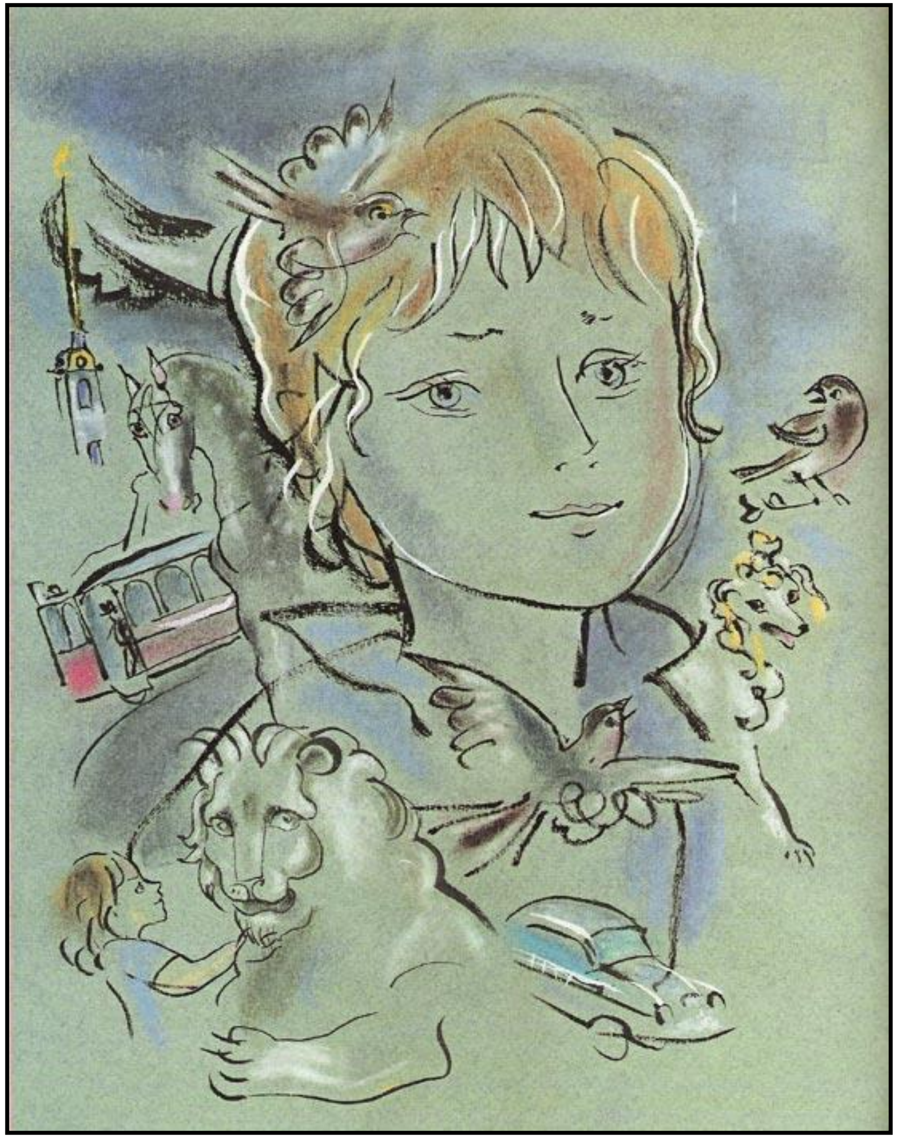
Recognition Techniques, old and new


2021-May-23 14:32:32.672 (BST)

Today’s outline

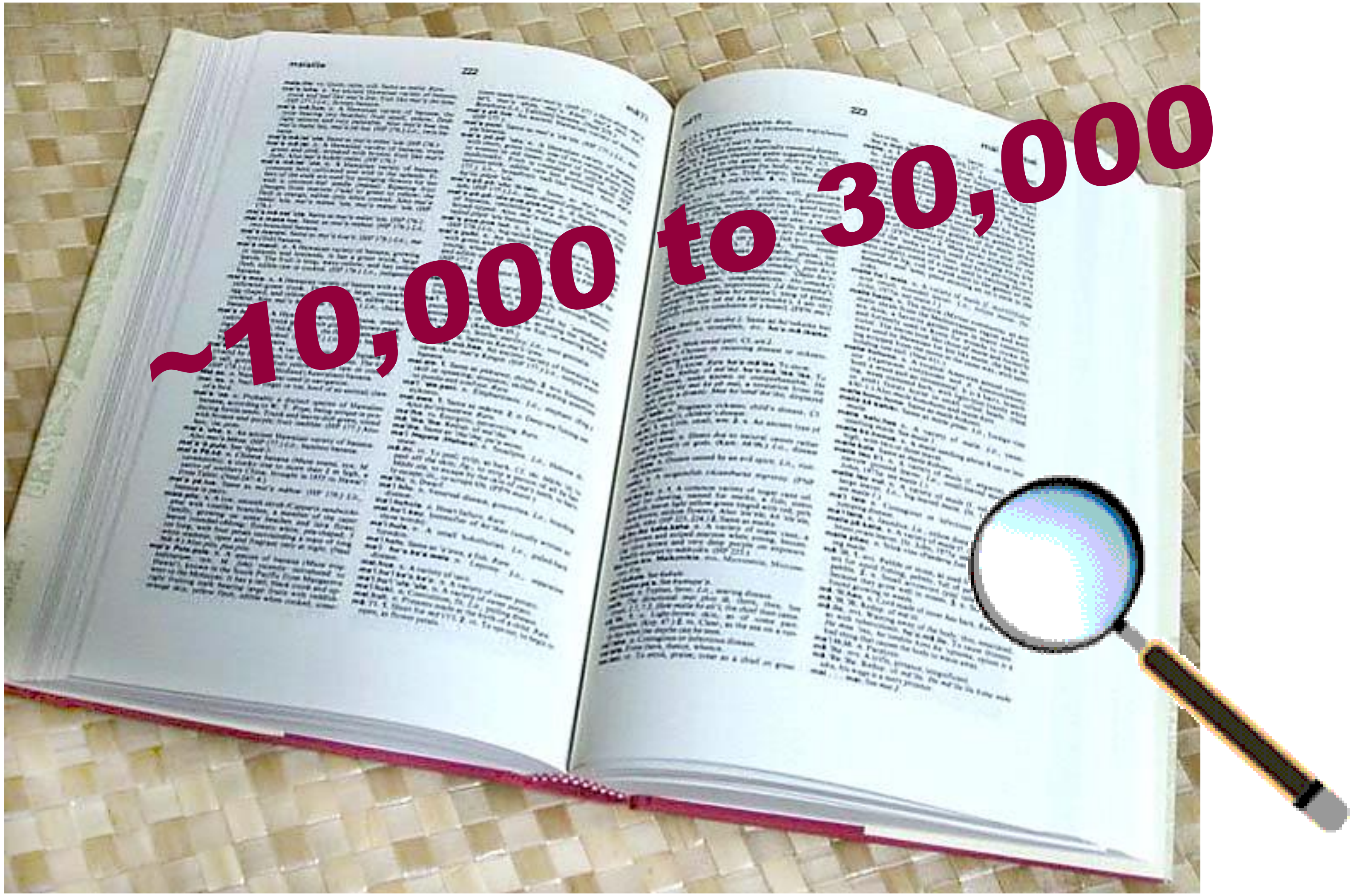
- We’ve covered Deep Convolutional Networks. But what did recognition techniques look like before AlexNet?
  - Bag of words models
  - Sliding window models
- What do more recent deep learning architectures look like?
  - VGG Net
  - Google Inception architectures
  - ResNet

Recognition: Overview and History



Slides from Lana Lazebnik, Fei-Fei Li, Rob Fergus, Antonio Torralba, and Jean Ponce

How many visual object categories are there?



Biederman 1987

But there's really a "heavy tail" of rarer object categories that humans can often understand after seeing few or no examples

And nothing really fits neatly into categories

Specific recognition tasks

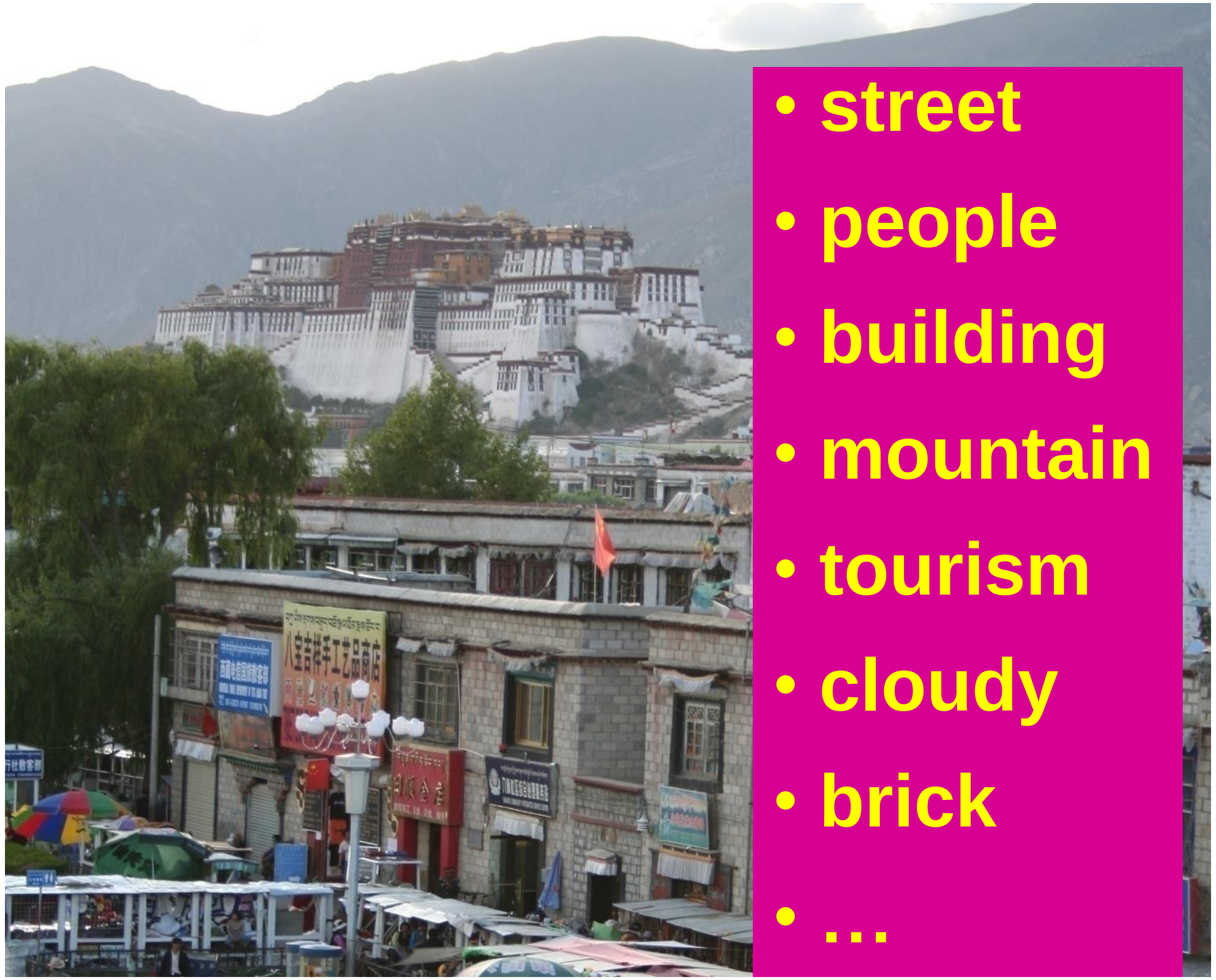
Svetlana Lazebnik

Scene categorization or classification

- **outdoor/indoor**
- **city/forest/factory/etc.**

Svetlana Lazebnik

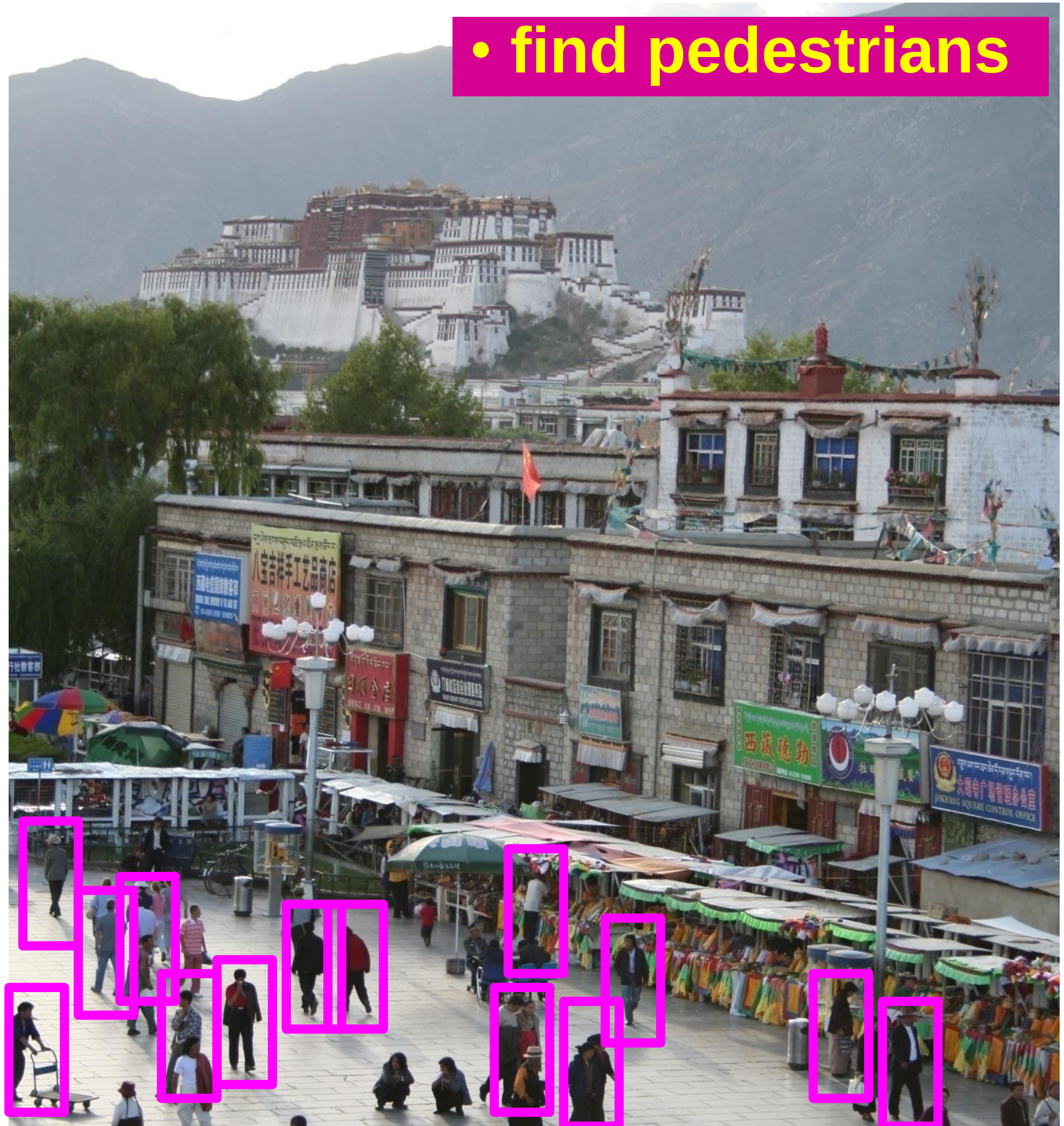
Image annotation / tagging / attributes



Categories are exclusive. An instance belongs to one category.
Attributes are not exclusive. An instance can have many or none.

Svetlana Lazebnik

Object detection



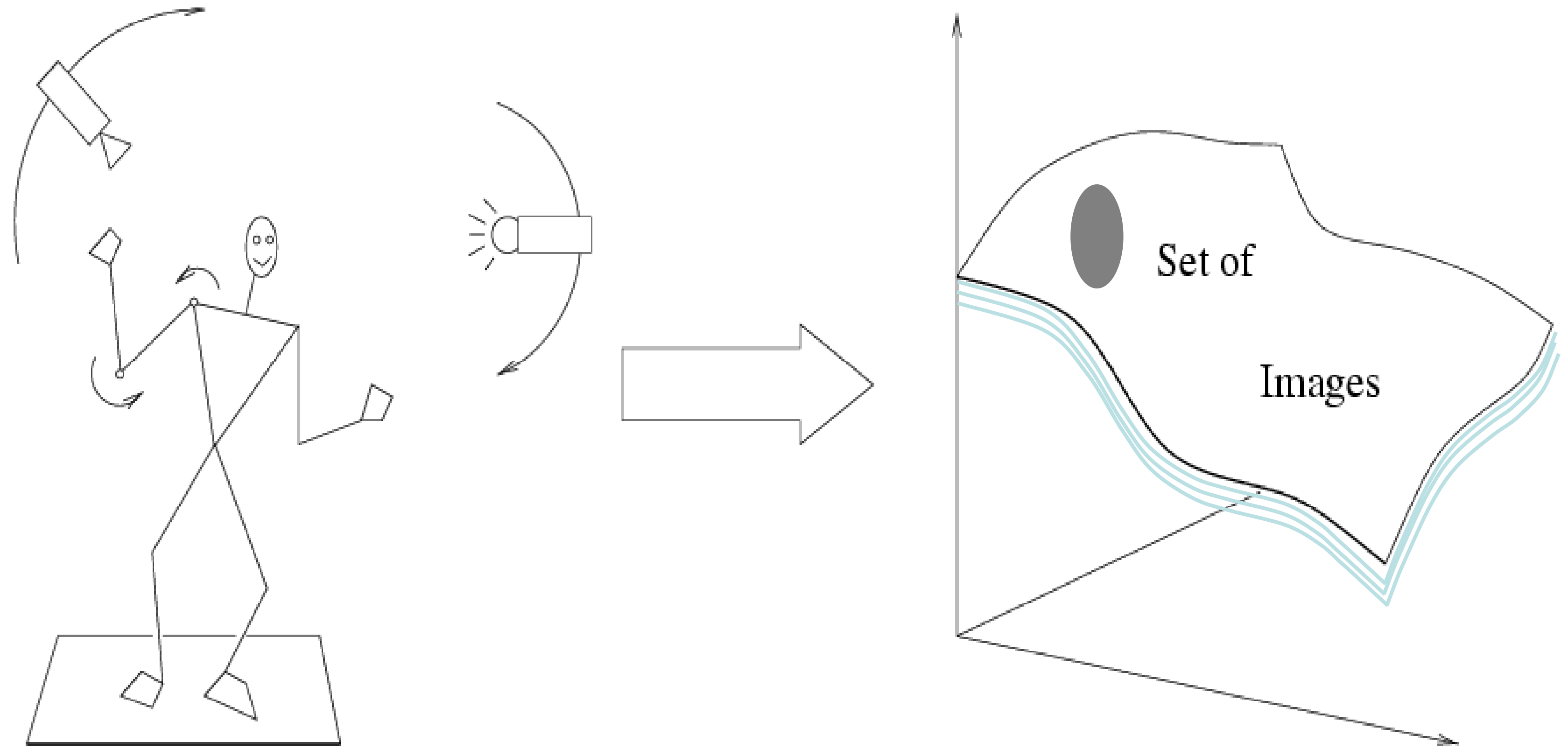
Svetlana Lazebnik

Image parsing / semantic segmentation

"Segmentation" doesn't count objects.
Not everything is a countable object, anyway

Svetlana Lazebnik

Recognition is all about modeling variability



Variability of a single object instance due to:

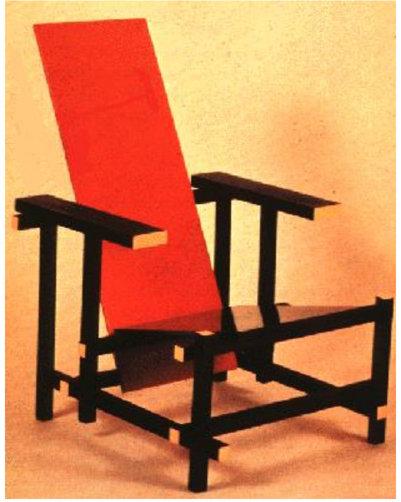
Camera position
Illumination
Shape parameters

Within-class variations among multiple object instances?

