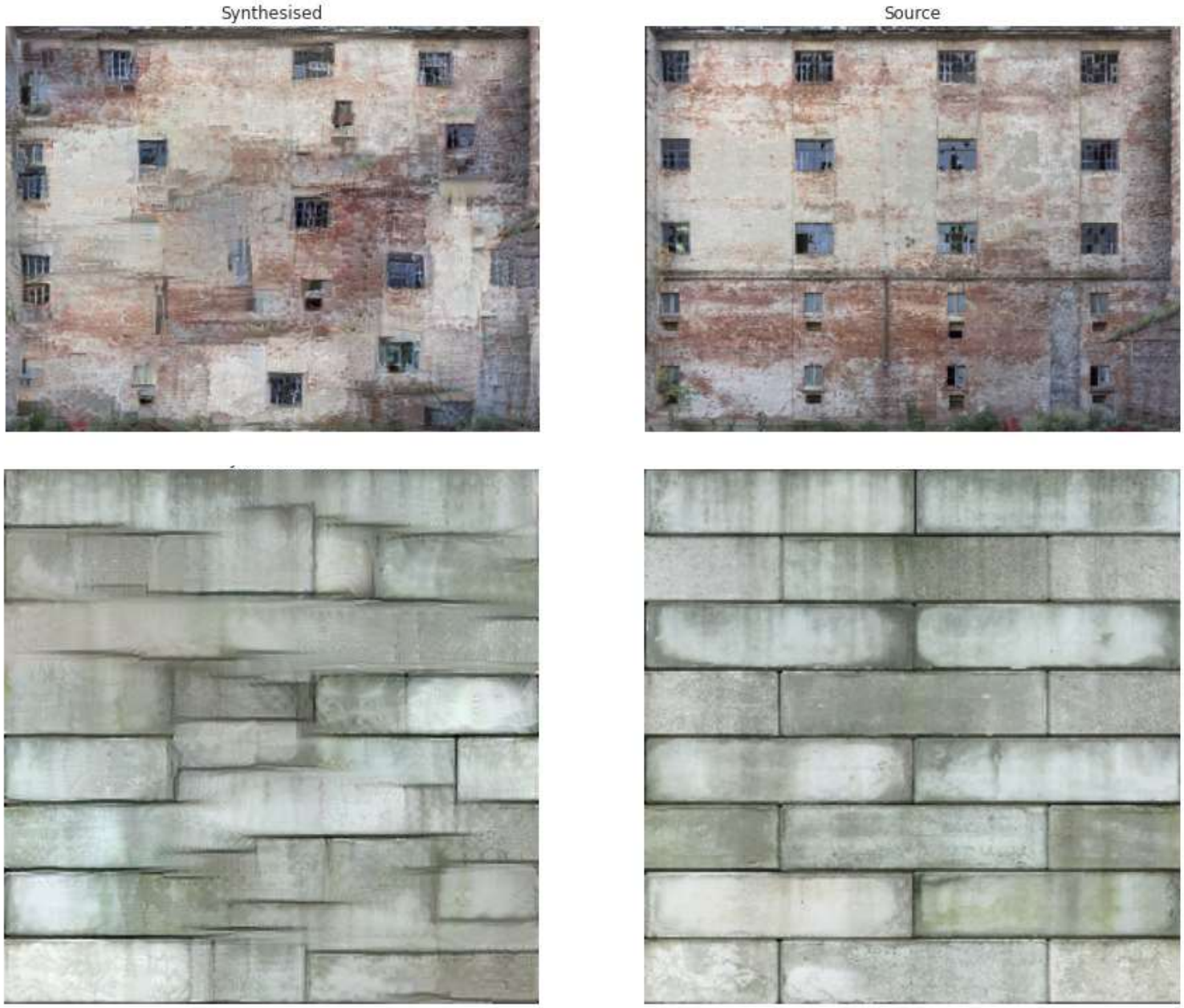
Source

Synthesised

Source

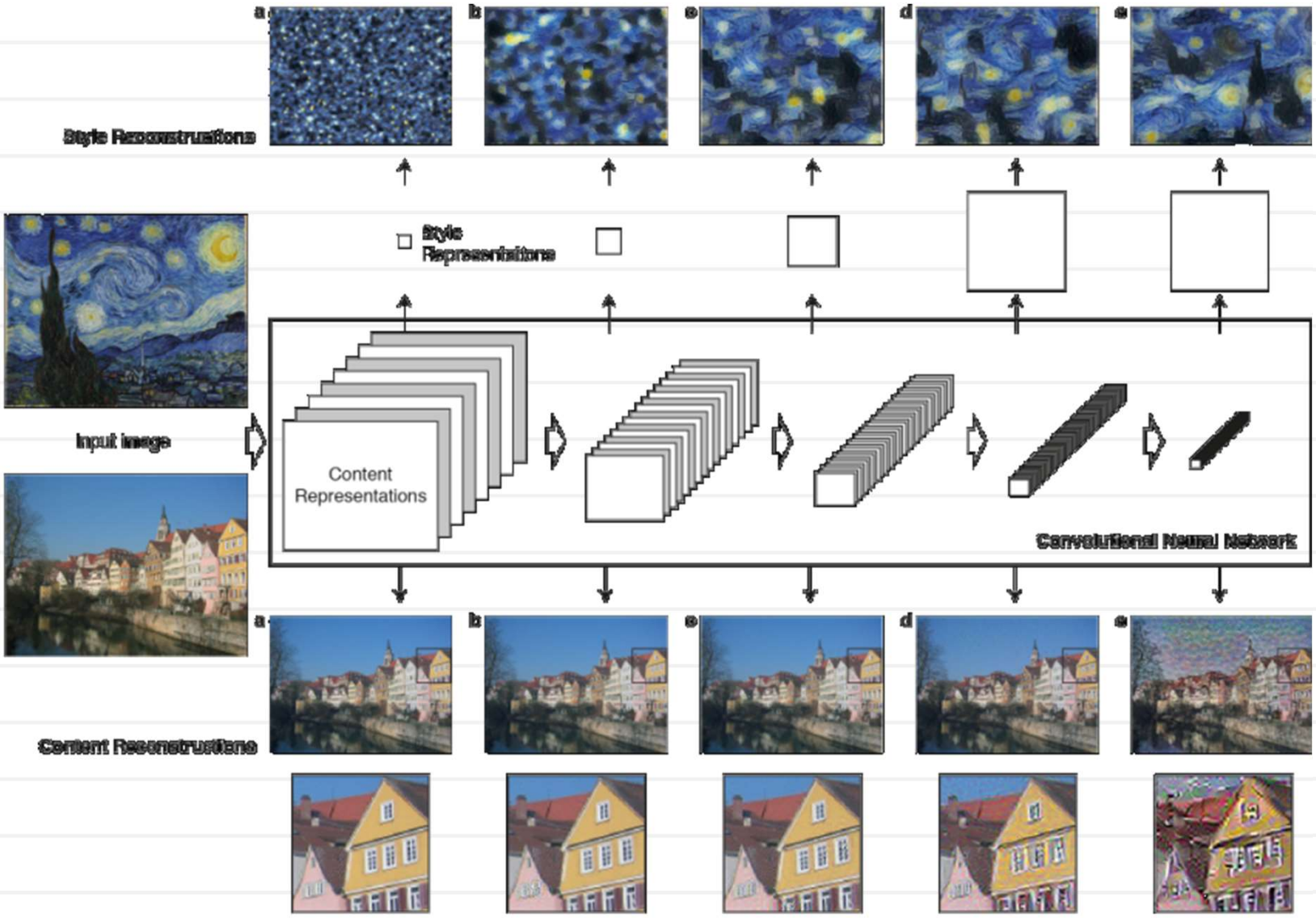
[Gatys, Ecker, Bethge 2015]