Svetlana Lazebnik

Within-class variations



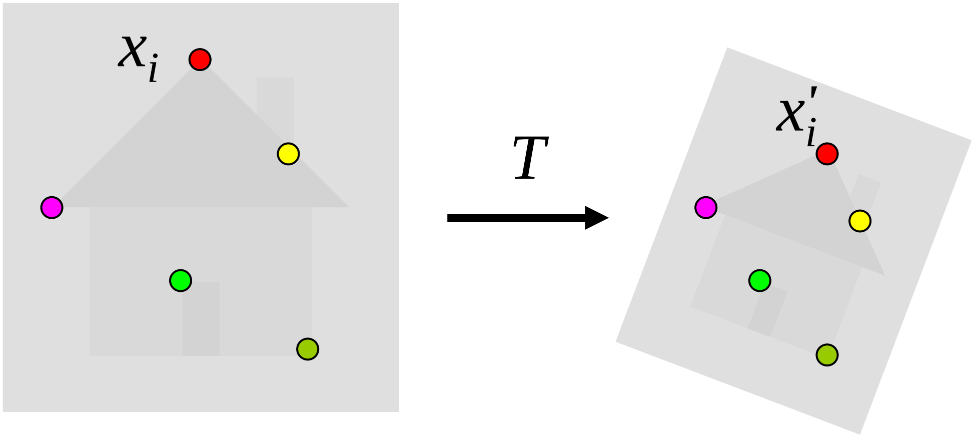
Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era

Recall: Alignment

- Alignment: fitting a model to a transformation between pairs of features (*matches*) in two images

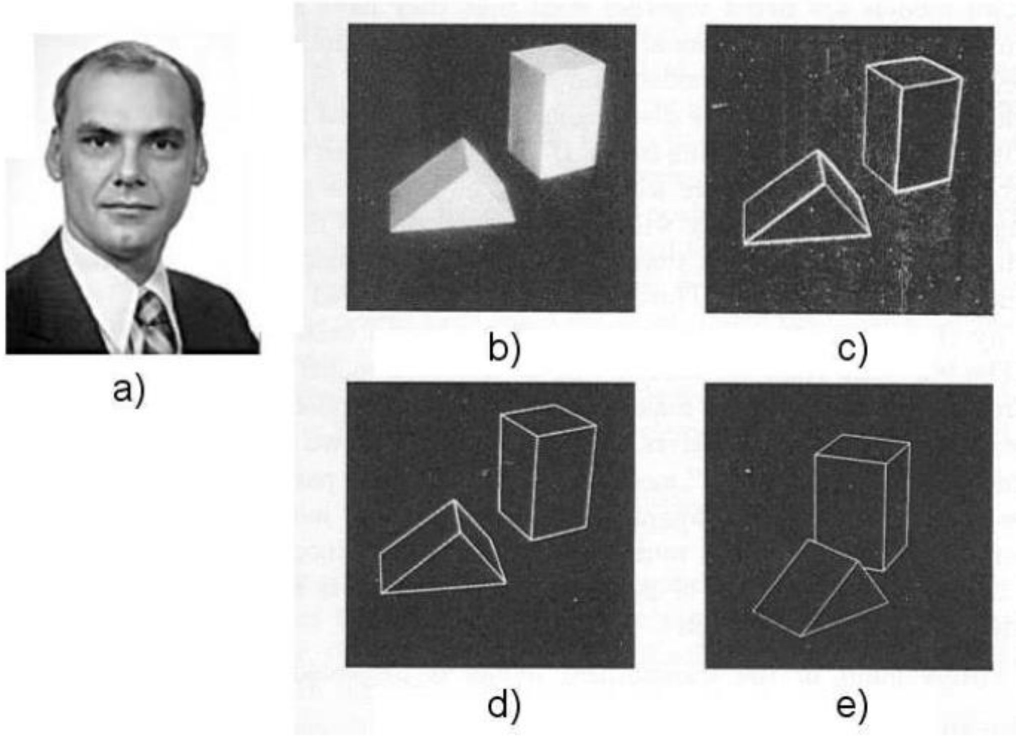


Find transformation T that minimizes

$$\sum_i \text{residual}(T(x_i), x_i')$$

Svetlana Lazebnik

Recognition as an alignment problem: Block world



L. G. Roberts, *Machine Perception of Three Dimensional Solids,* Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - e) are taken from [64] with permission MIT Press.)

J. Mundy, Object Recognition in the Geometric Era: a Retrospective, 2006

ABSTRACT

In order to make it possible for a computer to construct and display a three-dimensional array of solid objects from a single two-dimensional photograph, the rules and assumptions of depth perception have been carefully analyzed and mechanized. It is assumed that a photograph is a perspective projection of a set of objects which can be constructed from transformations of known three-dimensional models, and that the objects are supported by other visible objects or by a ground plane. These assumptions enable a computer to obtain a reasonable, three-dimensional description from the edge information in a photograph by means of a topological, mathematical process.

A computer program has been written which can process a photograph into a line drawing, transform the line drawing into a three-dimensional representation, and finally, display the three-dimensional structure with all the hidden lines removed, from any point of view. The 2-D to 3-D construction and 3-D to 2-D display processes are sufficiently general to handle most collections of planar-surfaced objects and provide a valuable starting point for future investigation of computer-aided three-dimensional systems.

referred to. Let us fix the real world coordinates by assuming that the focal plane is the $x=0$ plane and the focal point is at $x=f$, $y=0$, $z=0$. In order that the picture not be a reflection, we choose the focal plane in front of the camera. Then the objects seen will be in the $-x$ half space. Thus, the focal plane is really the plane of the print, not the negative. Figure 1 shows this arrangement.

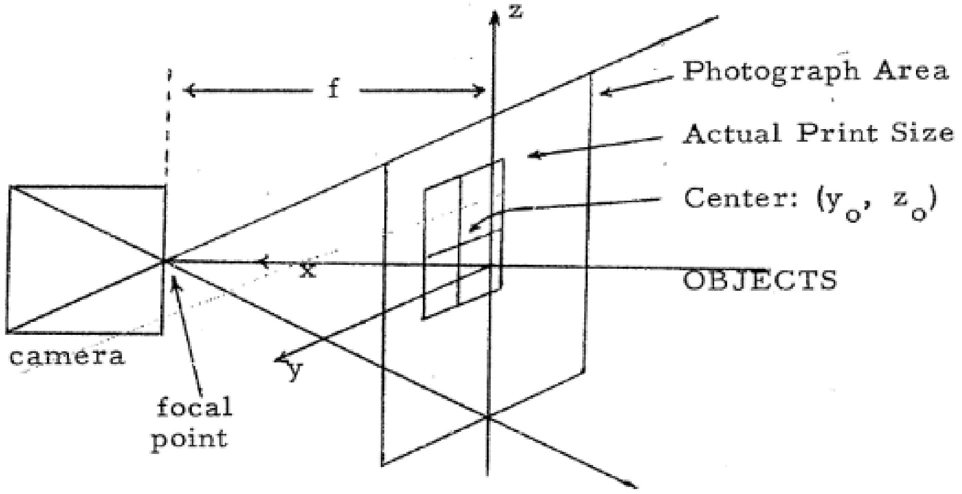


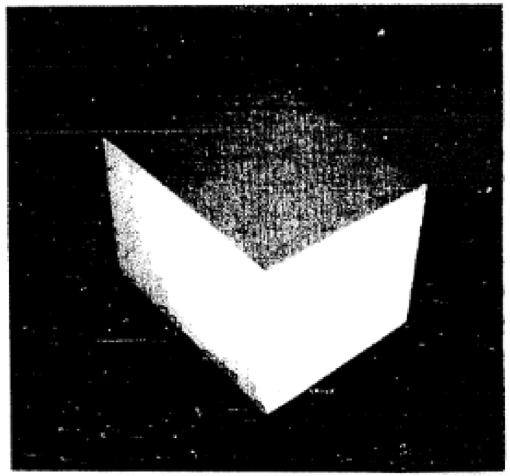
Figure 1: Camera Transformation

A particular camera will have some focal distance f. We will consider the square on the focal plane which was enlarged to create the print. The center of this square will be at some coordinates y_o, z_o, and the size of the square from the center to an edge will be some

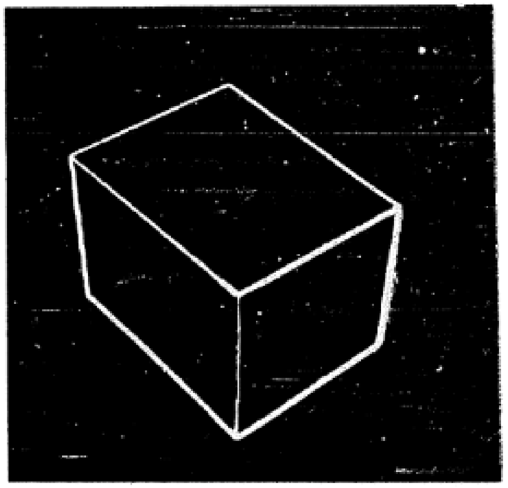
It is not necessary to know the variables y_o, z_o, f, and S since they can be computed from the picture given other assumptions later on. However, for the sake of simplicity we will assume S/f is known and that $y_o = z_o = 0$. The numerical values of S and f alone are not necessary since this just affects the scale of the real world. Thus, we can assume $S = 1$ and with $r = S/f$ obtain a simple transformation P.

$$P = \begin{bmatrix} 1 & & & -r \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$$

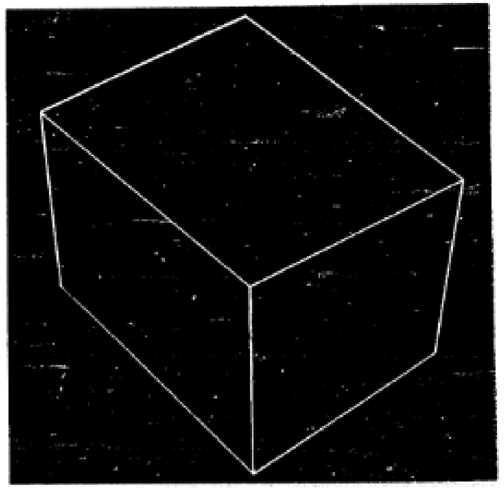
If $\bar{v}$ is a point in real space, then $\bar{v}P$ is a point in a perspective space such that its Y and Z coordinates are the original point's projection on the picture plane. The X coordinate of $\bar{v}P$ is also obtained and will be useful for hidden line computation during display of 3-D objects.



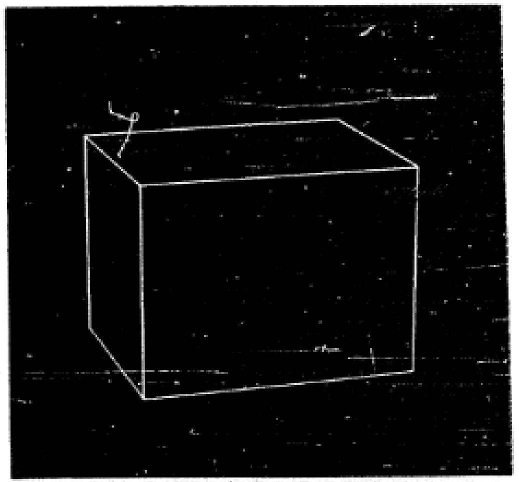
A. Original Picture



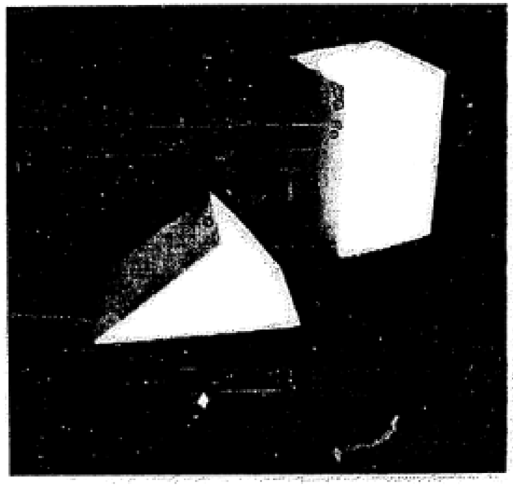
B. Differentiated Picture



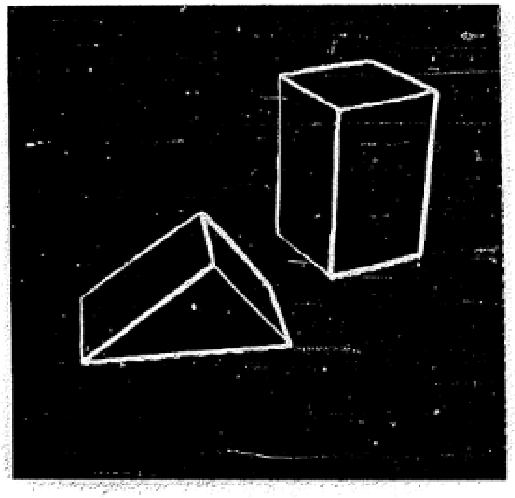
C. Line Drawing



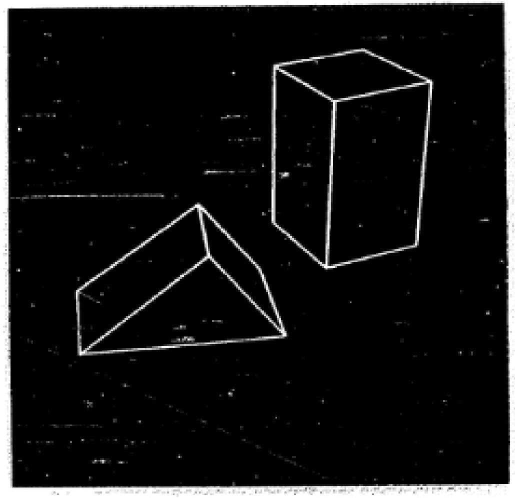
D. Rotated View



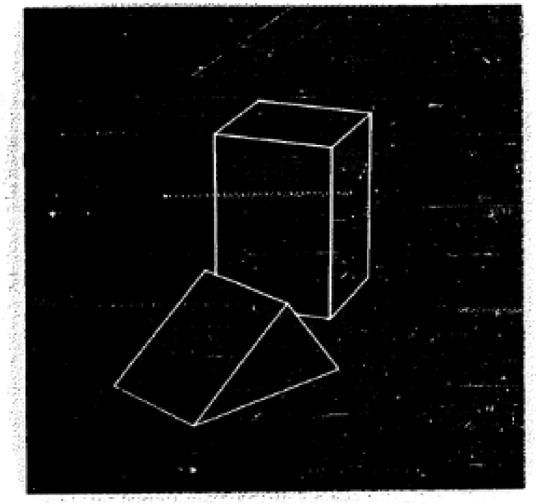
A. Original Picture



B. Differentiated Picture



C. Line Drawing



D. Rotated View

The input program has about 5000 instructions and uses over 40,000 registers of data storage for its pictures and lists. It takes about one minute to process a picture into a line drawing of which half is for differentiation. The 3-D construction and display programs are each about 3000 instructions and use from 5000 to 40,000 registers of data storage depending upon the number of objects. Both construction and display take about one second per object. All told, a rotated view of the objects in a photograph might be obtained in two minutes.

As far as machine depth perception, the only work I know of is on binocular images. Julesz has reported a procedure which shifts the binocular pictures to find the areas at different depths.[5] This procedure uses only texture, not edges, to develop the depth information and shows that the binocular information alone is sufficient for depth perception. This work is similar in goal but completely different in procedure from mine. Other work in machine photograph processing has mainly been in the field of information reduction for bandwidth compression and my paper in this area summarizes this work.[6]

1. Somerville, D.M.Y., Analytical Geometry of Three Dimensions, Cambridge University Press, 1959.

2. Roberts, L.G., "Pattern Recognition With An Adaptive Network," IRE International Convention Record, Pt. 2, pp. 66-70, 1960.

3. Selfridge, O.G., and U. Neisser, "Pattern Recognition by Machine," Scientific American, Vol. 203, No. 3, pp 60-68, August, 1960.

4. Hodes, L., Machine Processing of Line Drawings," Lincoln Laboratory, MIT, Lexington, Mass., Group Report 54G0028, March, 1961.

5. Julesz, B., "Toward the Automation of Binocular Depth Perception," Proceedings of the I.F.I.P. Congress, Munich, 1962.

6. Roberts, L.G., "Picture Coding Using Pseudo-Random Noise," IRE Trans. on Information Theory, Vol. IT-8, No. 2, pp 145-154, February, 1962.

7. Gibson, J.J., The Perception of the Visual World, H. Mifflin Company, Boston, Mass., 1950.

8. Ittelson, W.H., "Size As a Cue to Distance," American J. Psychology, Vol. 64, pp 54-67, 1951.

9. Attneave, F. and Arnoult, "The Quantitative Study of Shape and Pattern Perception," Psychological Bull., Vol. 53, p 452, 1956.

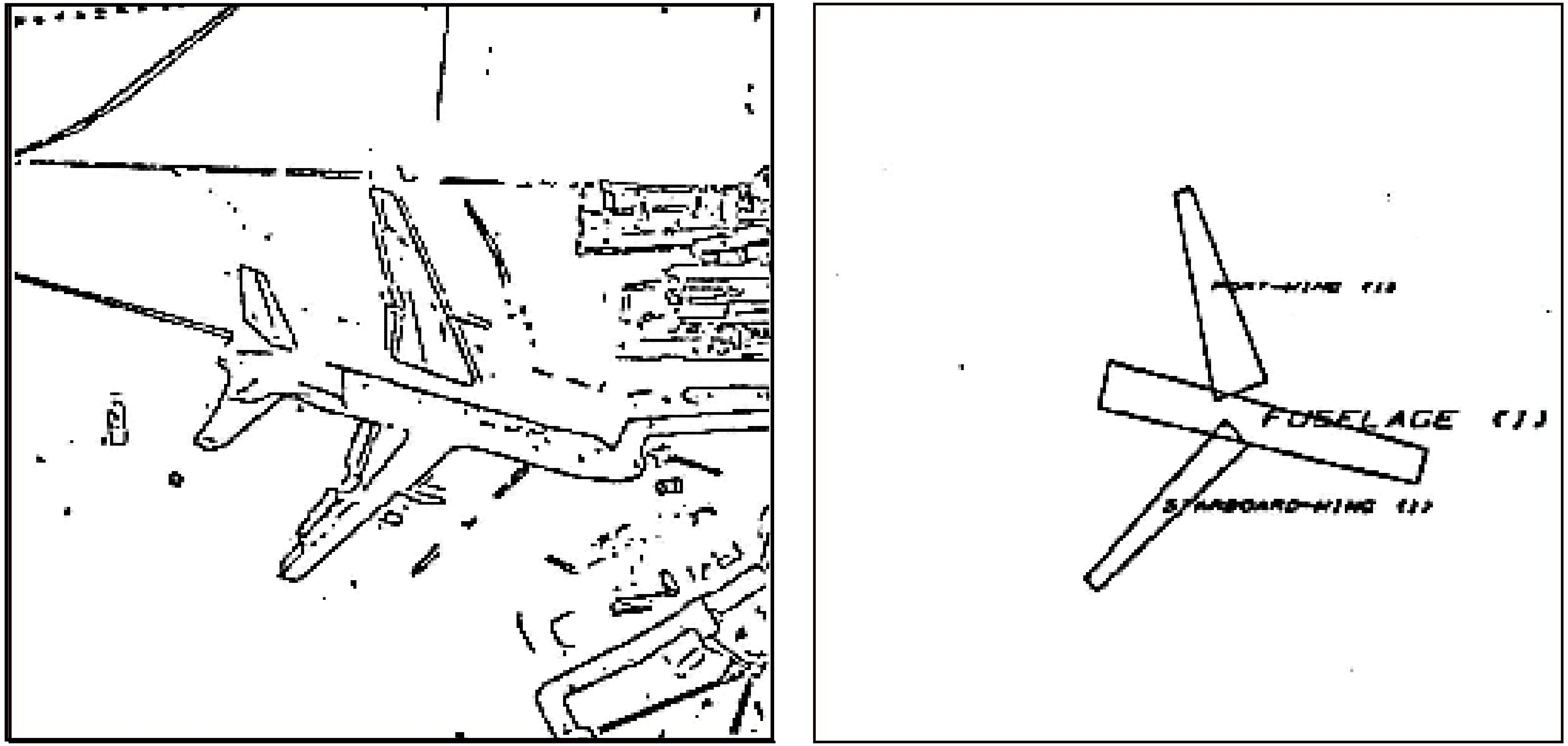
10. Langdon, J., "The Perception of 3-D Solids," Quar. J. Exp. Psychology, Vol. 7, 1955.

11. Stevens, S.S., "The Psychophysiology of Vision," in Sensory Communication, W. Rosenblith, Editor, MIT Press, Cambridge, and John Wiley and Sons, New York, N.Y., p. 13, 1961.

12. Sutherland, I.E., Sketchpad, A Man-Machine Graphical Communication System, Ph.D. Thesis, Massachusetts Institute of Technology, Electrical Engineering Department, Cambridge, Mass., February, 1963.

13. Johnson, T., Sketchpad III, 3-D, Graphical, Communication with a Digital Computer, Masters Thesis, Massachusetts Institute of Technology, Mechanical Engineering Department, Cambridge, Mass., June, 1963.

Representing and recognizing object categories is harder...



ACRONYM (Brooks and Binford, 1981)

Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

Russia Covering Aircraft With Tires Is About Confusing Image-Matching Missile Seekers U.S. Military Confirms

Russia's efforts to befuddle cruise missiles and drones with imaging-matching seeker capabilities speaks to issues that go beyond the war in Ukraine.

JOSEPH TREVITHICK / UPDATED ON SEP 13, 2024 7:50 PM EDT / 153

Schuyler Moore, U.S. Central Command's (CENTCOM) first-ever Chief Technology Officer, mentioned the Russian use of tires to disrupt incoming attacks on air bases during a broader live-streamed roundtable talk on artificial intelligence (AI) and related technologies that the Center for Strategic & International Studies (CSIS) think tank hosted today. Before taking up her current role, Moore had been Chief Strategy Officer for U.S. Naval Forces Central Command's (NAVCENT) Task Force 59, which is tasked with experimenting with integrating new AI-driven and uncrewed capabilities into day-to-day naval operations in the Middle East.

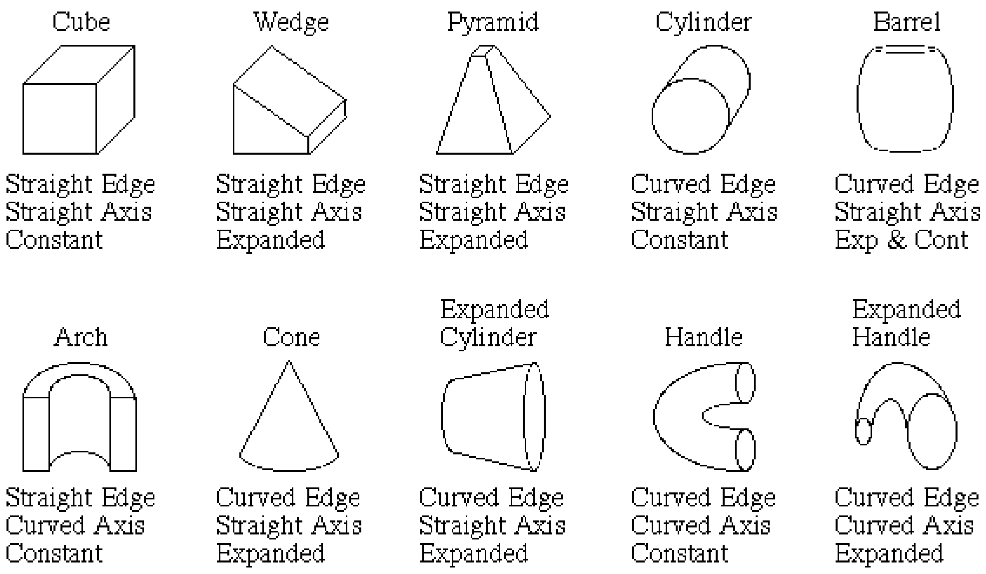
A "sort of classic unclassified example that exists is like a picture of a plane from the top, and you're looking for a plane, and then if you put tires on top of the wings, all of a sudden, a lot of computer vision models have difficulty identifying that that's a plane," Moore said as part of a larger discussion about AI models and data sets.

https://www.twz.com/air/russia-covering-its-aircraft-in-tires-is-about-befuddling-image-matching-seekers-u-s-military-confirms

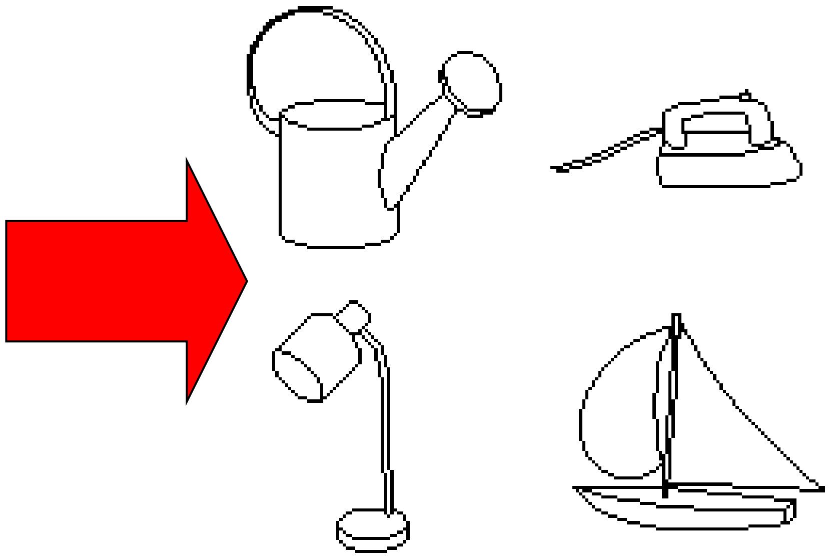
Recognition by components

Biederman (1987)

Primitives (geons)



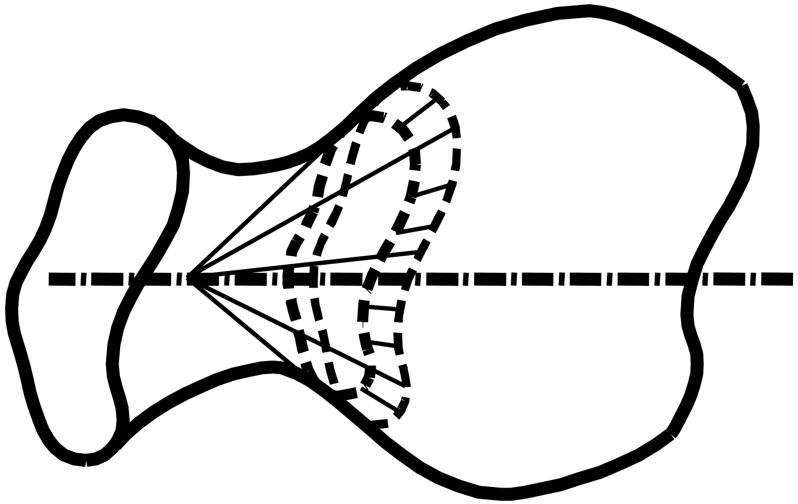
Objects



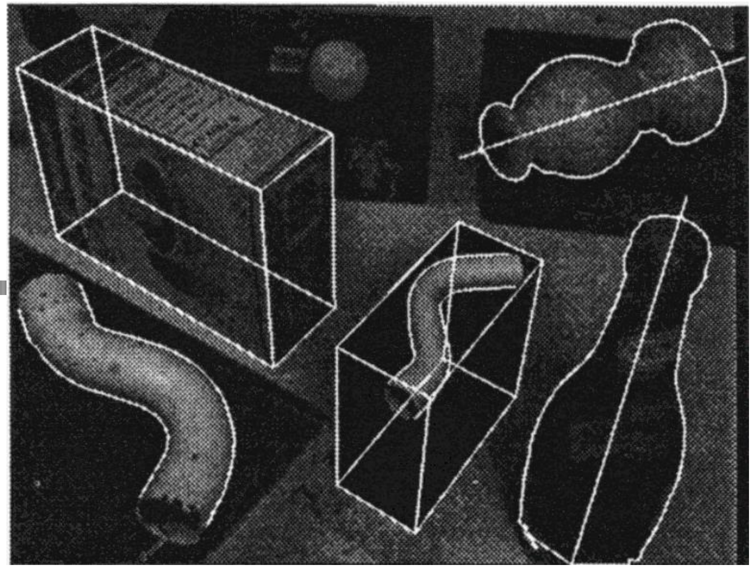
http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

Svetlana Lazebnik

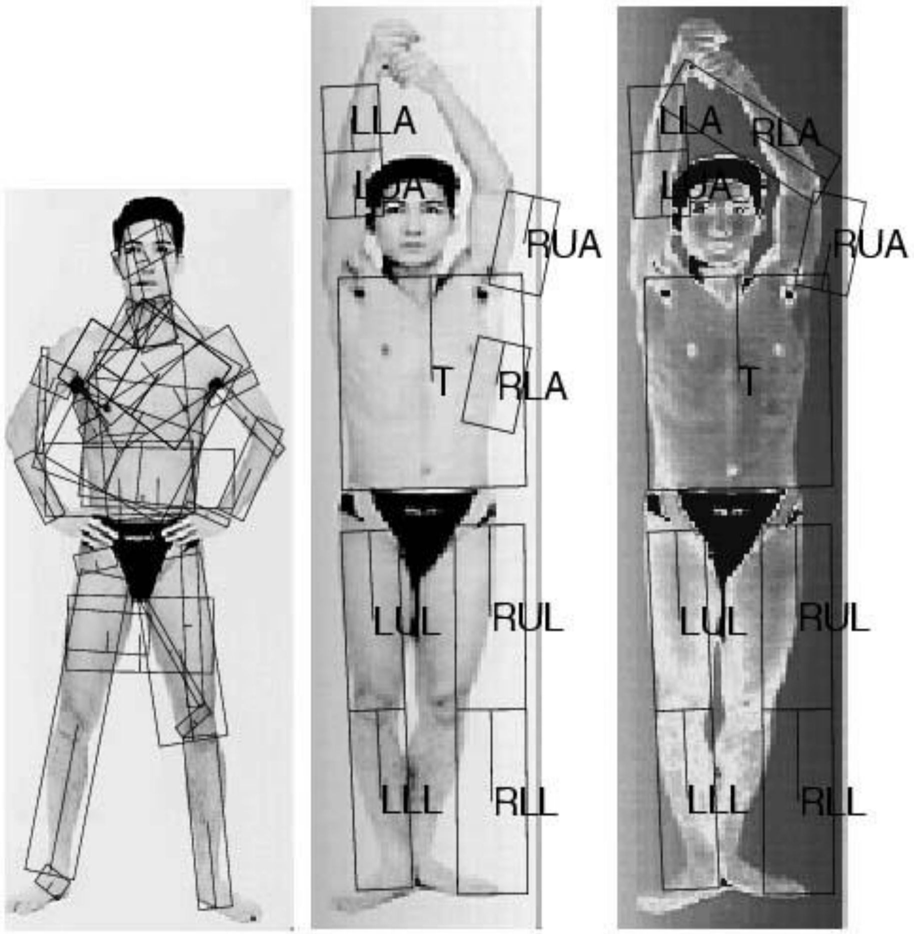
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



Zisserman et al. (1995)

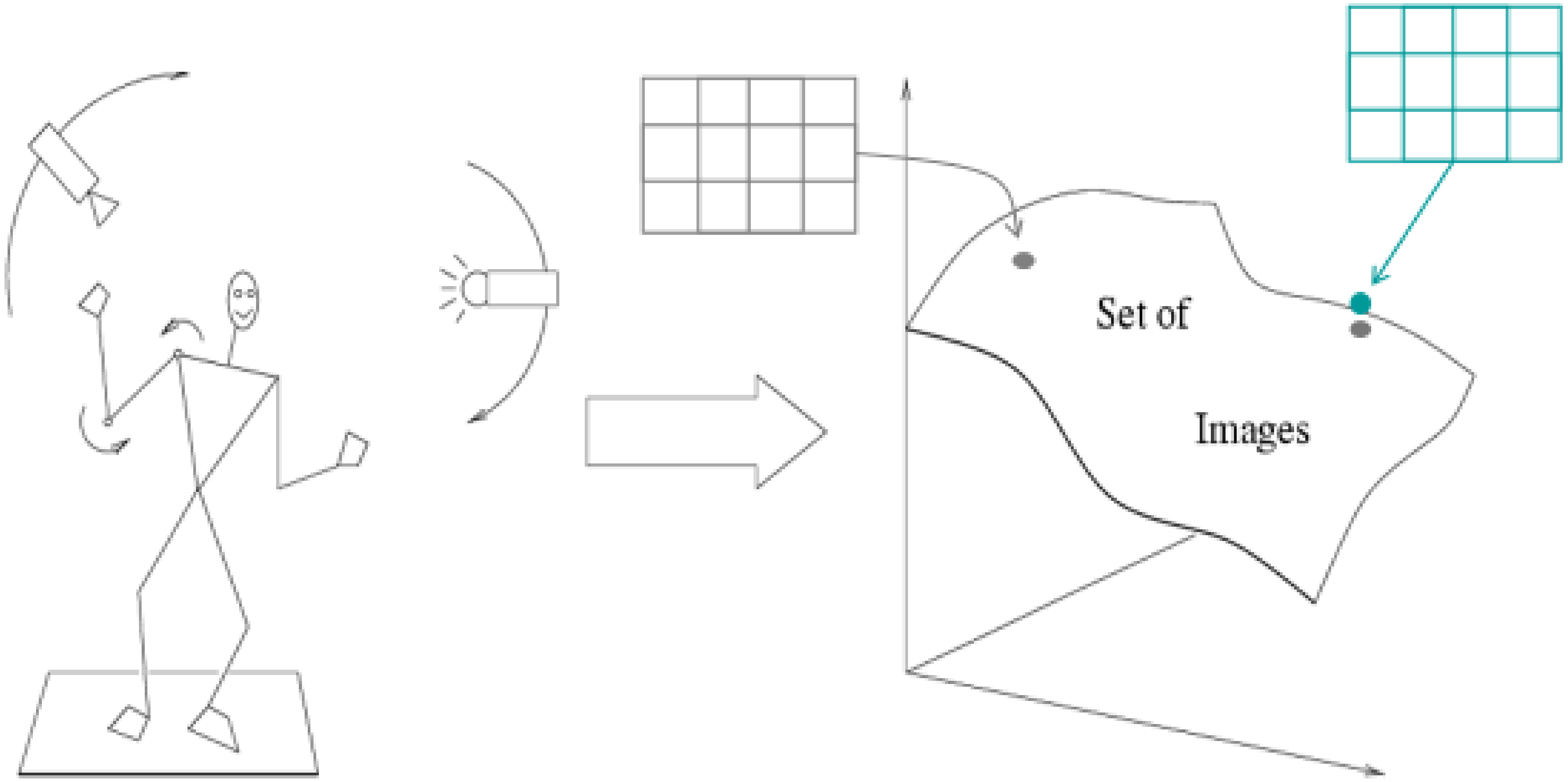


Forsyth (2000)

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models



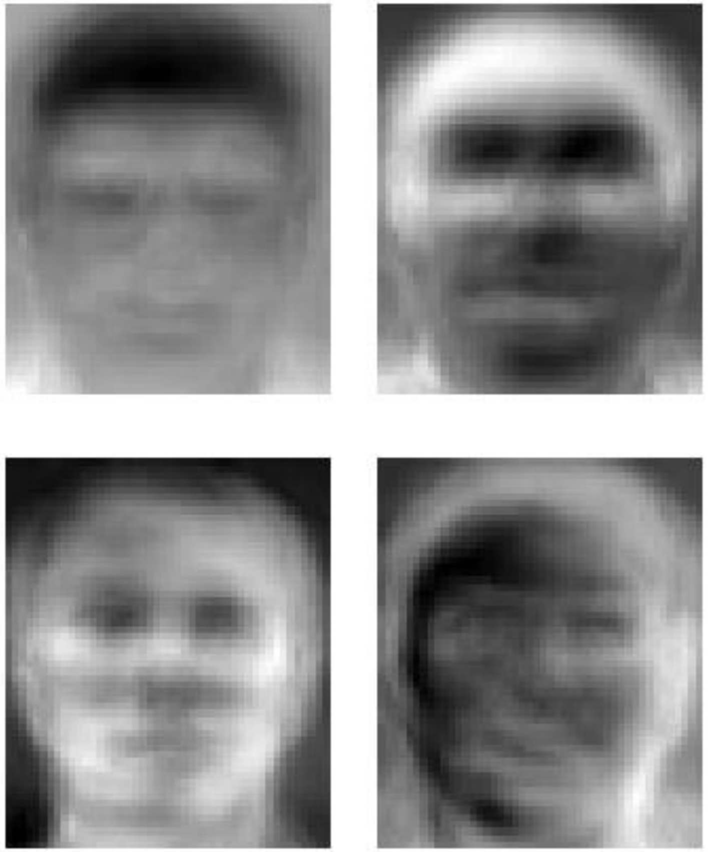
Empirical models of image variability

Appearance-based techniques

Turk & Pentland (1991); Murase & Nayar (1995); etc.

Svetlana Lazebnik

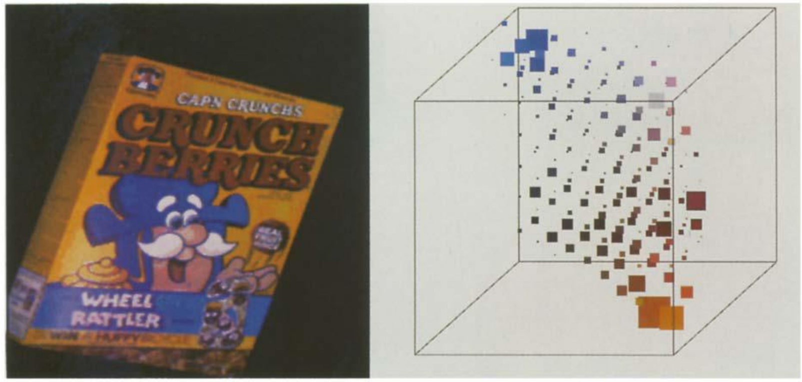
Eigenfaces (Turk & Pentland, 1991)

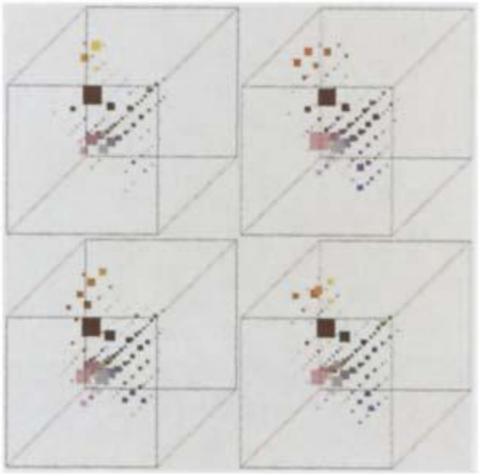


Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

Svetlana Lazebnik

Color Histograms





Swain and Ballard, Color Indexing, IJCV 1991.