# Results for Texture synthesis



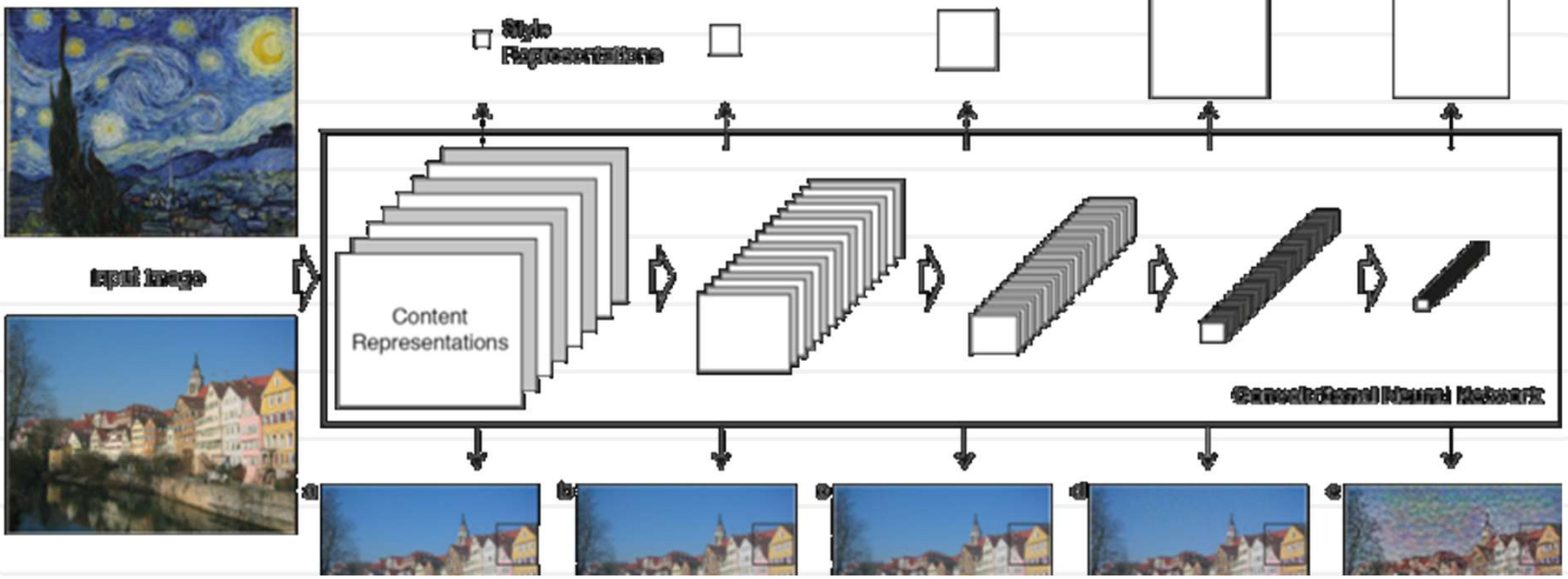

[Gatys, Ecker, Bethge 2015]

# Stylization with texture loss




[Gatys et al. 2015]

# Stylization with texture loss



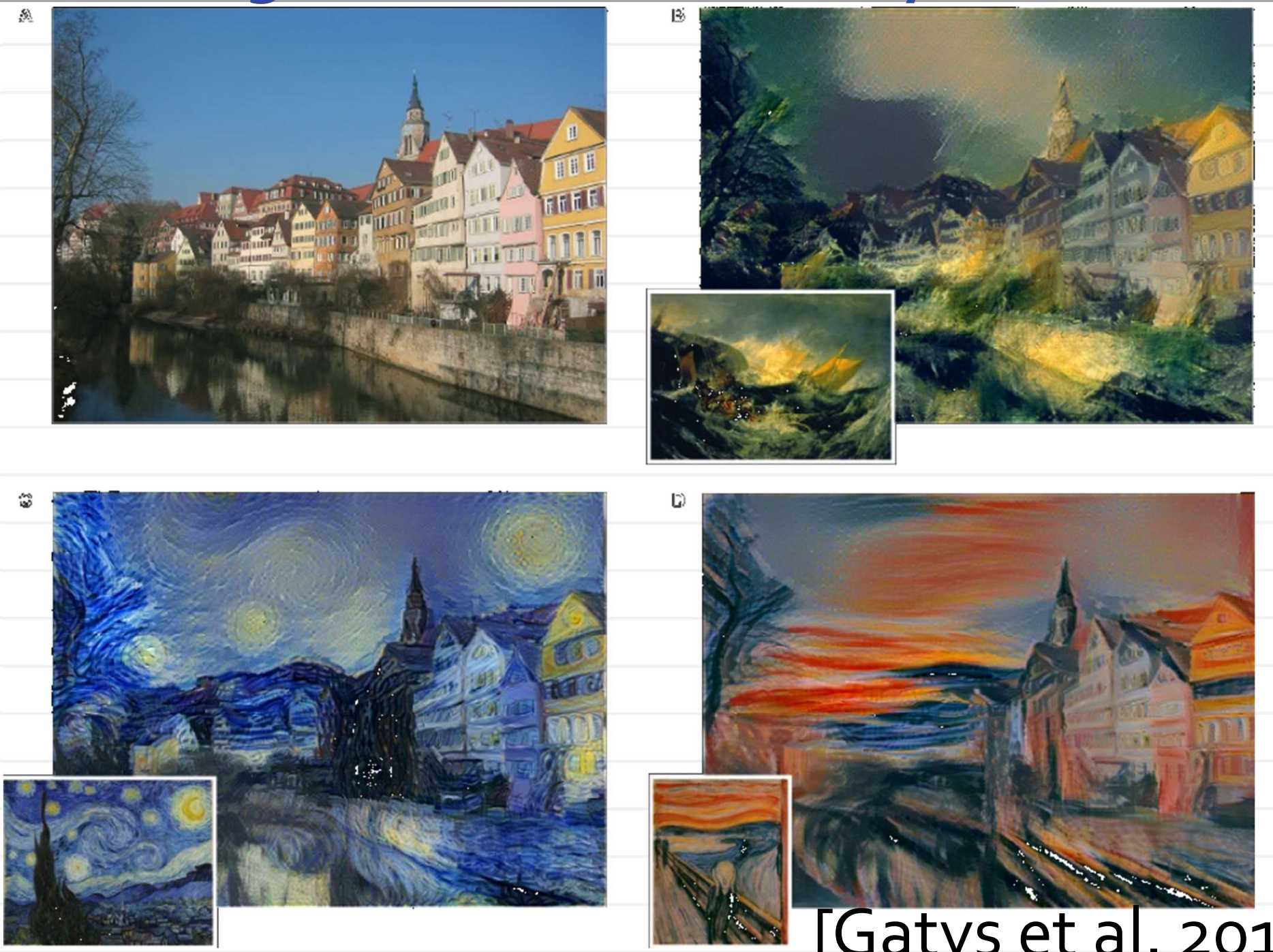

$$\|G^{\ell} - G_A^{\ell}\|^2$$

$$\|F^{\ell'} - F_B^{\ell'}\|^2$$

$$L = \|G^{\ell} - G_A^{\ell}\|^2 + \|F^{\ell'} - F_B^{\ell'}\|^2$$

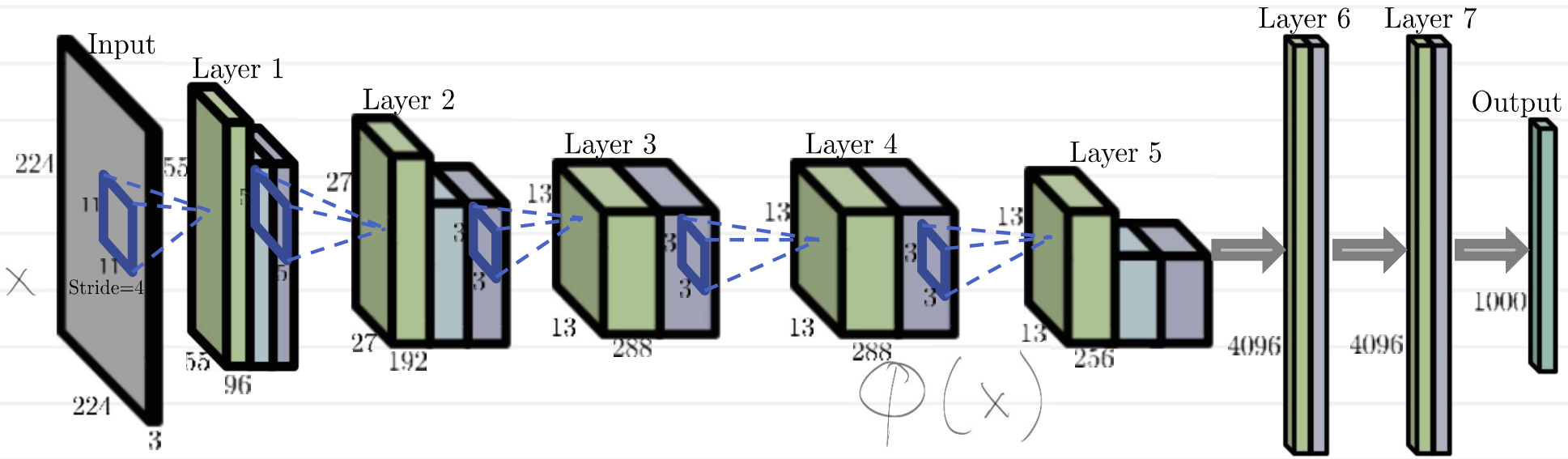
[Gatys et al. 2015]

# Neural algorithm for artistic style



[Gatys et al. 2015]

# Representations inside the neural network




Lots of important questions:

- What are their properties?
- Are they redundant?
- Are they invertible?
- Are the intermediate representations useful?

# Retrieval using learned representations

[Krizhevsky et al. NIPS12]

# Retrieving same objects/buildings

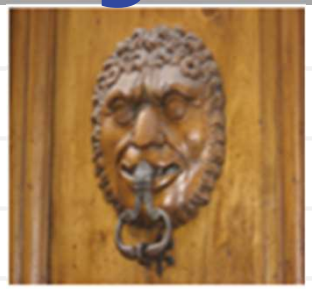
Query:

[Babenko et al. 2014]
comparing representations:



# Retrieving same objects/buildings

Query:



[Babenko et al. 2014]
comparing representations:



# Compact descriptors

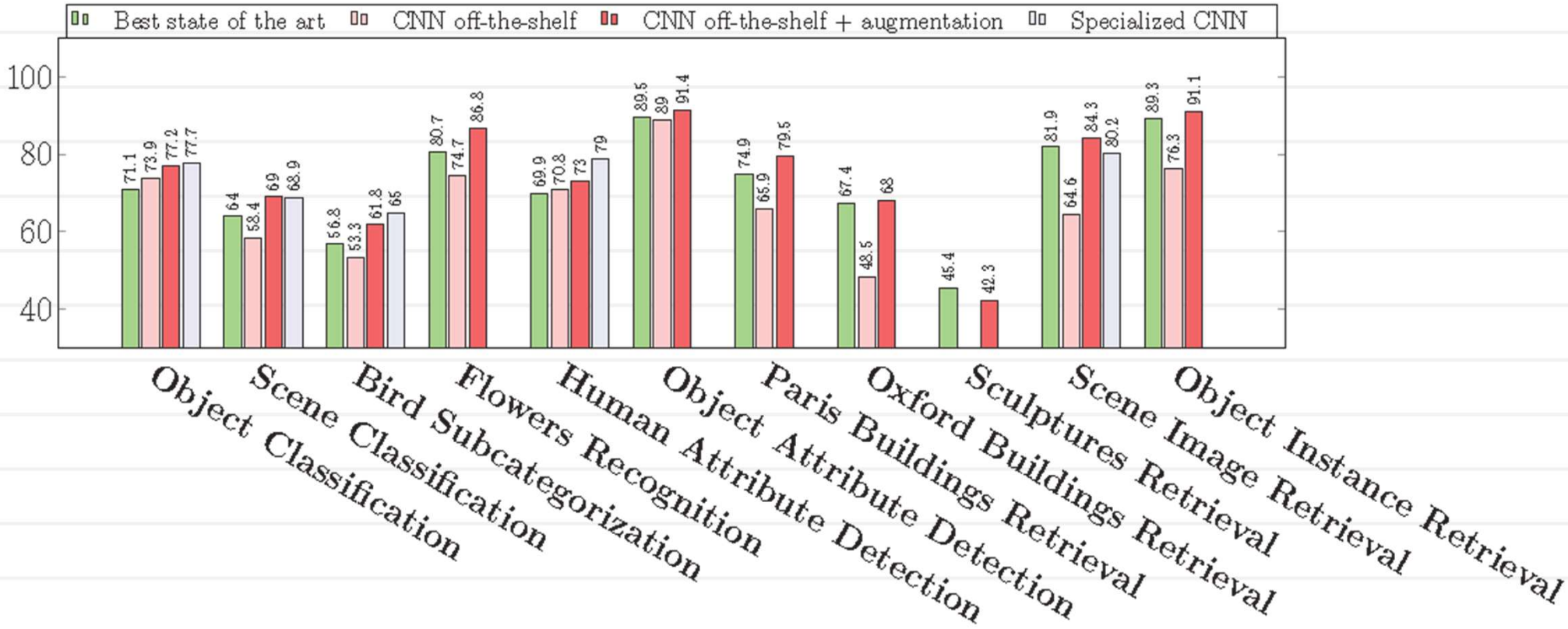
- Original descriptors can be compressed very well (e.g. to 128 dims)
- Even better representation for retrieval ([Babenko et al. ICCV2015]):

$$\psi(x) = \sum_k \sum_\ell \psi_{LC}(x)^{k,\ell}$$

$$\psi'(x) = \mathrm{diag}(s_1 \dots s_N)^{-1} M_{PCA} \psi(x)$$

$$\psi''(x) = \frac{\psi'(x)}{\|\psi'(x)\|_{L_2}}$$

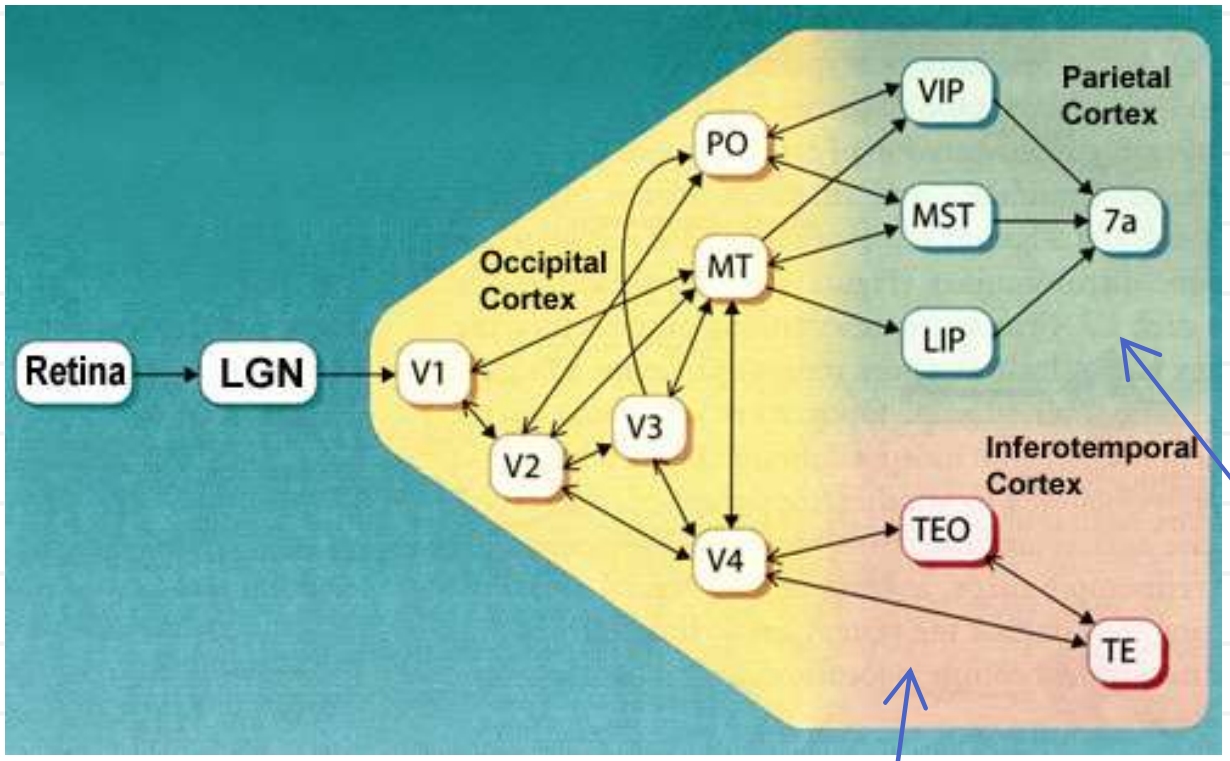
# Deep features as generic representation