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- 1990s – present: sliding window approaches

Svetlana Lazebnik

Sliding window approaches

Sliding window approaches

- Turk and Pentland, 1991
- Belhumeur, Hespanha, & Kriegman, 1997
- Schneiderman & Kanade 2004
- Viola and Jones, 2000

- Schneiderman & Kanade, 2004
- Argawal and Roth, 2002
- Poggio et al. 1993

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features

Svetlana Lazebnik

Local features for object instance recognition

D. Lowe (1999, 2004)

Large-scale image search

Combining local features, indexing, and spatial constraints

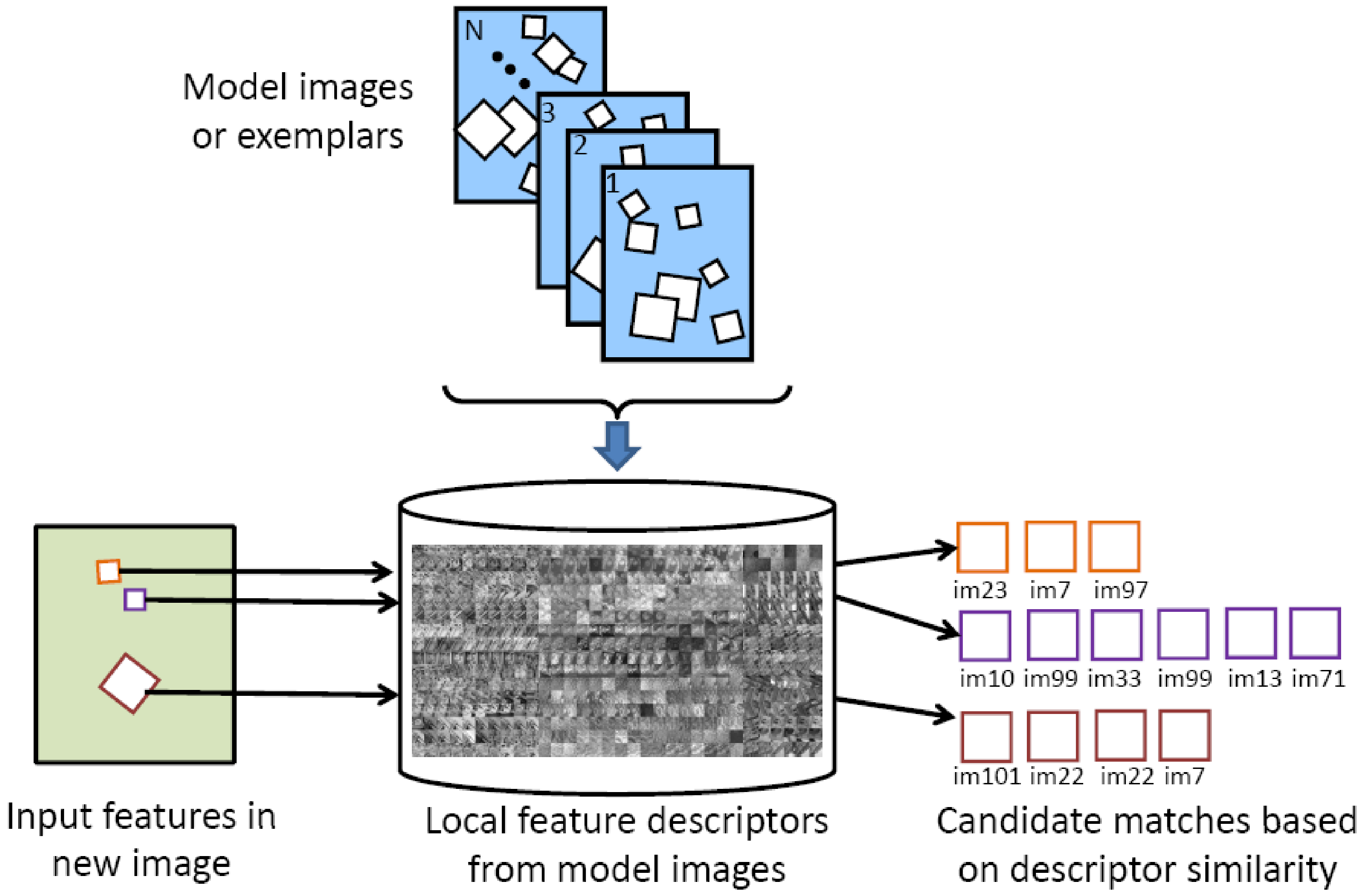
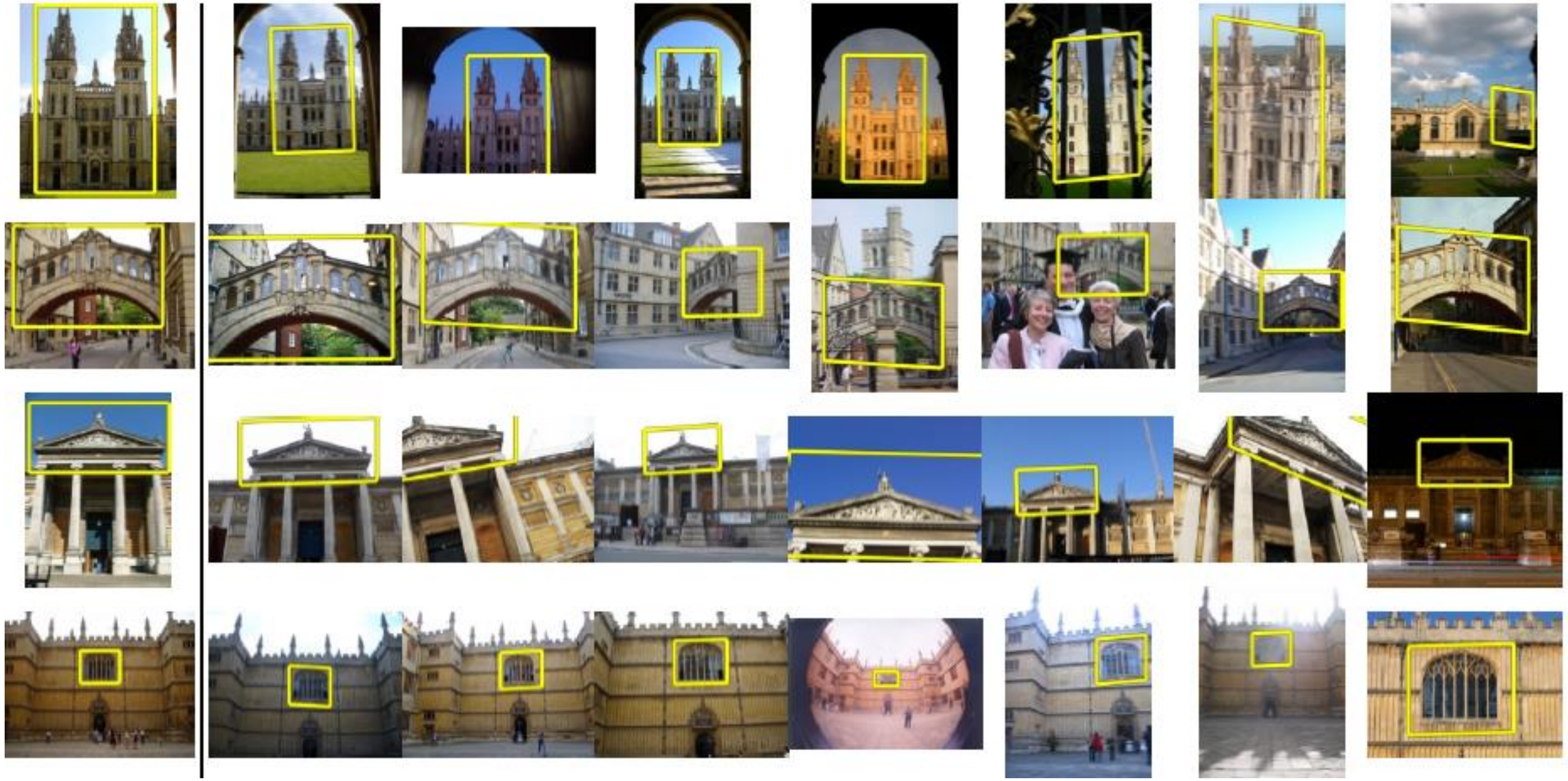


Image credit: K. Grauman and B. Leibe

Large-scale image search

Combining local features, indexing, and spatial constraints



Philbin et al. ‘07

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models

Parts-and-shape models

- Model:
  - Object as a set of parts
  - Relative locations between parts
  - Appearance of part

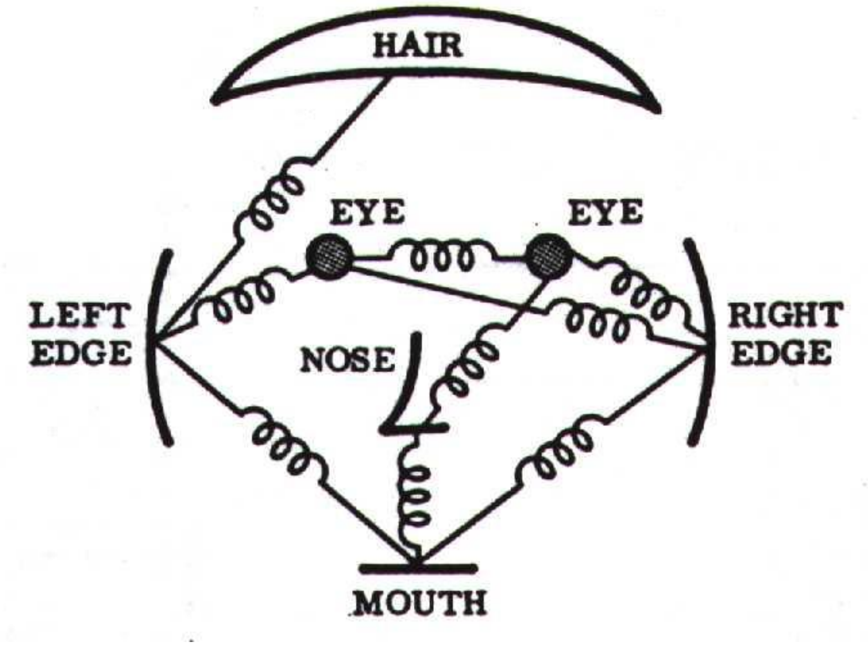
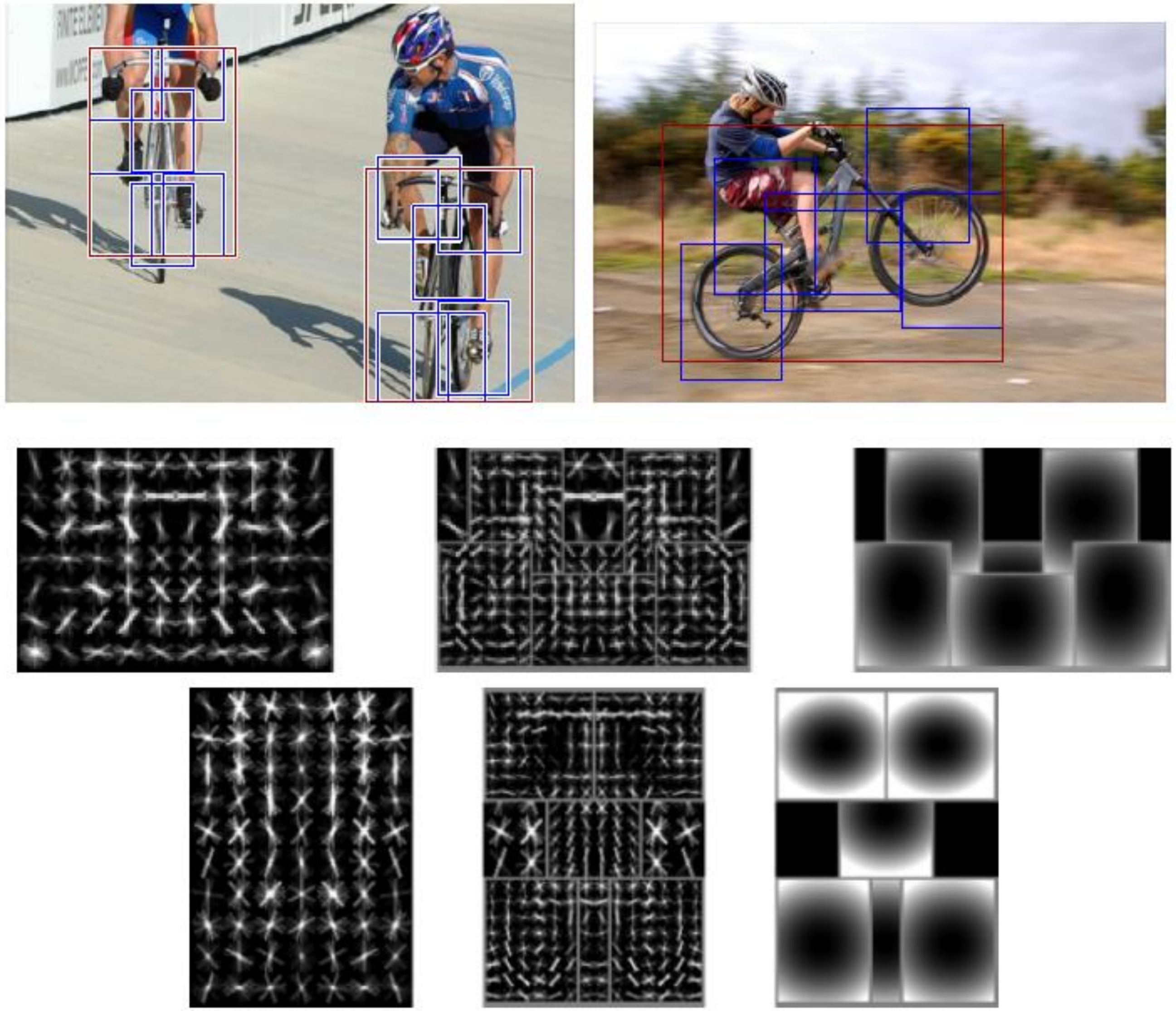


Figure from [Fischler & Elschlager 73]

Discriminatively trained part-based models



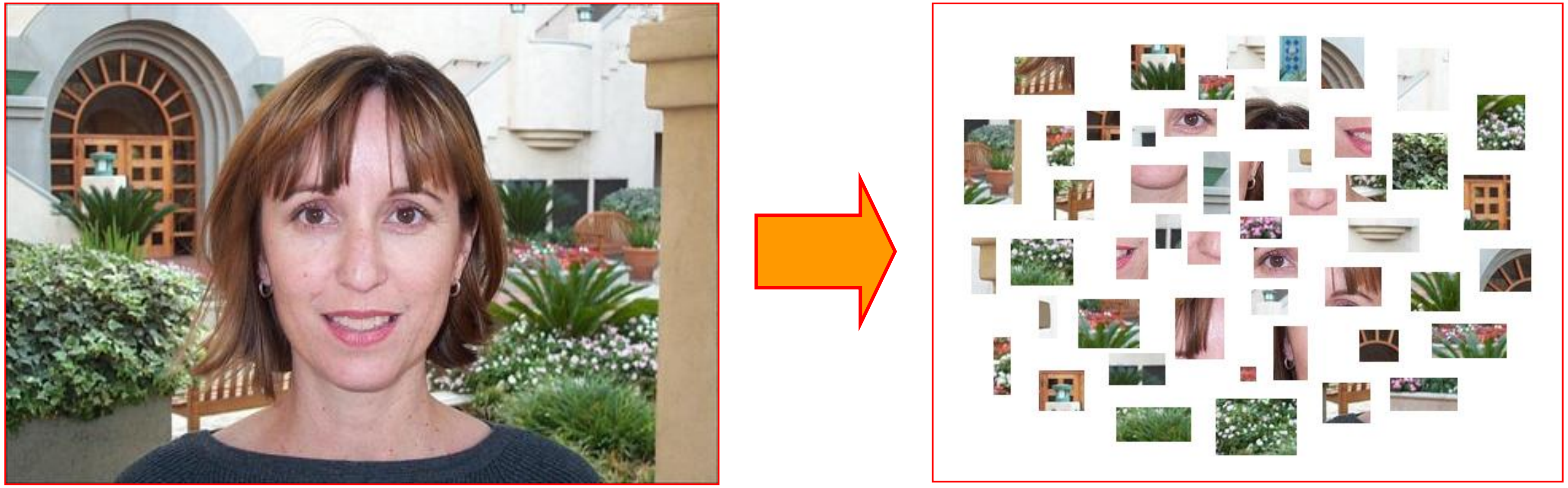
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, **"Object Detection with Discriminatively Trained Part-Based Models,"** PAMI 2009

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features

Svetlana Lazebnik

Bag-of-features models



Svetlana Lazebnik

Bag-of-features models

Object → **Bag of ‘words’**

Svetlana Lazebnik

Objects as texture

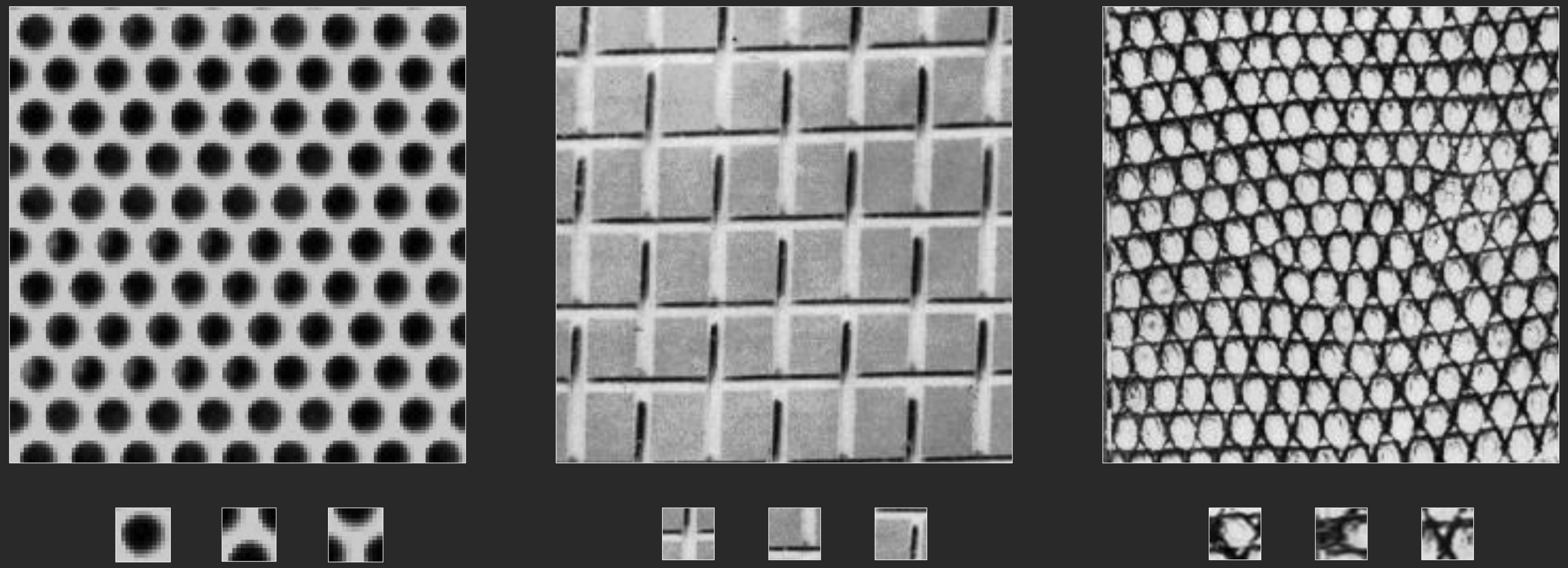
- All of these are treated as being the same

- No distinction between foreground and background. No concern about spatial layout.

Svetlana Lazebnik

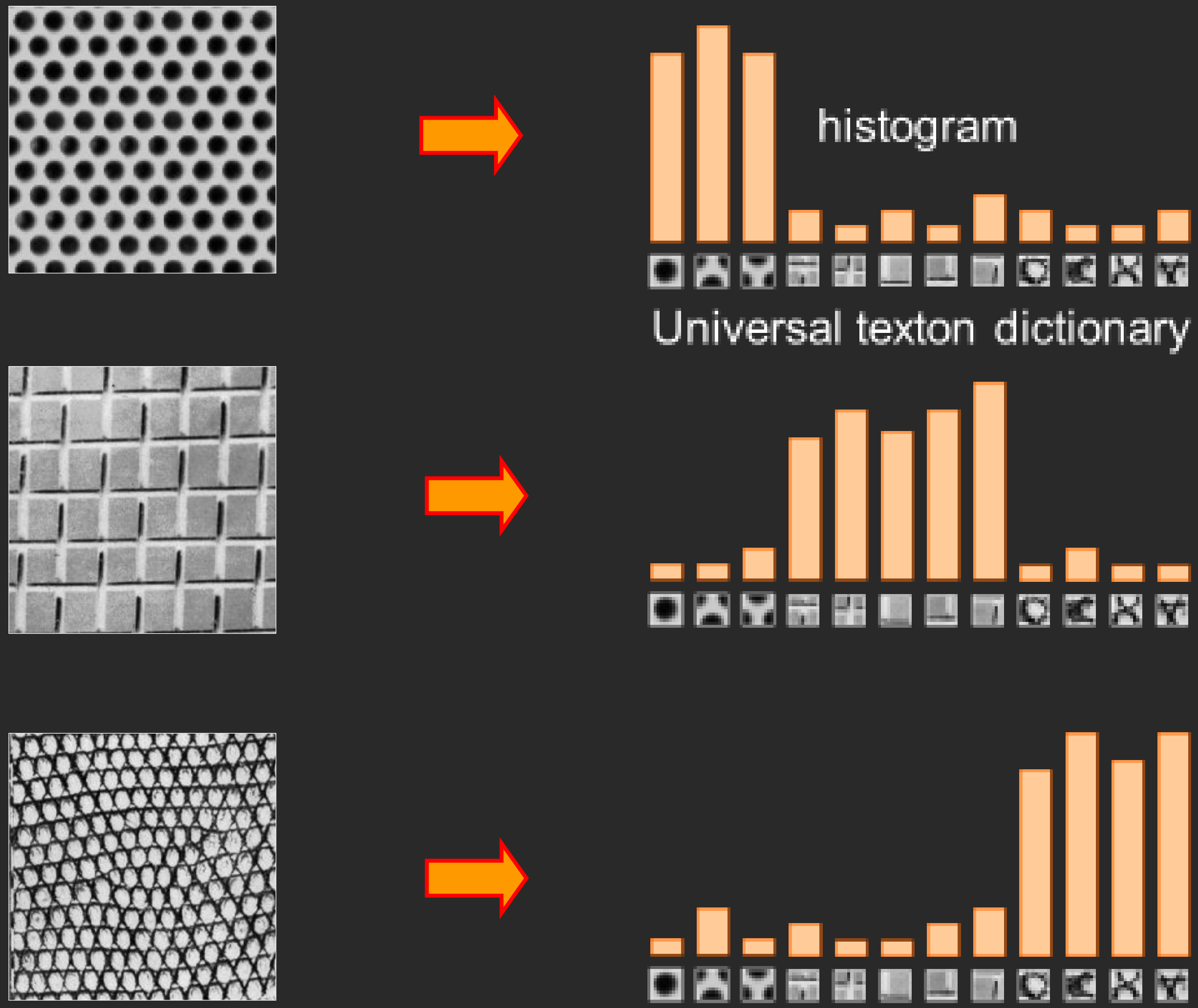
Origin 1: Texture recognition

- Texture is characterized by the repetition of basic elements or *textons*
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Origin 1: Texture recognition



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

US Presidential Speeches Tag Cloud
`http://chir.ag/phernalia/preztags/`

Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary   Salton & McGill (1983)

US Presidential Speeches Tag Cloud
`http://chir.ag/phernalia/preztags/`

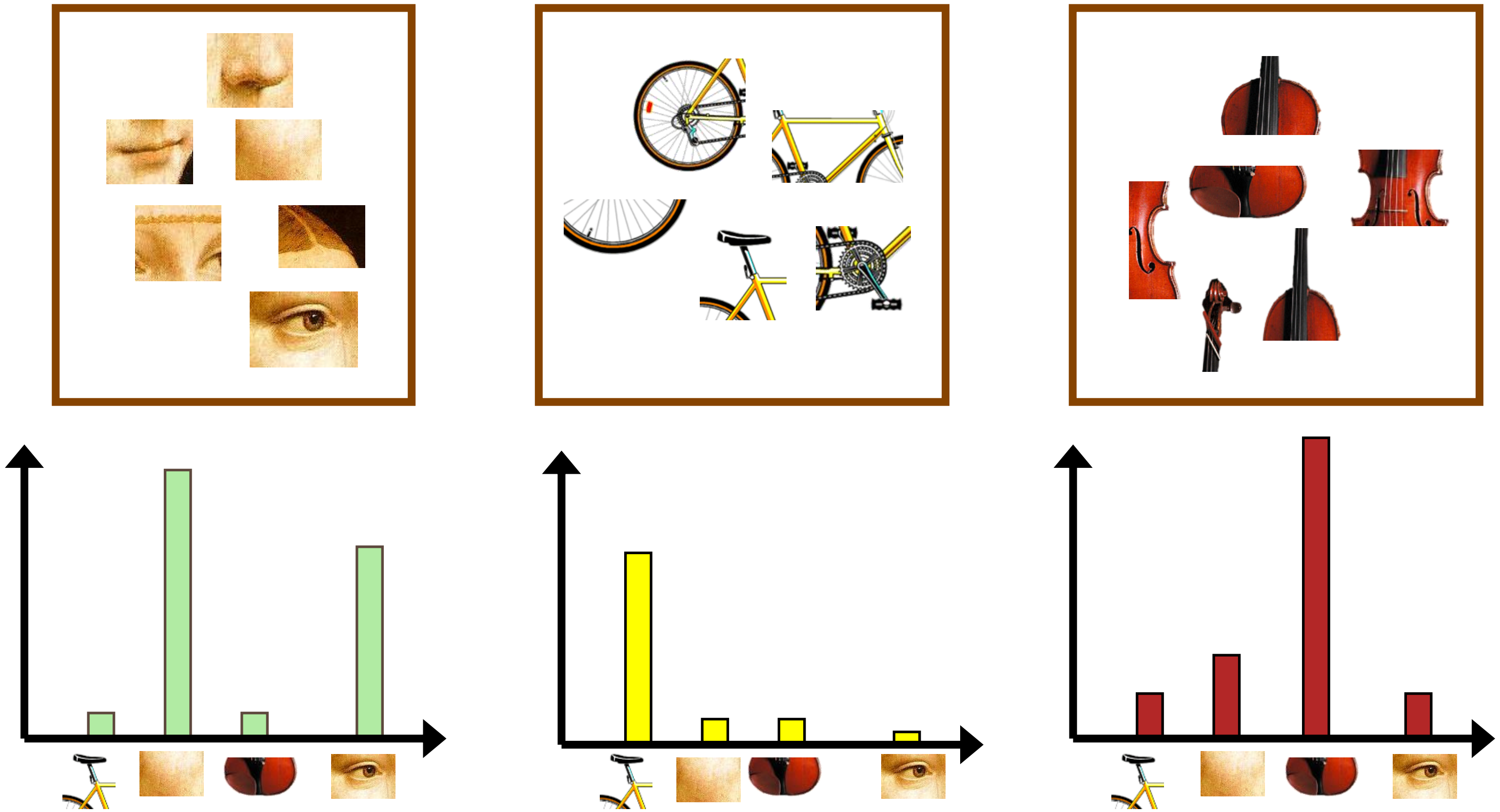
Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

US Presidential Speeches Tag Cloud
`http://chir.ag/phernalia/preztags/`

Bag-of-features steps

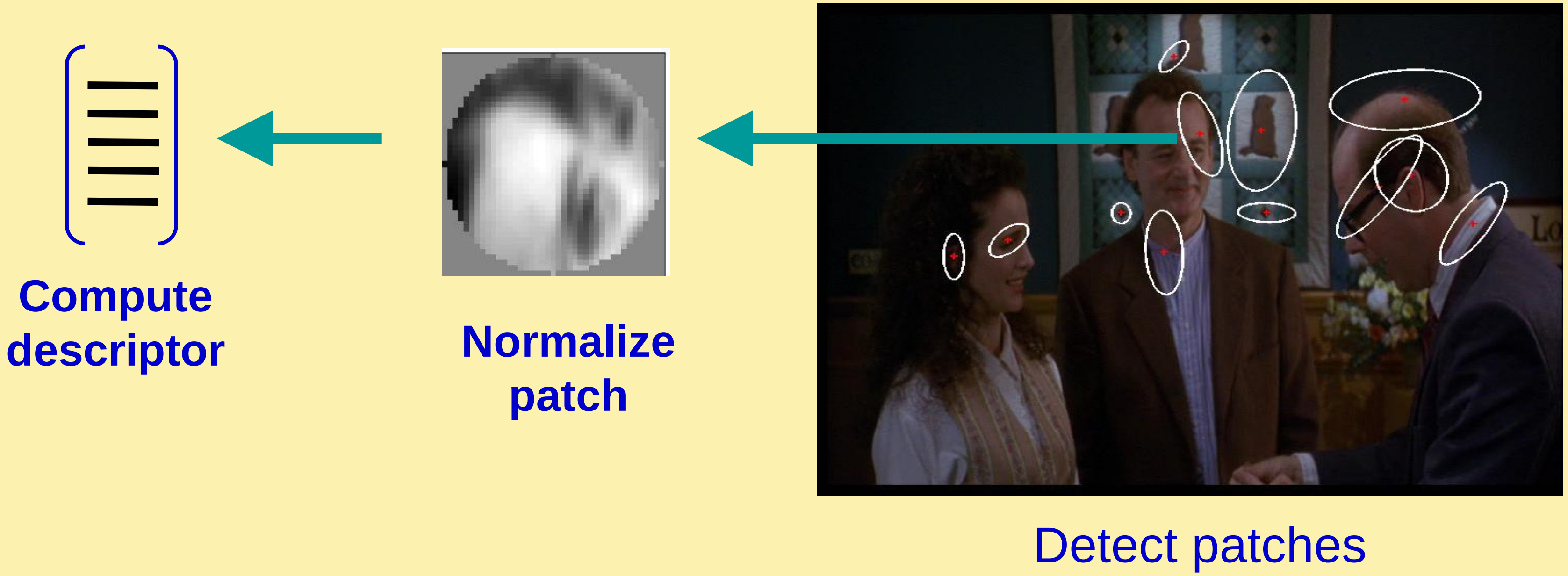
1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”