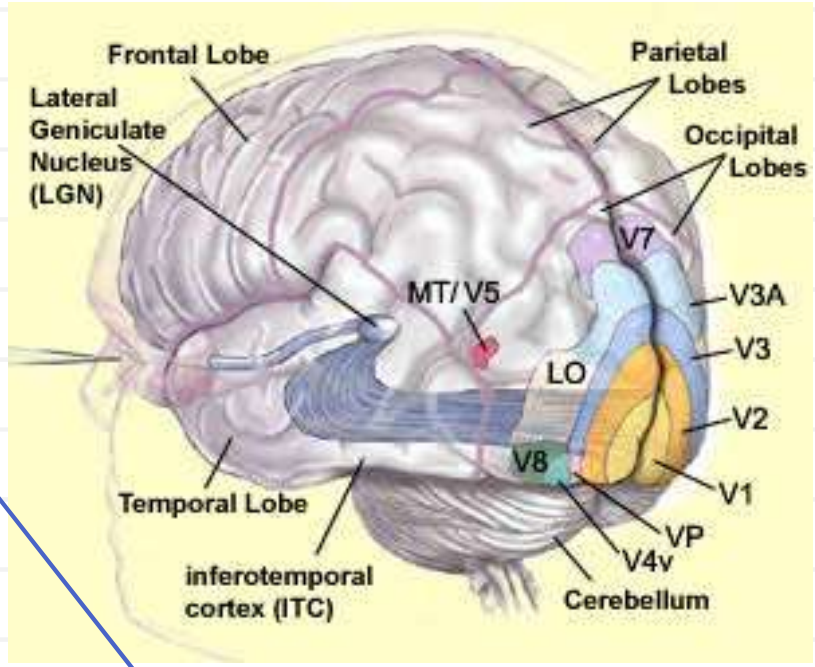



[Razavian et al. 2014]: ImageNet-trained network -> representations-> linear SVM or L2 NN-search

"Augmentation" = L2-normalize+PCA+ whitening+L2-normalize+power (all standard things used to postprocess shallow features)

# Visual cortex



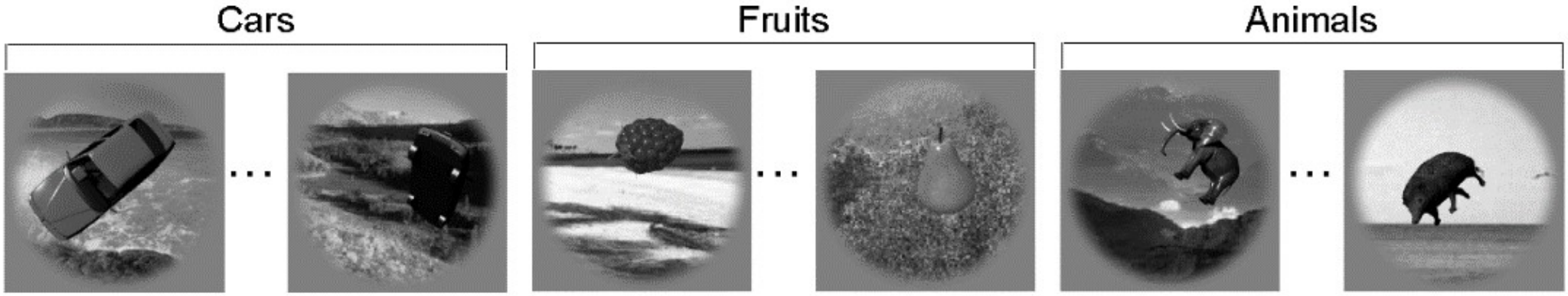

What? (objects, faces)

Where? (localizations,motions, actions)

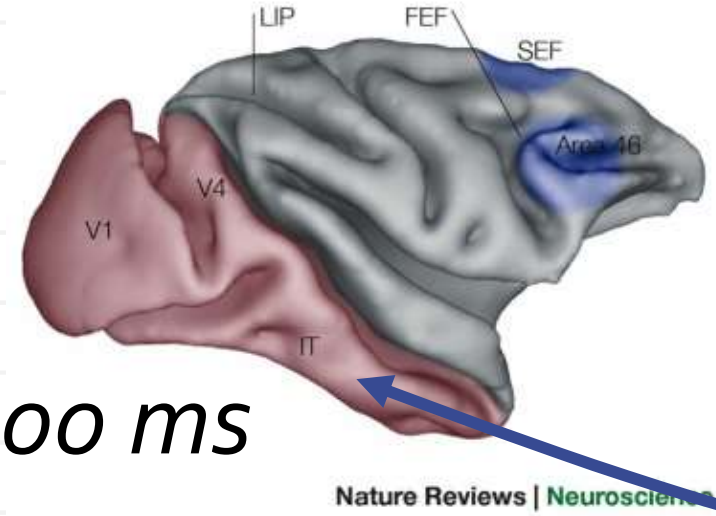
Source: "The brain from top to bottom" website

# Human vs machine

[Cadieu et al. 2014]




7 generic classes




*100 ms*

vs.

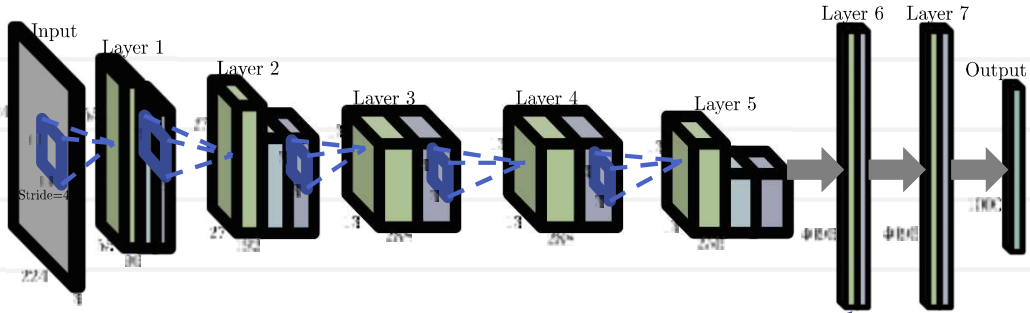
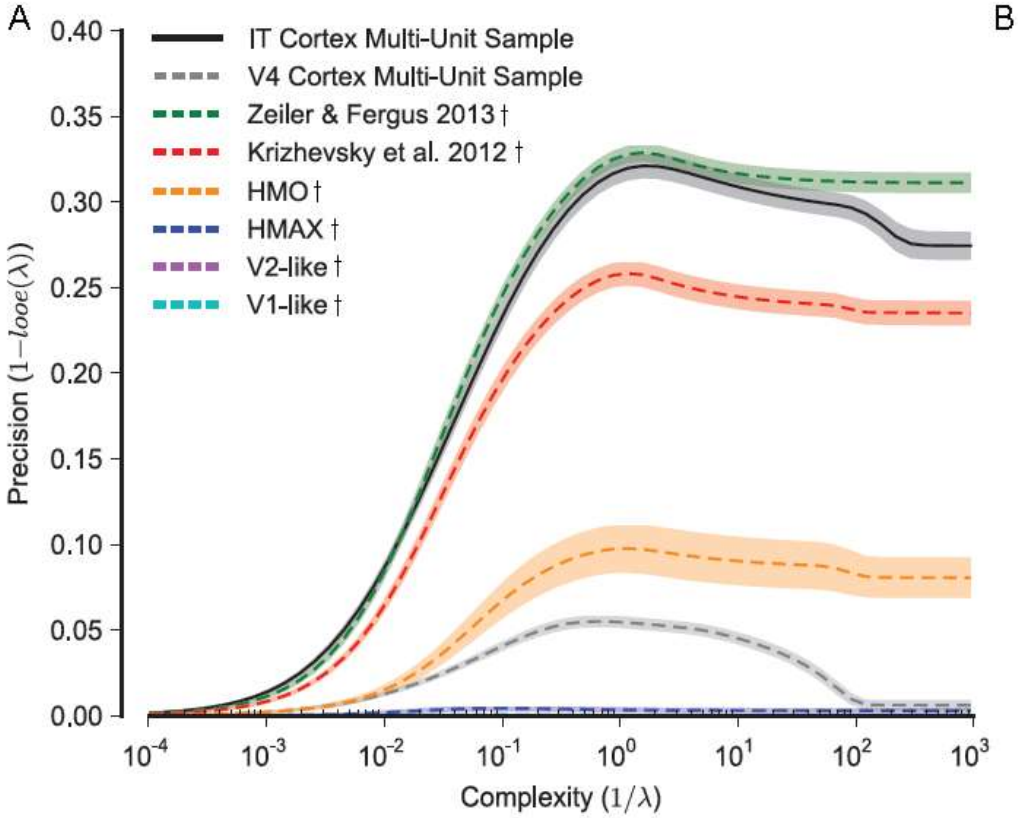



Image: [Sugrue et al. 2005]

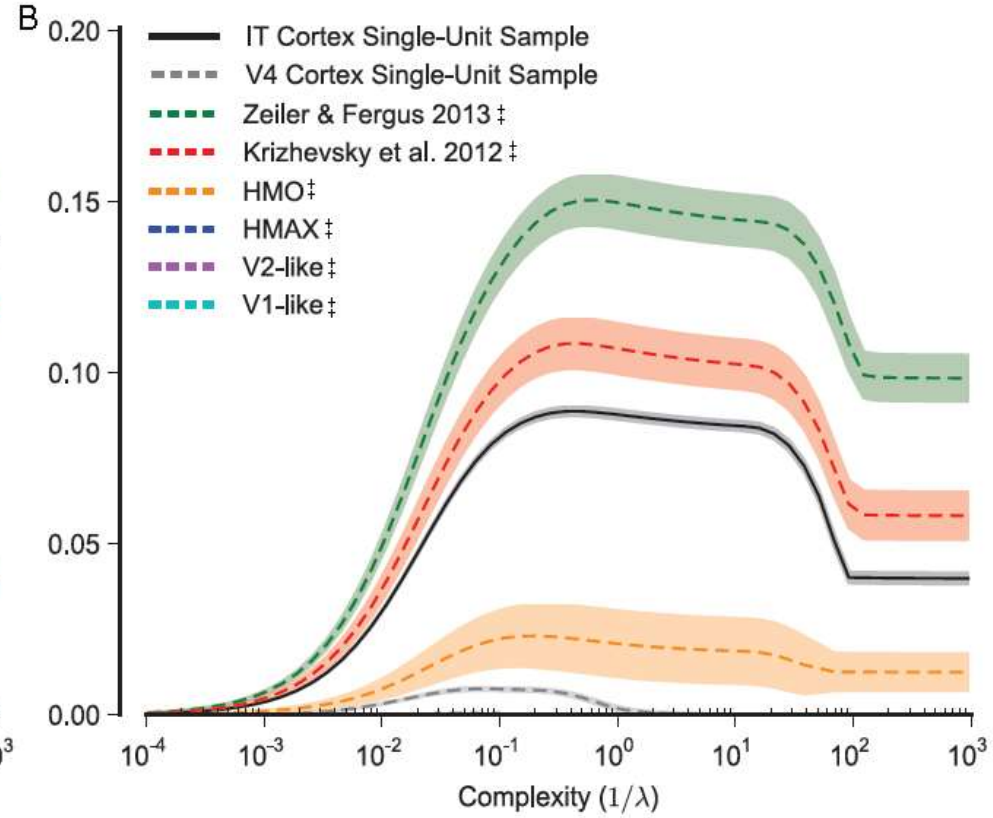
# Human vs machine

[Cadieu et al. 2014]