1. Feature extraction

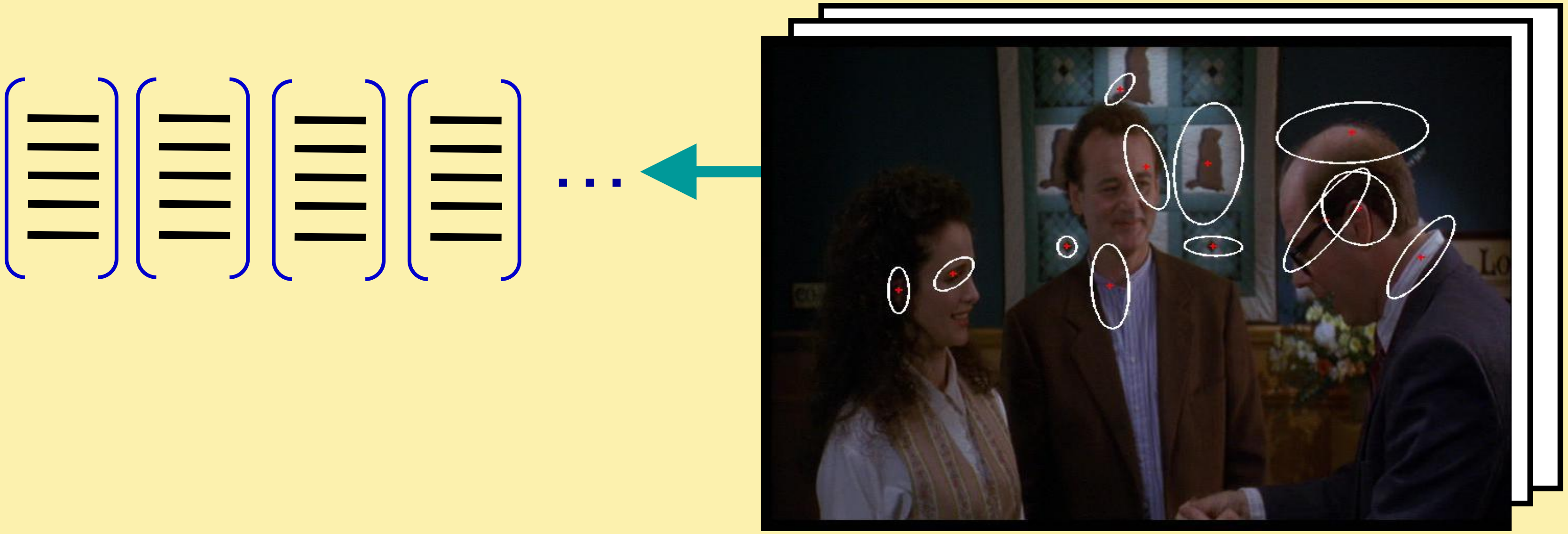
- Regular grid or interest regions

1. Feature extraction



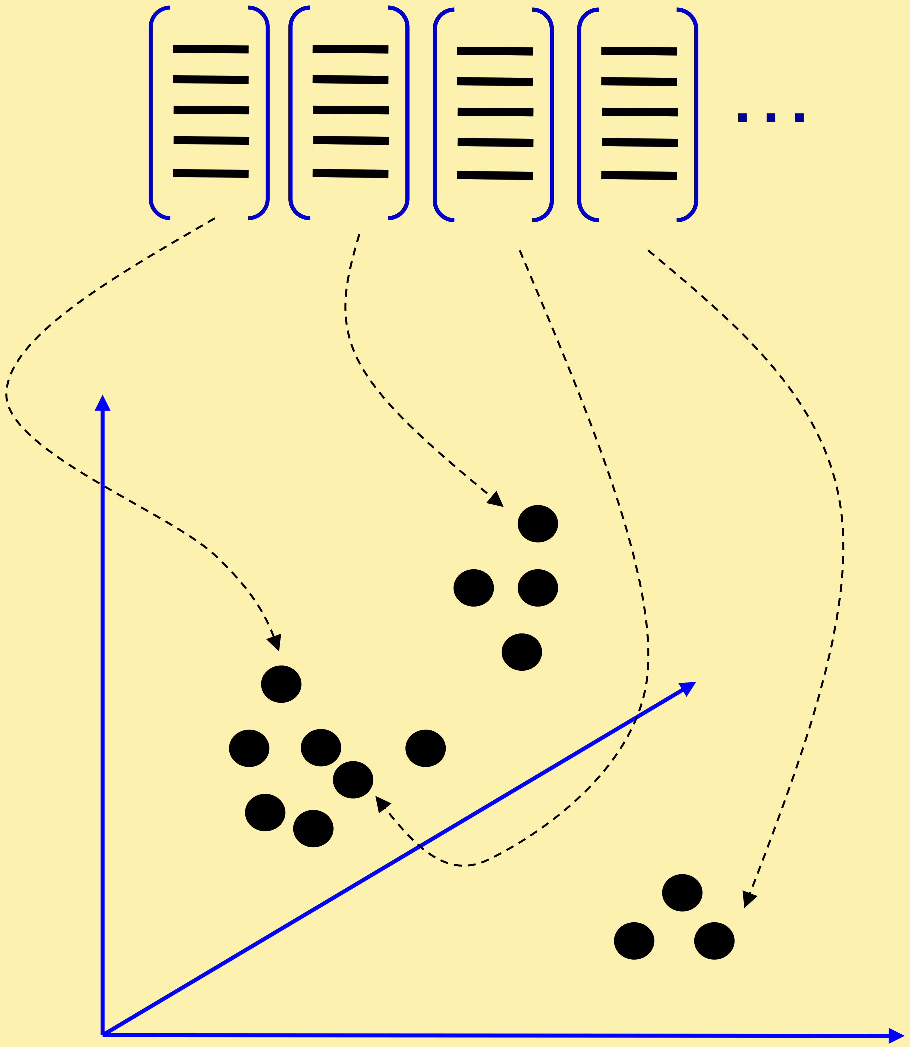
Slide credit: Josef Sivic

1. Feature extraction



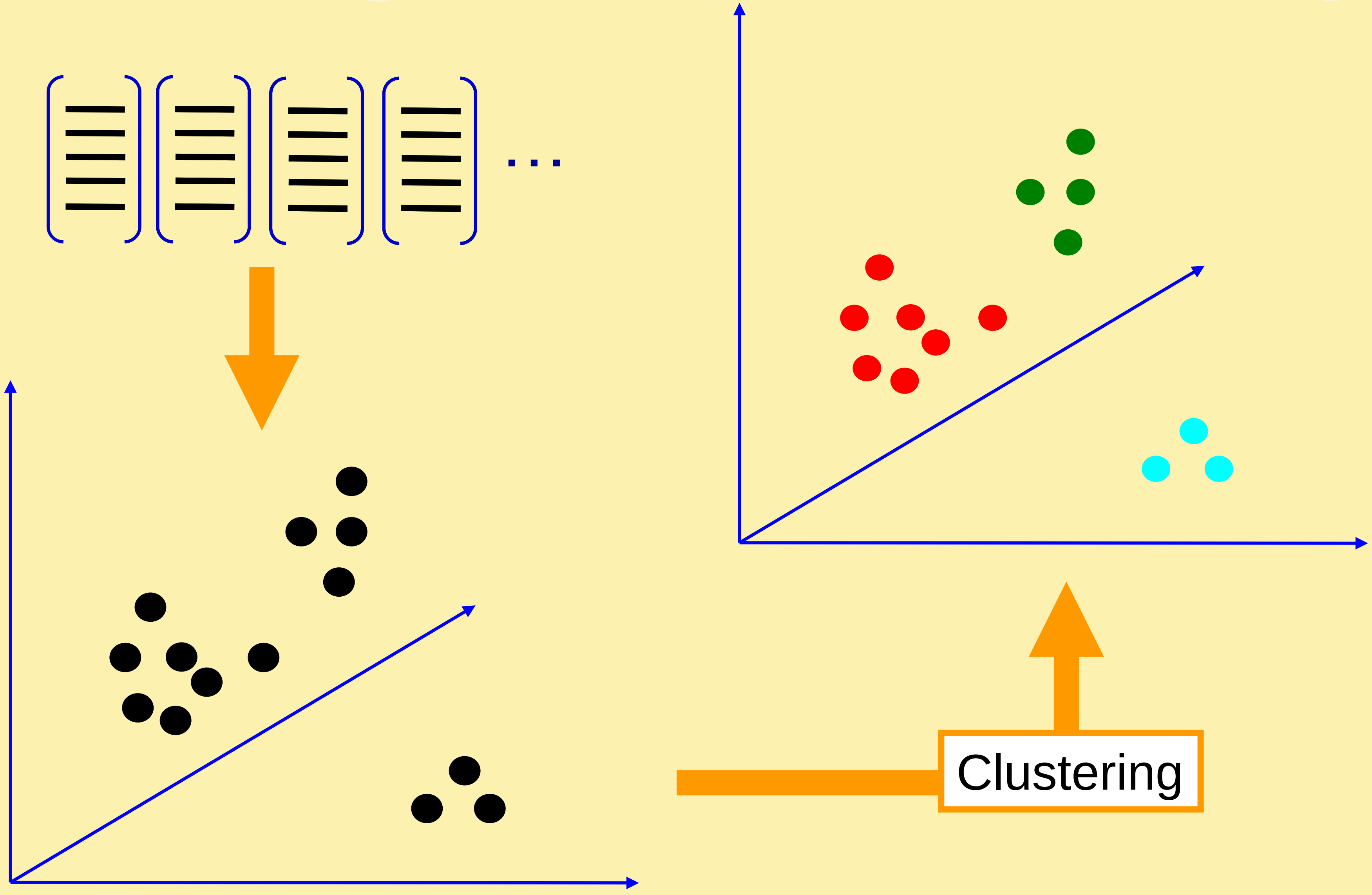
Slide credit: Josef Sivic

2. Learning the visual vocabulary



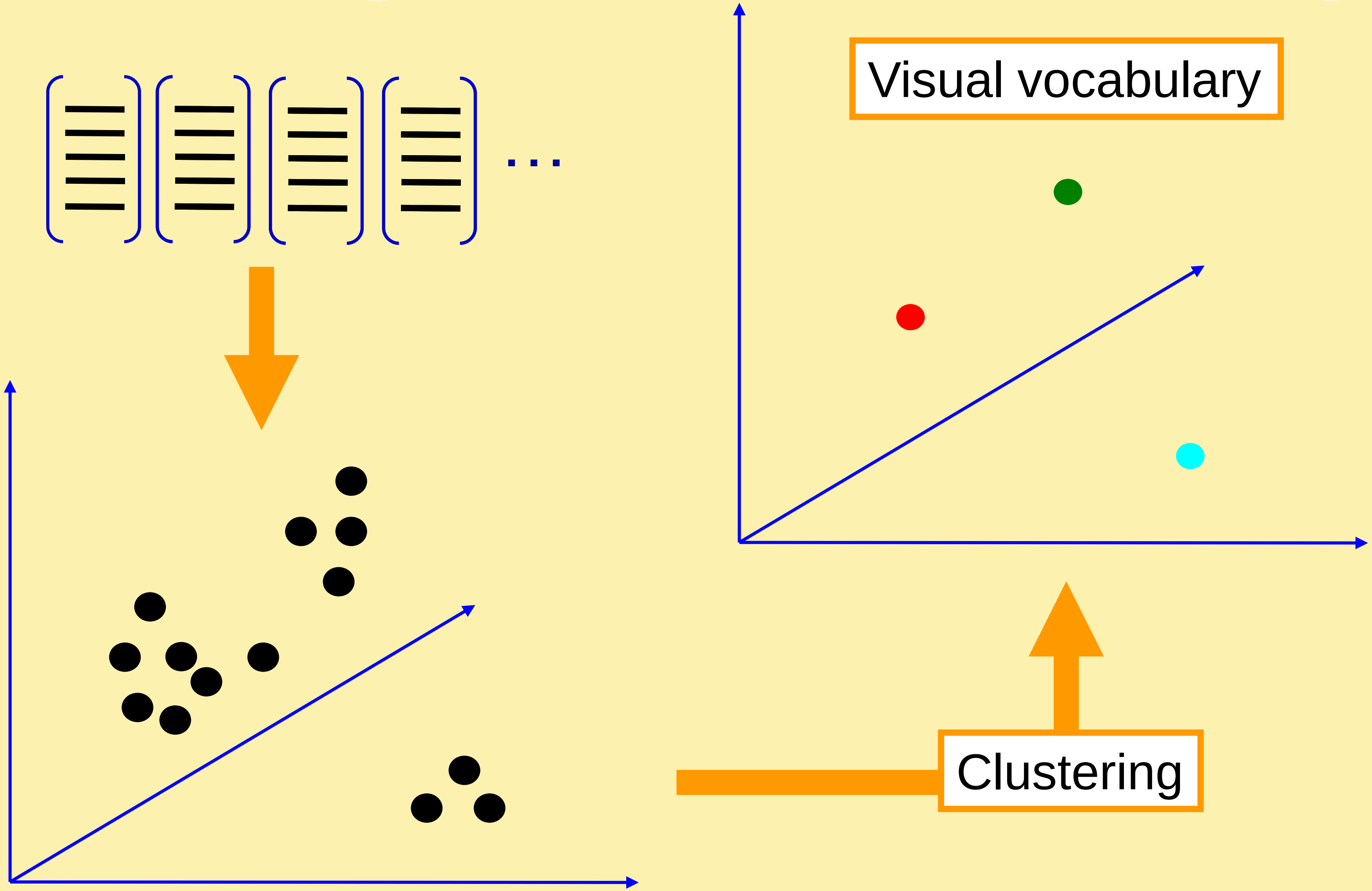
Slide credit: Josef Sivic

2. Learning the visual vocabulary



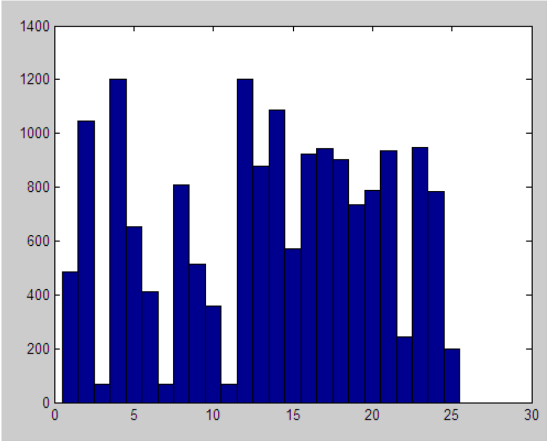
Slide credit: Josef Sivic

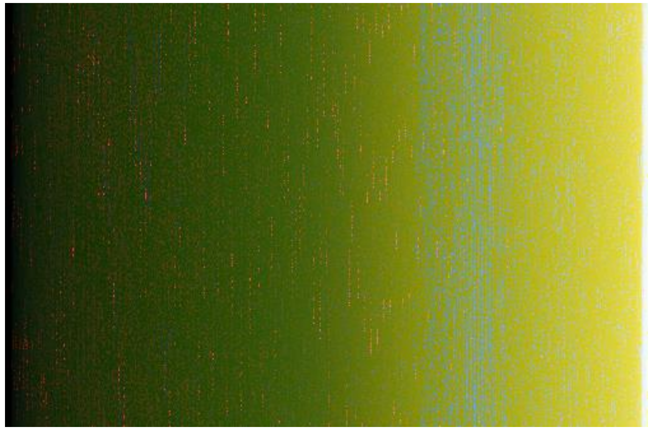
2. Learning the visual vocabulary



Slide credit: Josef Sivic

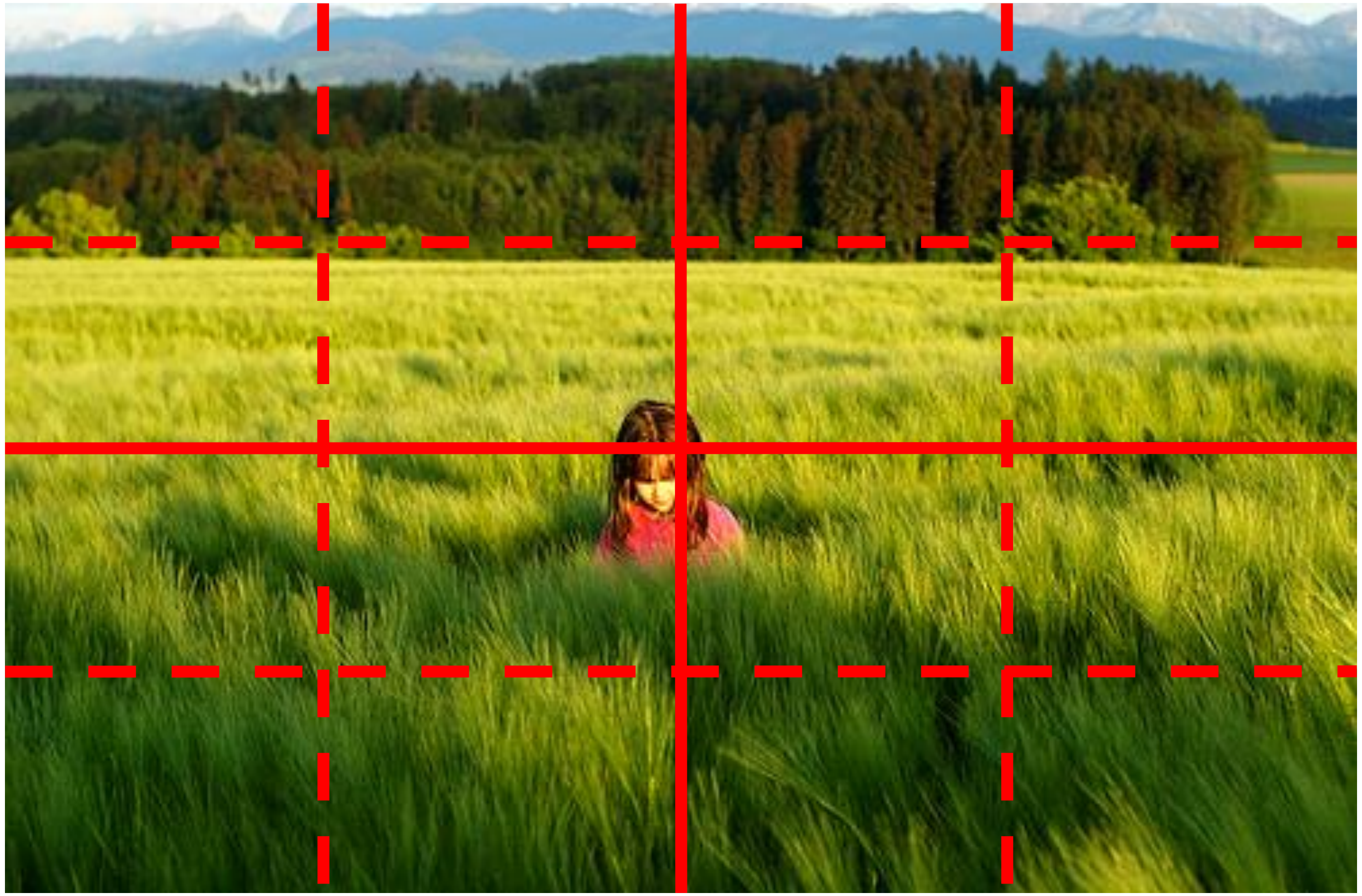
But what about layout?





All of these images have the same color histogram

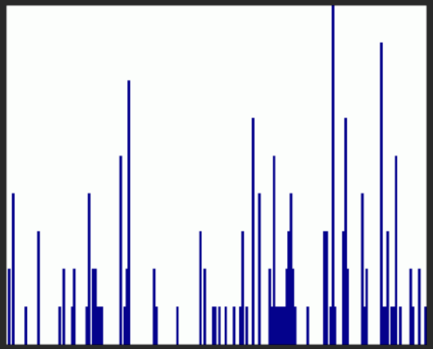
Spatial pyramid



Compute histogram in each spatial bin

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

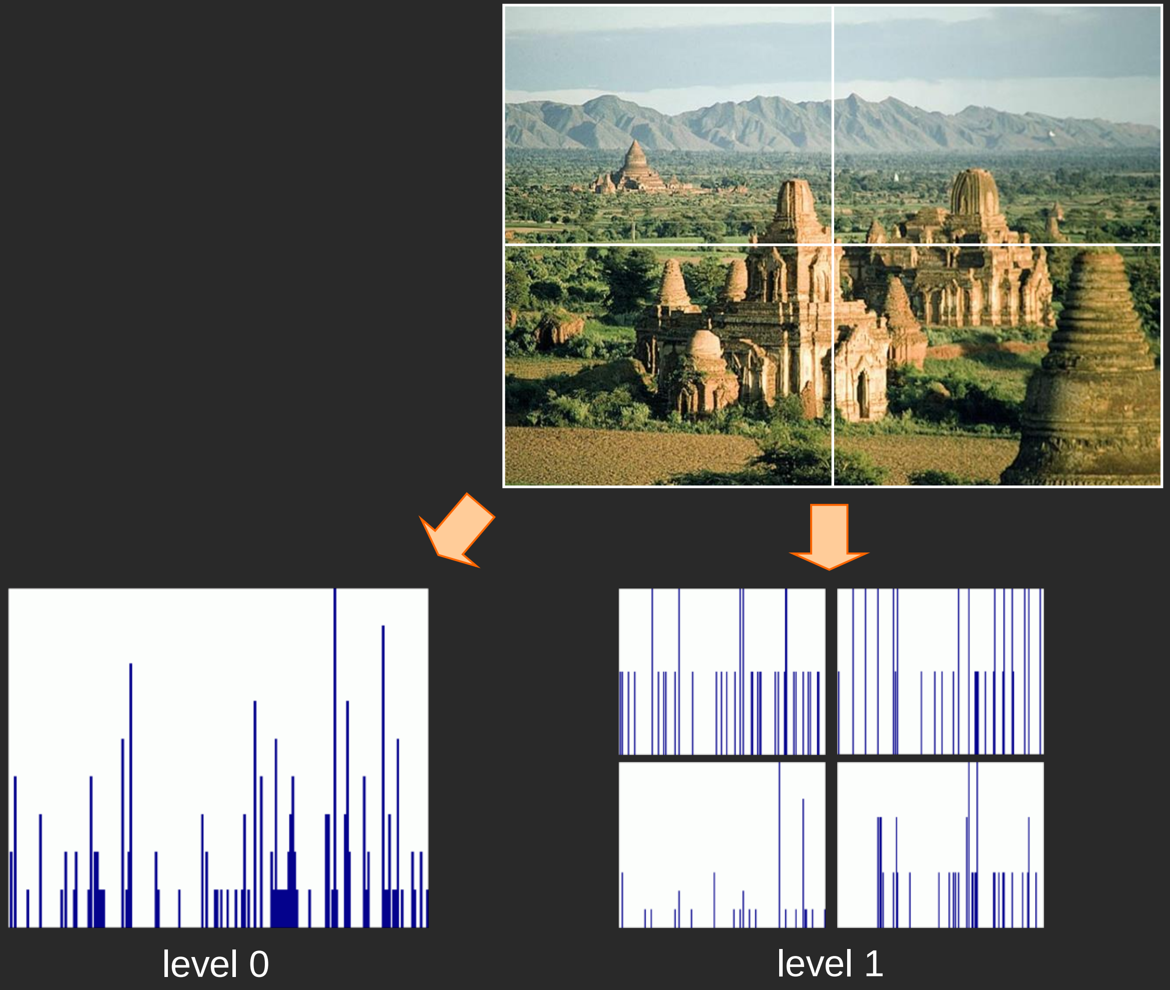


level 0

Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

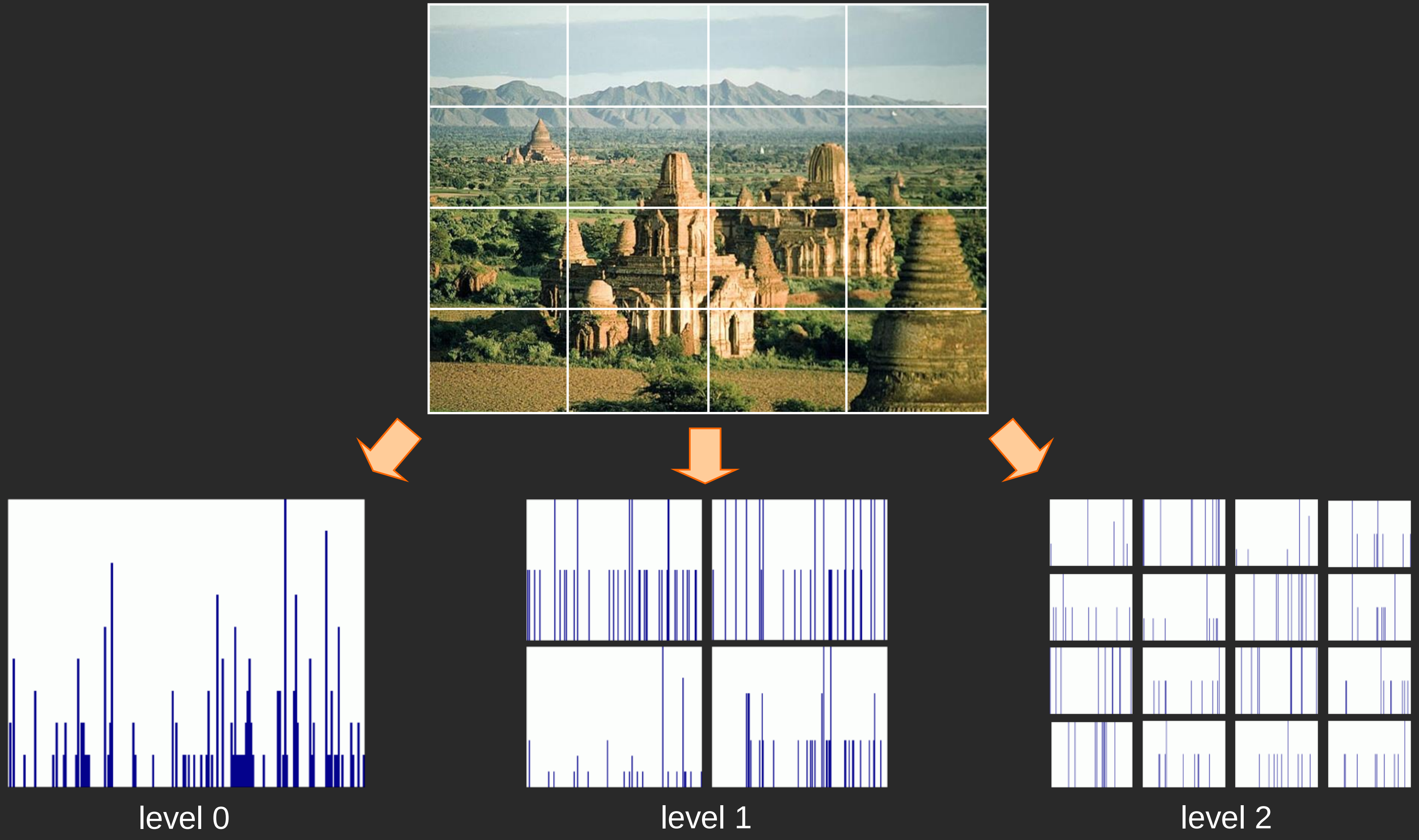
- Extension of a bag of features
- Locally orderless representation at several levels of resolution



Lazebnik, Schmid & Ponce (CVPR 2006)

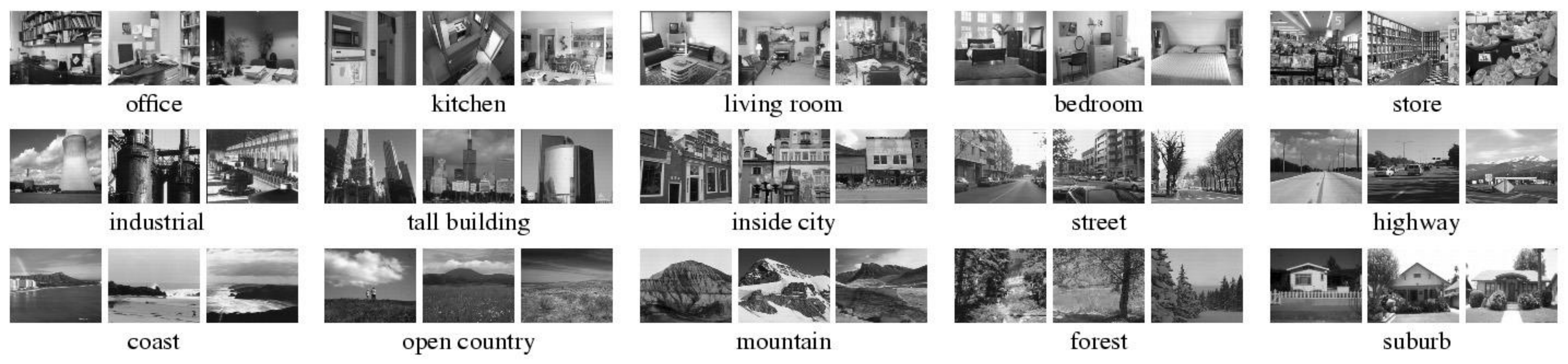
Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution



Lazebnik, Schmid & Ponce (CVPR 2006)

Scene category dataset



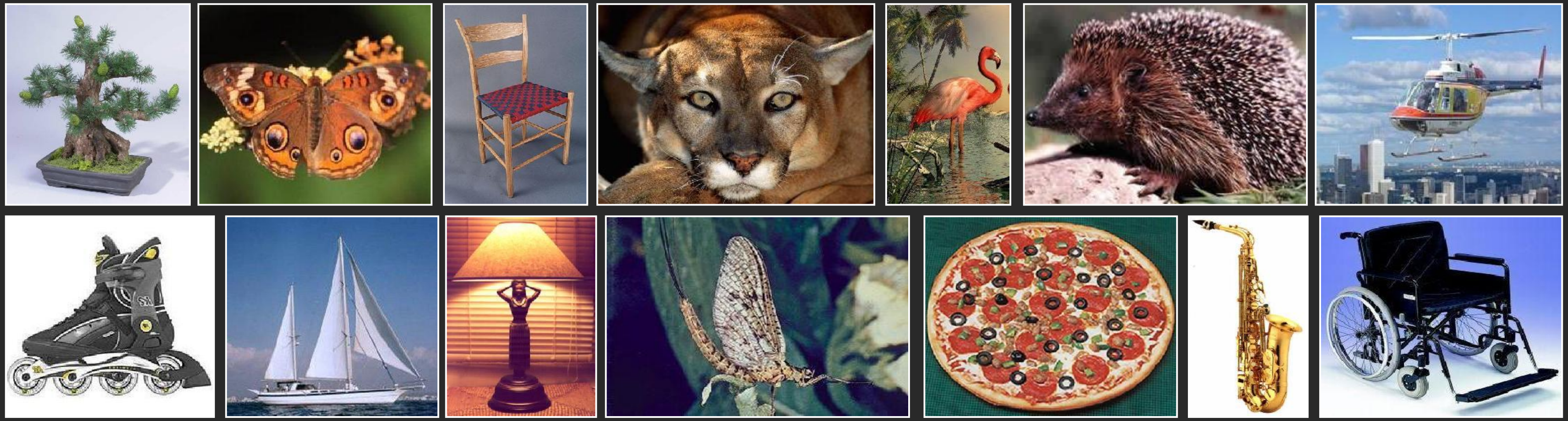
Multi-class classification results

(100 training images per class)

	Strong features (vocabulary size: 200)	
Level	Single-level	Pyramid
0 (1×1)	72.2 ± 0.6	
1 (2×2)	77.9 ± 0.6	79.0 ± 0.5
2 (4×4)	79.4 ± 0.3	**81.1** ± 0.3
3 (8×8)	77.2 ± 0.4	80.7 ± 0.3

Caltech101 dataset

http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html



Multi-class classification results (30 training images per class)

	Strong features (200)	
Level	Single-level	Pyramid
0	41.2 ±1.2	
1	55.9 ±0.9	57.0 ±0.8
2	63.6 ±0.9	**64.6** ±0.8
3	60.3 ±0.9	64.6 ±0.7

Modern recognition systems have < 3% error on CalTech 101

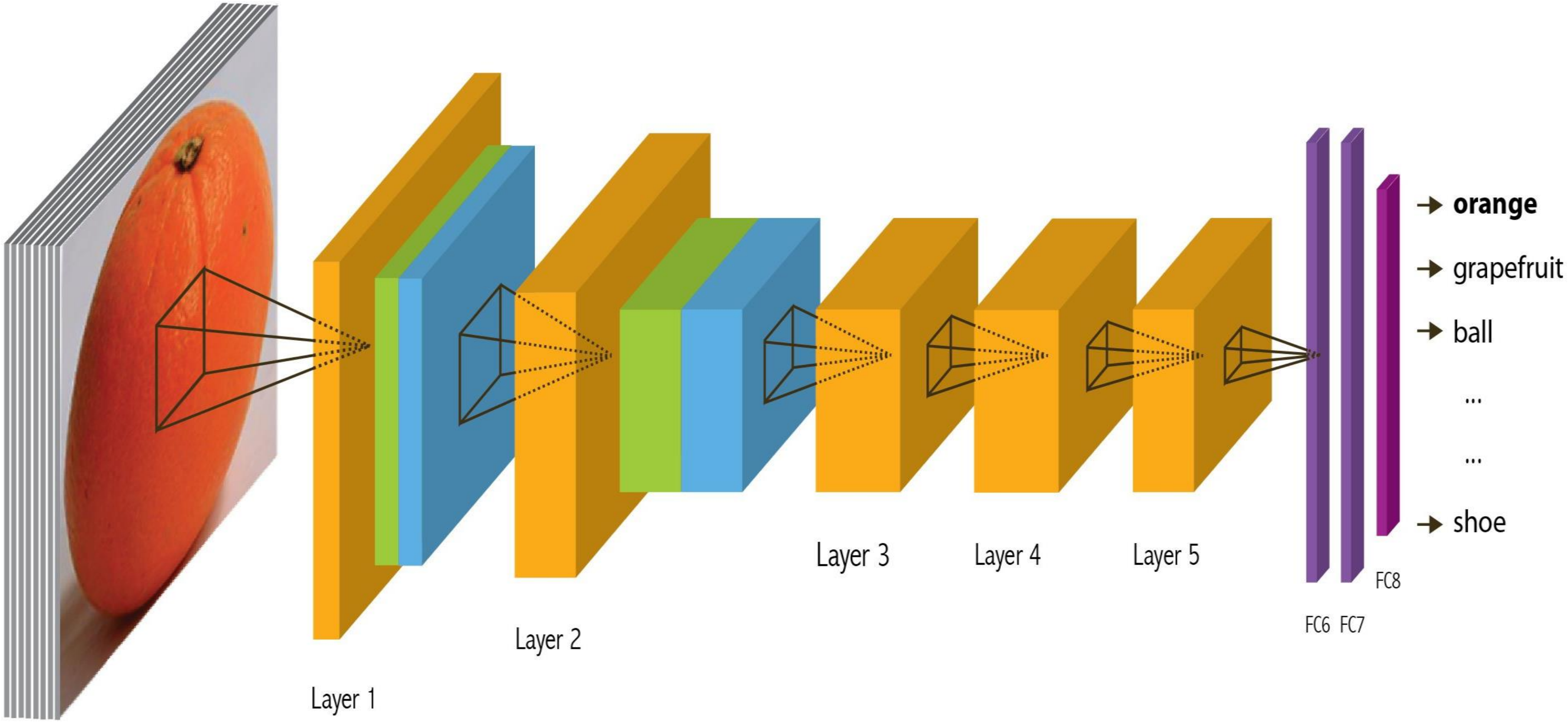
History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features
- Present trends: *deep learning*

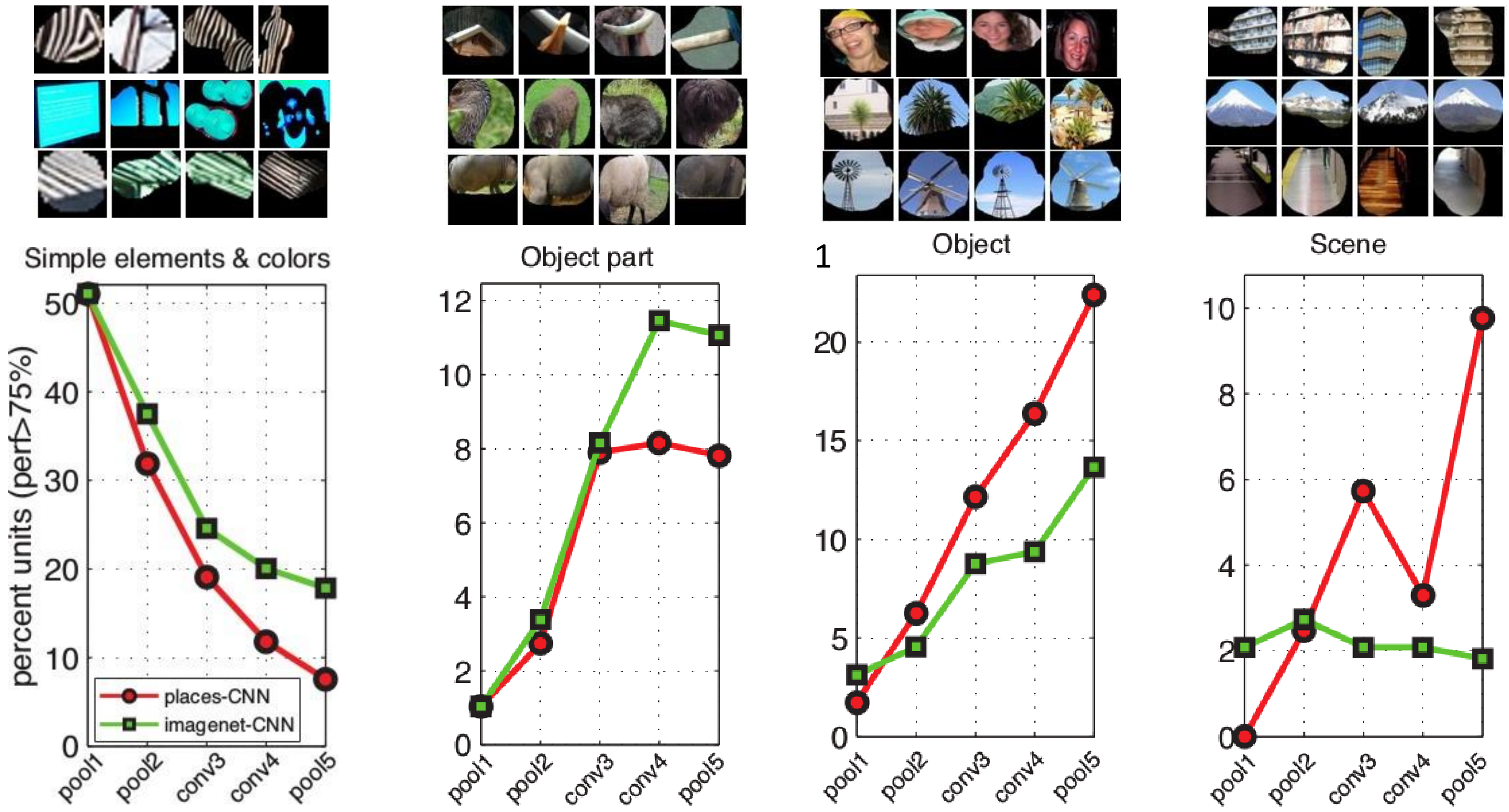
Svetlana Lazebnik

Beyond AlexNet

Recap: Convolutional Network, AlexNet



Recap: Convolutional Network Interpretation



Object detectors emerge within CNN trained to classify scenes, without any object supervision!

Beyond AlexNet

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan & Andrew Zisserman 2015

These are the "VGG" networks.
"Perceptual Loss" in generative deep learning refers to these networks

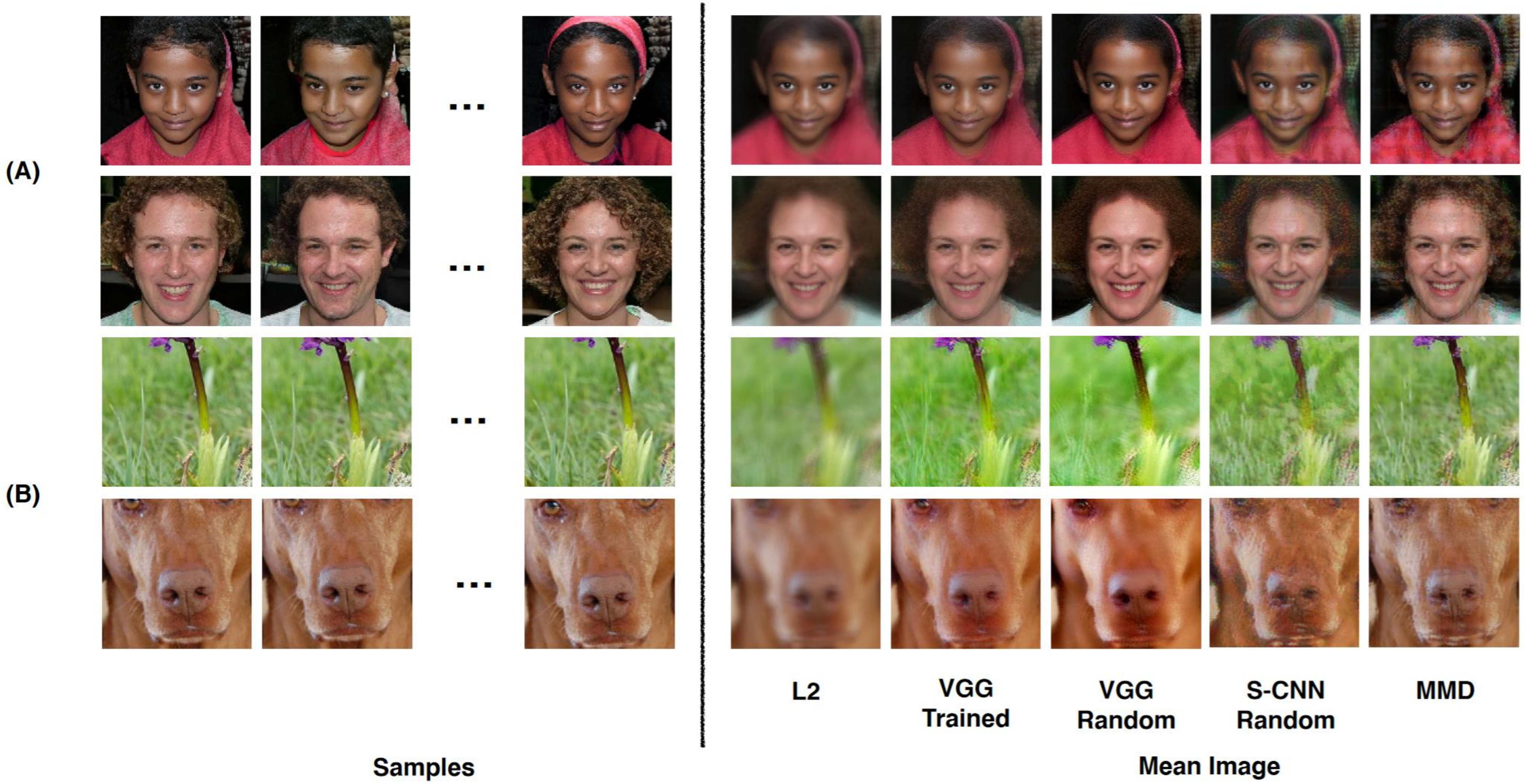
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Table 4: **ConvNet performance at multiple test scales.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	$[256; 512]$	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	$[256; 512]$	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	$[256; 512]$	256,384,512	**24.8**	**7.5**



“VGG” networks are commonly used as the basis for “Perceptual Loss”.
The images on the right are as close as possible to all images on the left in various feature spaces.

Understanding and Simplifying Perceptual Distances. Dan Amir and Yair Weiss. CVPR 2021

Generative Image Dynamics

Zhengqi Li, Richard Tucker, Noah Snavely, Aleksander Holynski
Google Research

https://generative-dynamics.github.io/

CVPR 2024 Best Paper Award

Paper arXiv Demo Supp

We jointly train the feature extractor and synthesis networks with start and target frames (I_0, I_t) randomly sampled from real videos, using the estimated flow field from I_0 to I_t to warp encoded features from I_0, and supervising predictions $\hat{I}_t$ against I_t with a VGG perceptual loss [49].

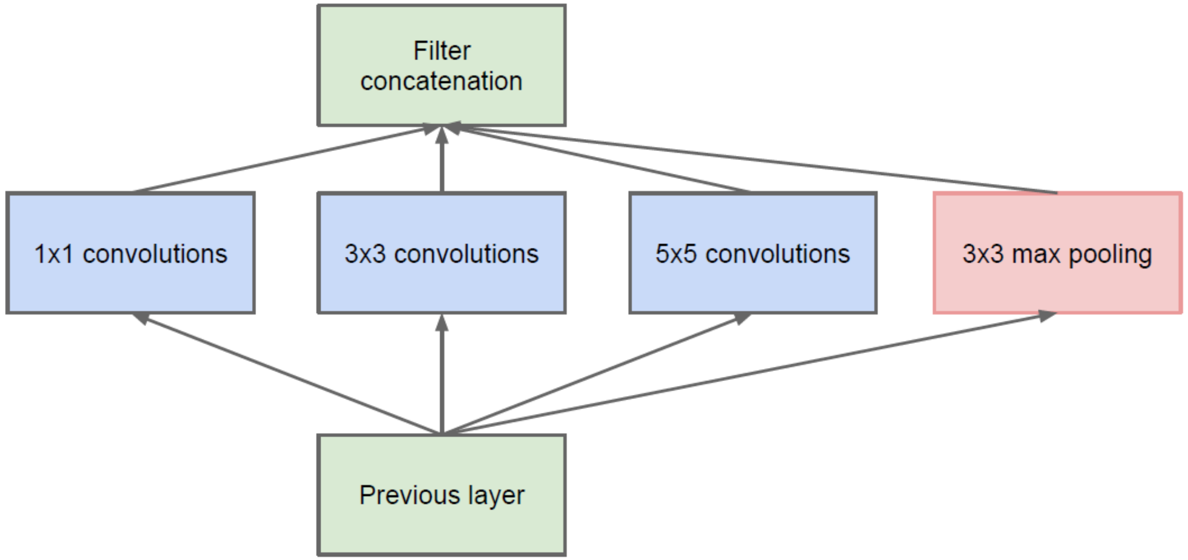
Our method automatically turns single still images into seamless looping videos.