80 ("multi-unit") vs 80




40 (single neurons) vs 40

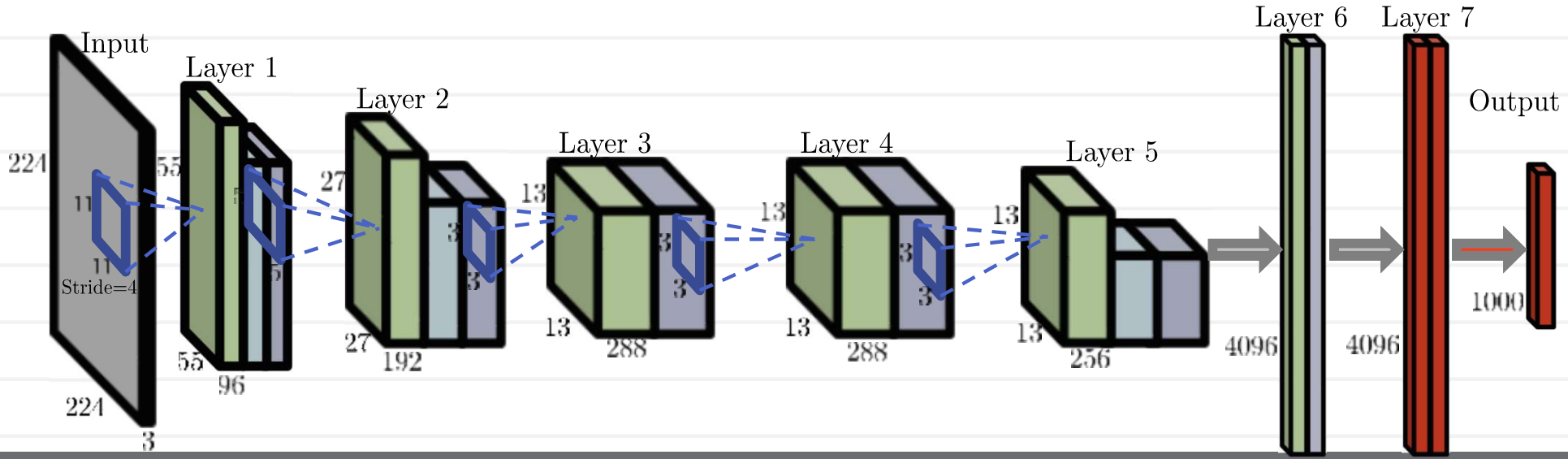



- Macaques were fixating on the images
- 100ms display, spiking rate recorded after that
- Averaging over 6 displays for each image

# Fine-tuning

Lots of papers (e.g. [Oquab et al. CVPR 14], [Babenko et al. ECCV 14], [Yosinski et al. NIPS14]) "fine-tune" the network on smaller datasets:

- Chop off top layers
- Initialize new top layers
- Run SGD



# Fine-tuning example

- Image-Net -> “Landmarks”

Oxford Buildings:

0.388->0.523

INRIA Holidays:

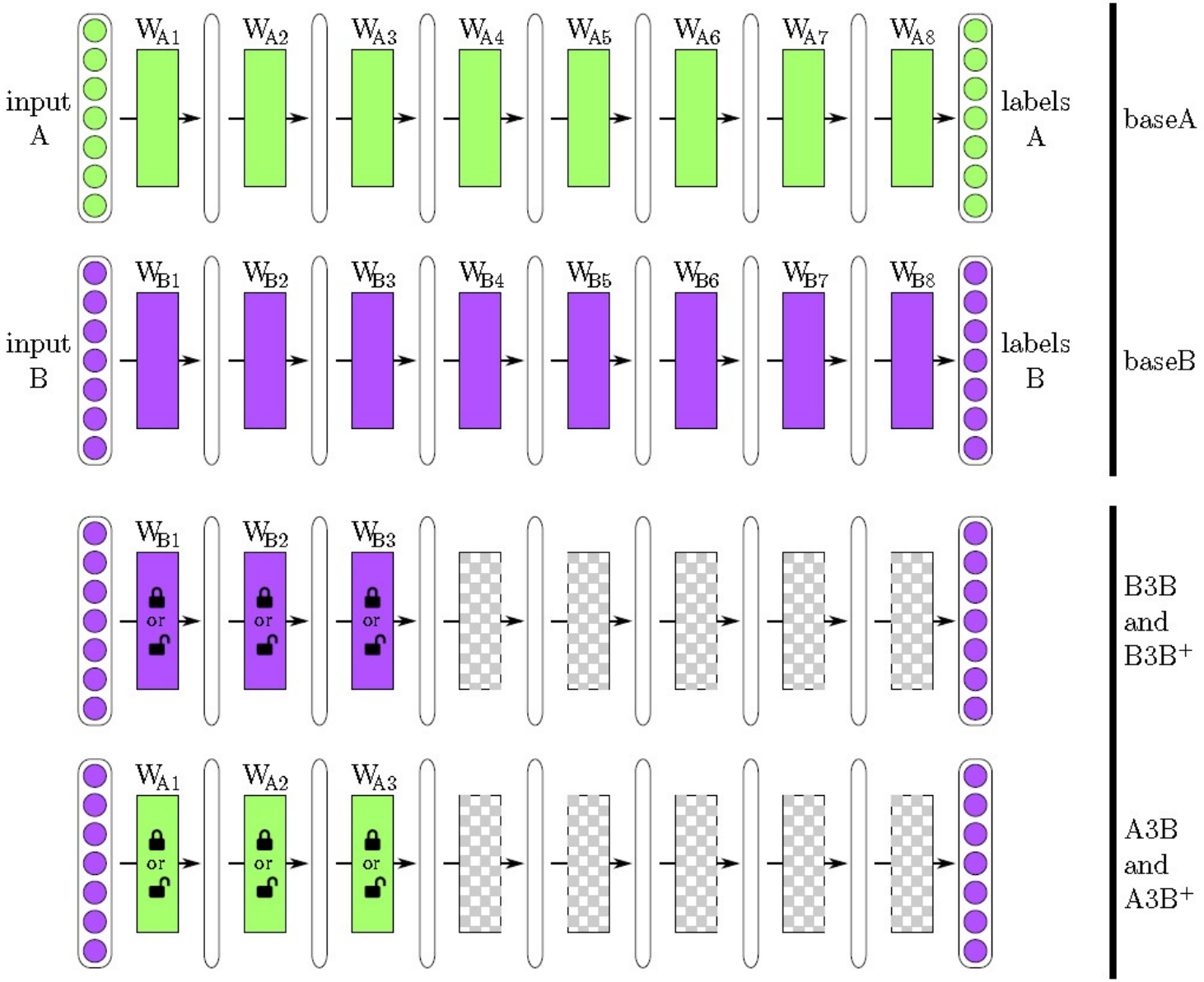
0.727->0.769

[Babenko et al. 2014]

# Fine-tuning examples

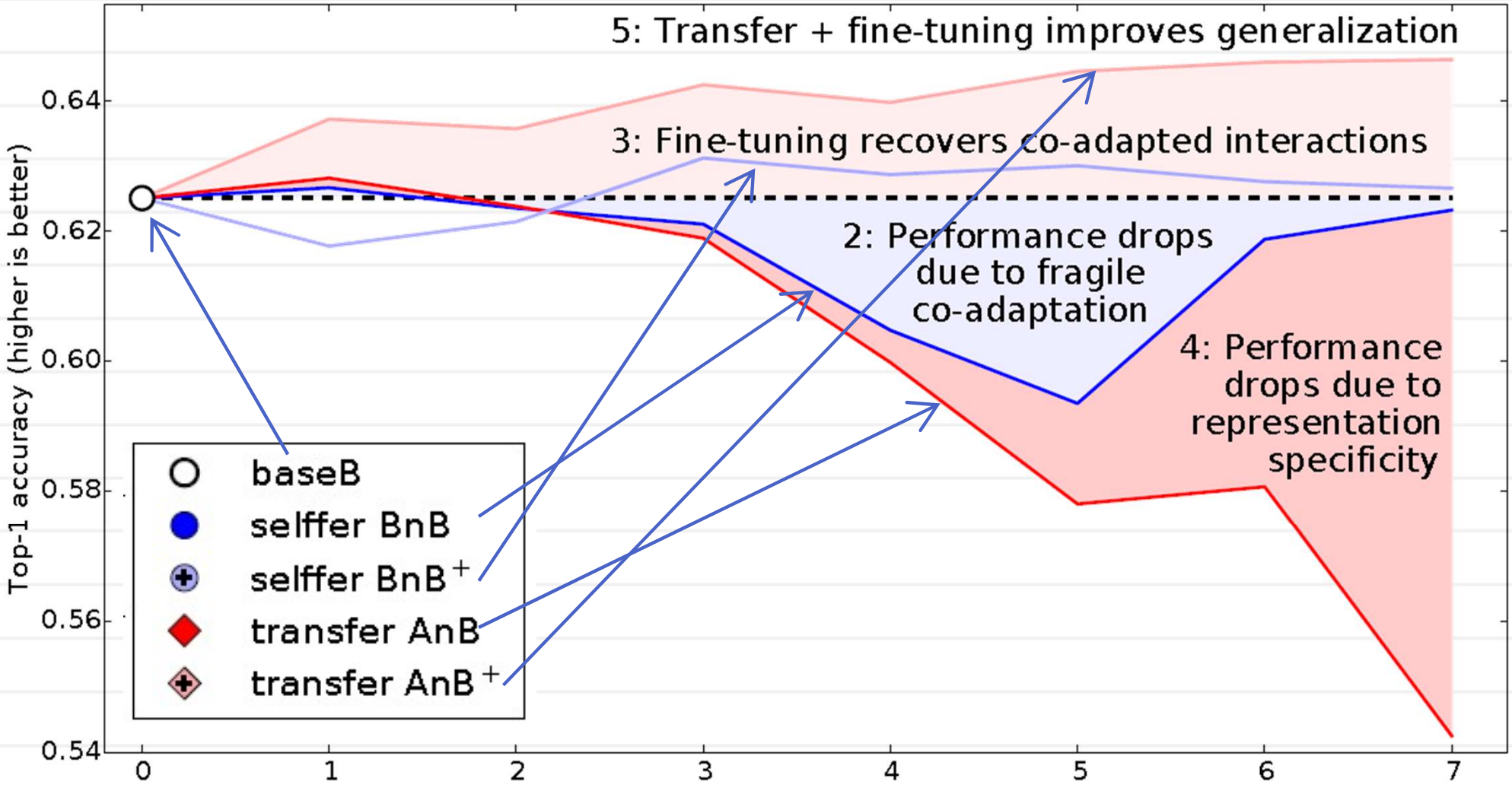
[Babenko et al. 2014]

# How to fine-tune




[Yosinski et al. 2014]: two halves of Image-NET

# How to fine-tune?




"+" = do not lock bottom layers

[Yosinski et al. NIPS2014]

# Applying CNN in practice

New problem


- Caffe zoo
- MatConvNet zoo
- Lasagne Recipes

# Literature


Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun:
Deep Residual Learning for Image Recognition. CoRR abs/1512.03385 (2015)

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton:
ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman:
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.CoRR abs/1312.6034 (2013)

Karen Simonyan, Andrew Zisserman:
Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556 (2014)

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan,Vincent Vanhoucke, Andrew Rabinovich:
Going deeper with convolutions. CVPR 2015: 1-9

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus:
Intriguing properties of neural networks. CoRR abs/1312.6199 (2013)

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke:
Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. CoRRabs/1602.07261 (2016)

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, Fei-Fei Li:
ImageNet Large Scale Visual Recognition Challenge. CoRR abs/1409.0575 (2014)

# Literature


Matthew D. Zeiler, Rob Fergus:
Visualizing and Understanding Convolutional Networks. ECCV (1) 2014: 818-833

Anh Mai Nguyen, Jason Yosinski, Jeff Clune:
Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.CVPR 2015: 427-436

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson:
CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. CVPR Workshops 2014: 512-519

Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson:
How transferable are features in deep neural networks? NIPS 2014, 3320-3328

Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic:
Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks.CVPR 2014: 1717-1724

Charles F. Cadieu, Ha Hong, Daniel Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj,James J. DiCarlo:
Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Computational Biology 10(12) (2014)

Artem Babenko, Victor S. Lempitsky:
Aggregating Local Deep Features for Image Retrieval. ICCV 2015: 1269-1277

Artem Babenko, Anton Slesarev, Alexander Chigorin, Victor S. Lempitsky:
Neural Codes for Image Retrieval. ECCV (1) 2014: 584-599

# Literature

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge:
Texture Synthesis Using Convolutional Neural Networks. NIPS 2015: 262-270

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge:
A Neural Algorithm of Artistic Style. CoRR abs/1508.06576 (2015)

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals:
Understanding deep learning requires rethinking generalization. CoRR abs/1611.03530 (2016)
[ICLR 2017]

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin A. Riedmiller:
Striving for Simplicity: The All Convolutional Net. CoRR abs/1412.6806 (2014)