Going Deeper with Convolutions

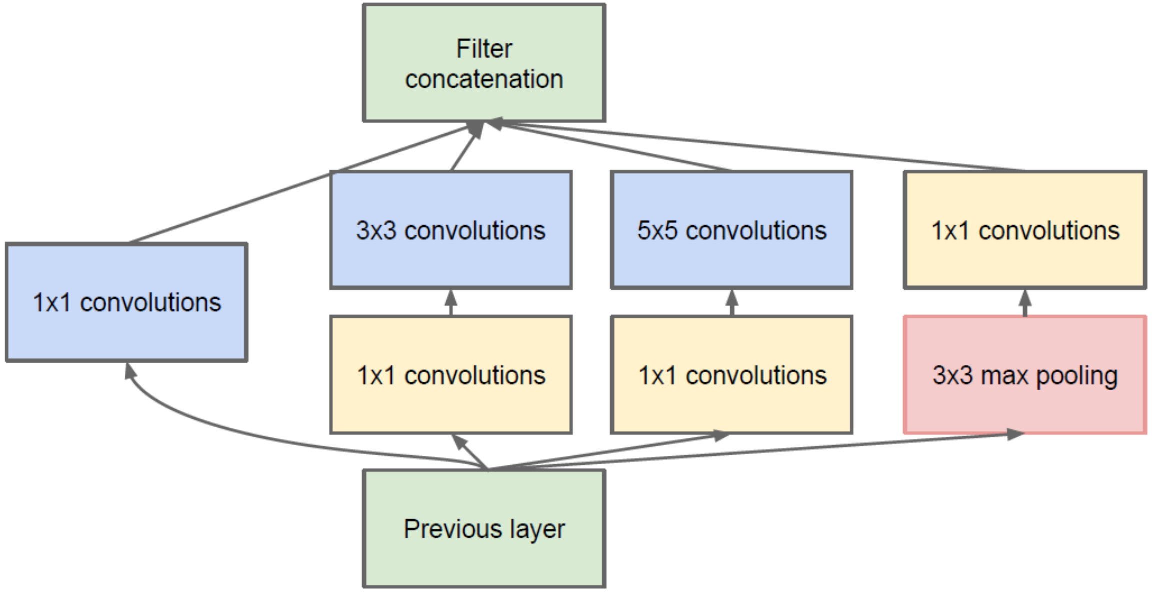
**Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed,
Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich**
2015

This is the "Inception" architecture or "GoogLeNet"

***The architecture blocks are called "Inception" modules
and the collection of them into a particular net is "GoogLeNet"**



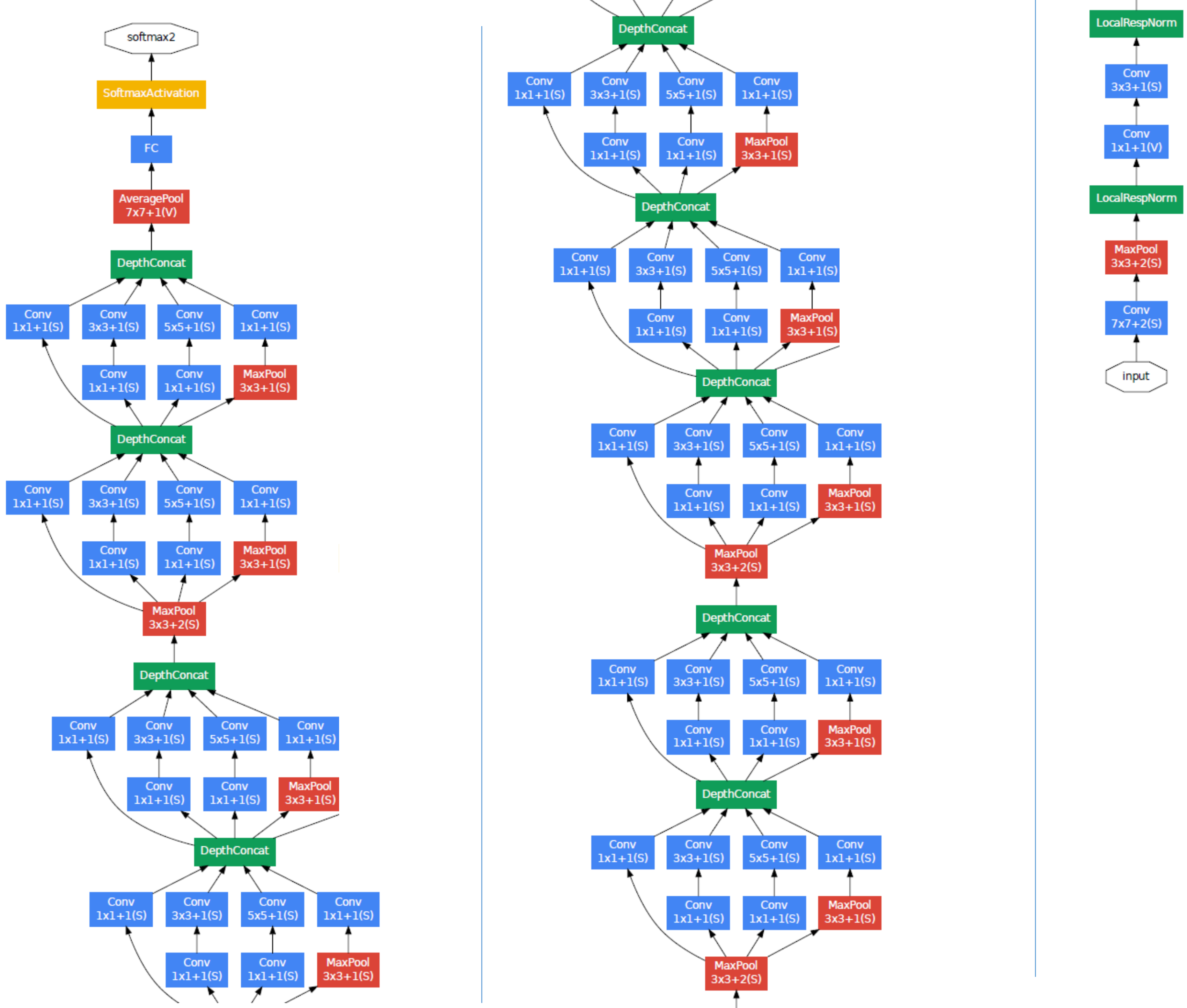
(a) Inception module, naïve version

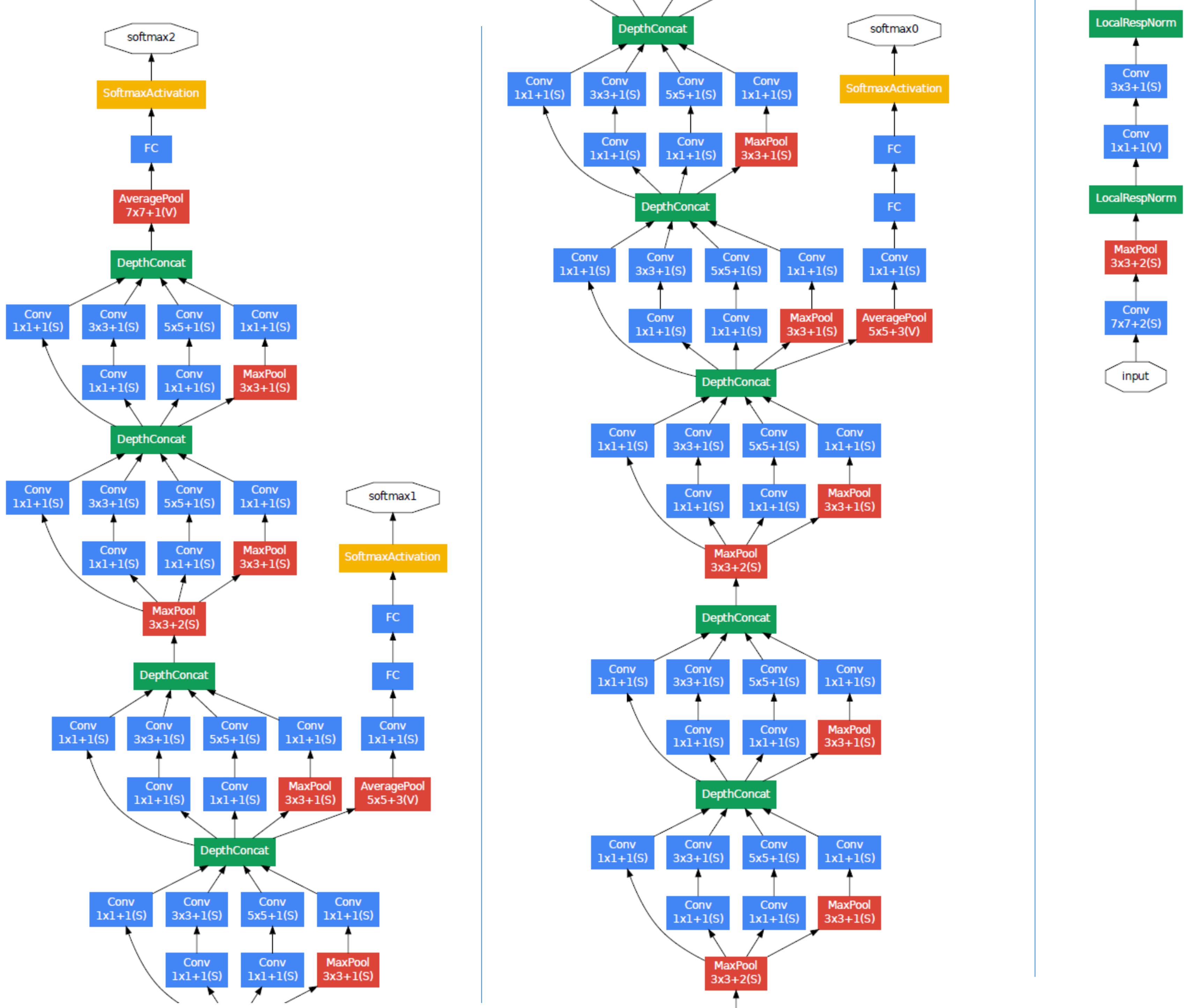


(b) Inception module with dimensionality reduction

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Only 6.8 million parameters. AlexNet ~60 million, VGG up to 138 million


softmax2
SoftmaxActivation
FC
AveragePool 7x7+1(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
LocalRespNorm
Conv 3x3+1(S)
Conv 1x1+1(V)
LocalRespNorm
MaxPool 3x3+2(S)
Conv 7x7+2(S)
input

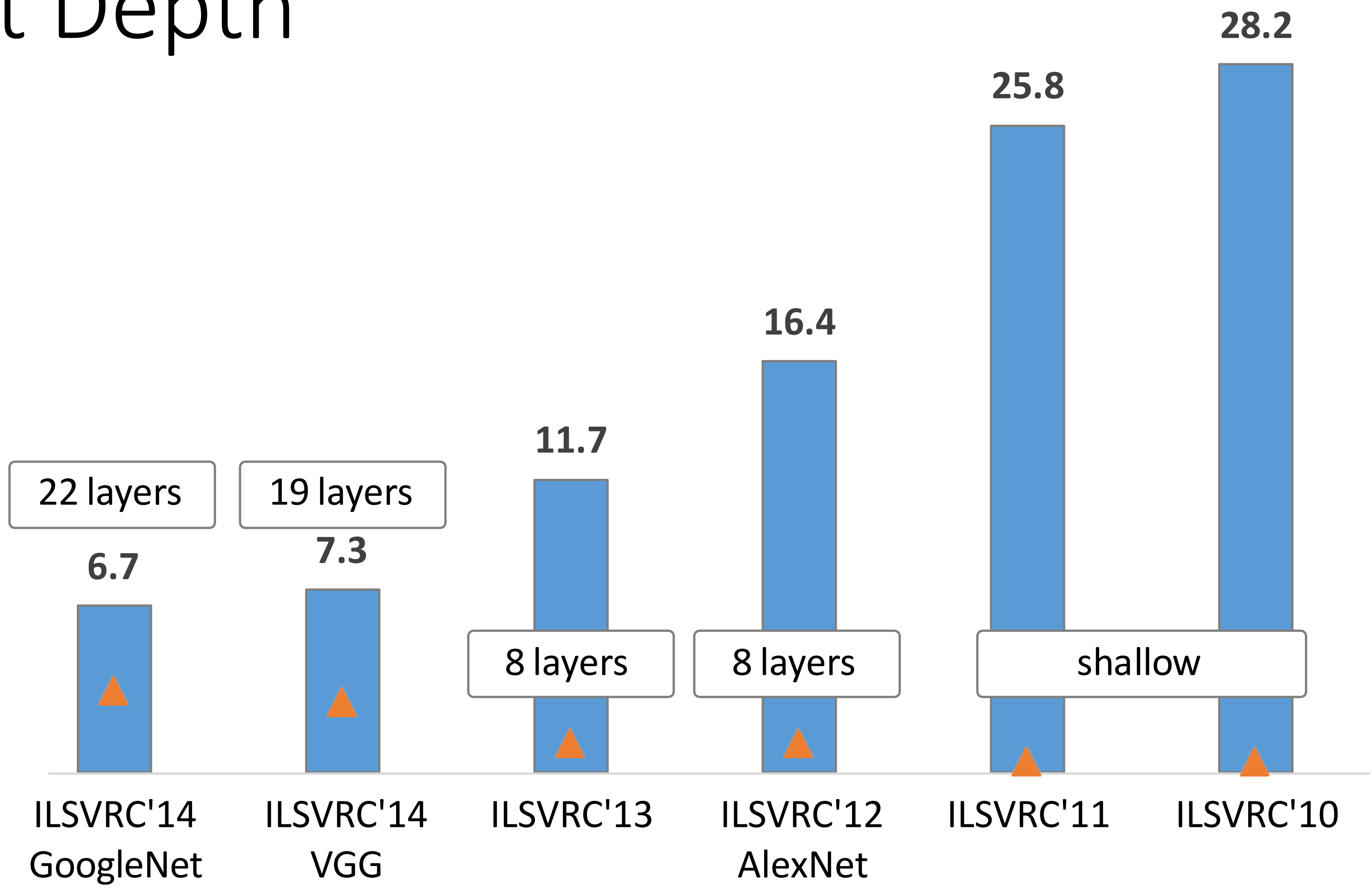

softmax2
SoftmaxActivation
FC
AveragePool 7x7+1(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
softmax1
SoftmaxActivation
FC
FC
Conv 1x1+1(S)
AveragePool 5x5+3(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
softmax0
SoftmaxActivation
FC
FC
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
Conv 1x1+1(S)
AveragePool 5x5+3(V)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
LocalRespNorm
Conv 3x3+1(S)
Conv 1x1+1(V)
LocalRespNorm
MaxPool 3x3+2(S)
Conv 7x7+2(S)
input

Team	**Year**	**Place**	**Error (top-5)**	**Uses external data**
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

Table 2: Classification performance.

Number of models	**Number of Crops**	**Cost**	**Top-5 error**	**compared to base**
1	1	1	10.07%	base
1	10	10	9.15%	-0.92%
1	144	144	7.89%	-2.18%
7	1	7	8.09%	-1.98%
7	10	70	7.62%	-2.45%
7	144	1008	6.67%	-3.45%

ConvNet Depth



ImageNet Classification top-5 error (%)

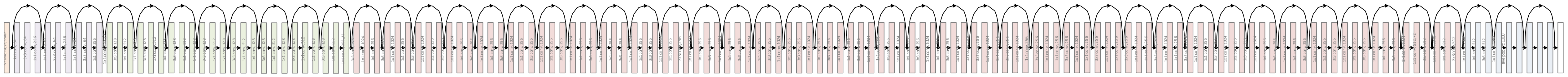
Surely it would be ridiculous to go any deeper…

Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

work done at

Microsoft Research Asia



Cited 255,503 times as of 3/5/2025

	Publication	h5-index	h5-median
1.	Nature	444	667
2.	The New England Journal of Medicine	432	780
3.	Science	401	614
4.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	389	627
5.	The Lancet	354	635
6.	Advanced Materials	312	418
7.	Nature Communications	307	428
8.	Cell	300	505
9.	International Conference on Learning Representations	286	533
10.	Neural Information Processing Systems	278	436
11.	JAMA	267	425
12.	Chemical Reviews	265	444
13.	Proceedings of the National Academy of Sciences	256	364
14.	Angewandte Chemie	245	332
15.	Chemical Society Reviews	244	386
16.	Journal of the American Chemical Society	242	344
17.	IEEE/CVF International Conference on Computer Vision	239	415
18.	Nucleic Acids Research	238	550
19.	International Conference on Machine Learning	237	421

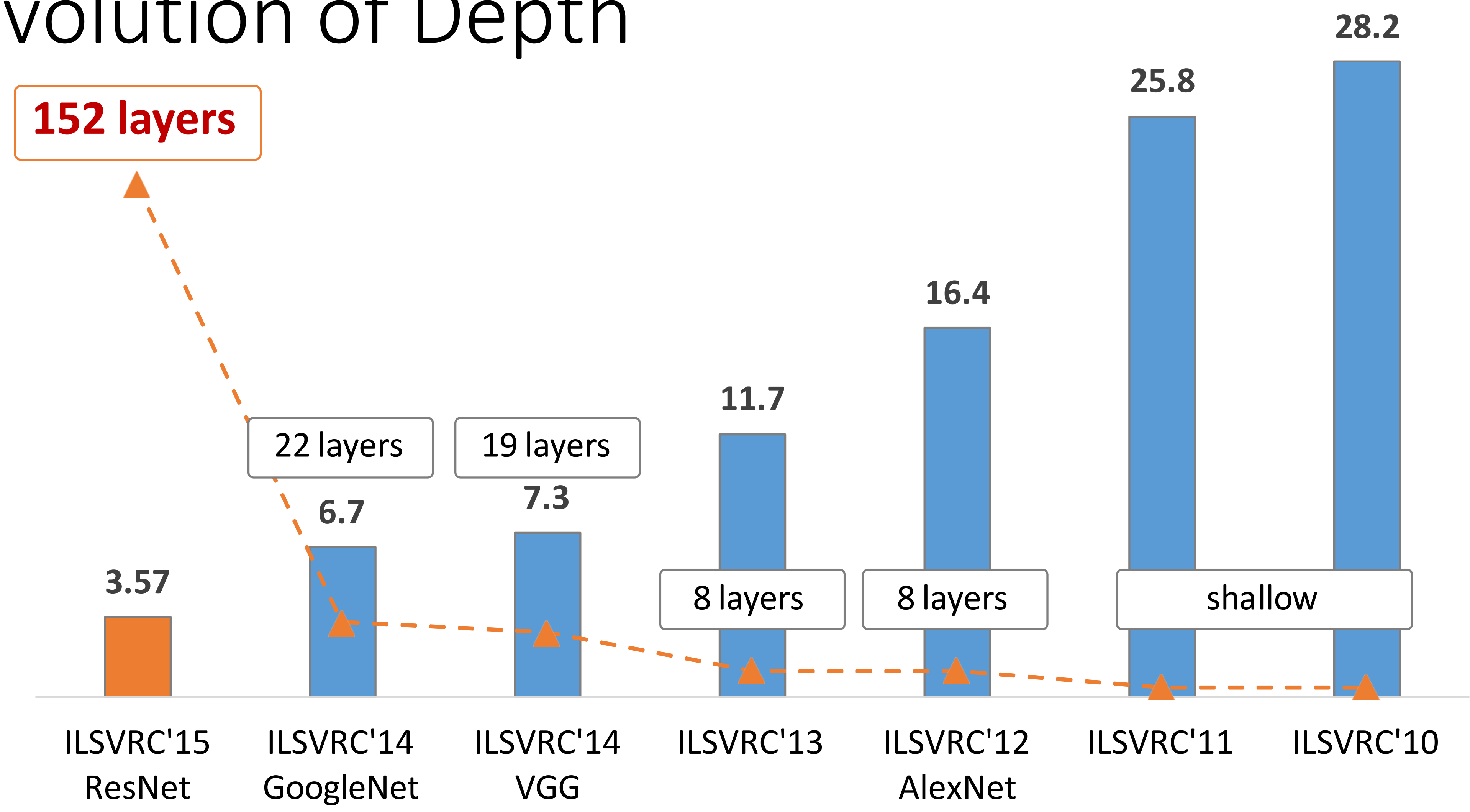
ResNet @ ILSVRC & COCO 2015 Competitions

1st places in all five main tracks

- ImageNet Classification: “*Ultra-deep*” 152-layer nets
- ImageNet Detection: 16% better than 2nd
- ImageNet Localization: 27% better than 2nd
- COCO Detection: 11% better than 2nd
- COCO Segmentation: 12% better than 2nd

*improvements are relative numbers

Revolution of Depth

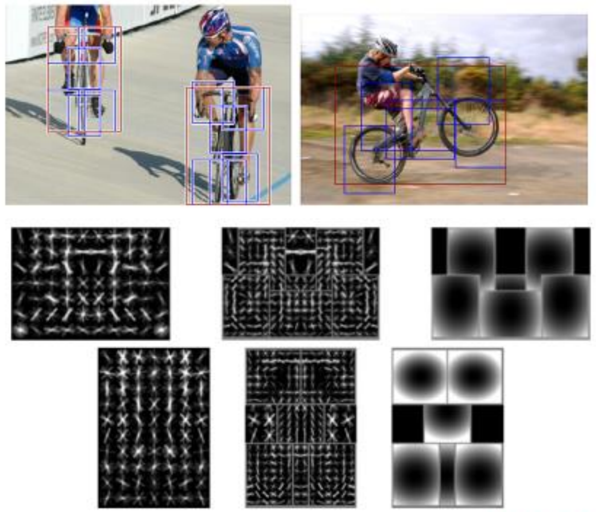


ImageNet Classification top-5 error (%)

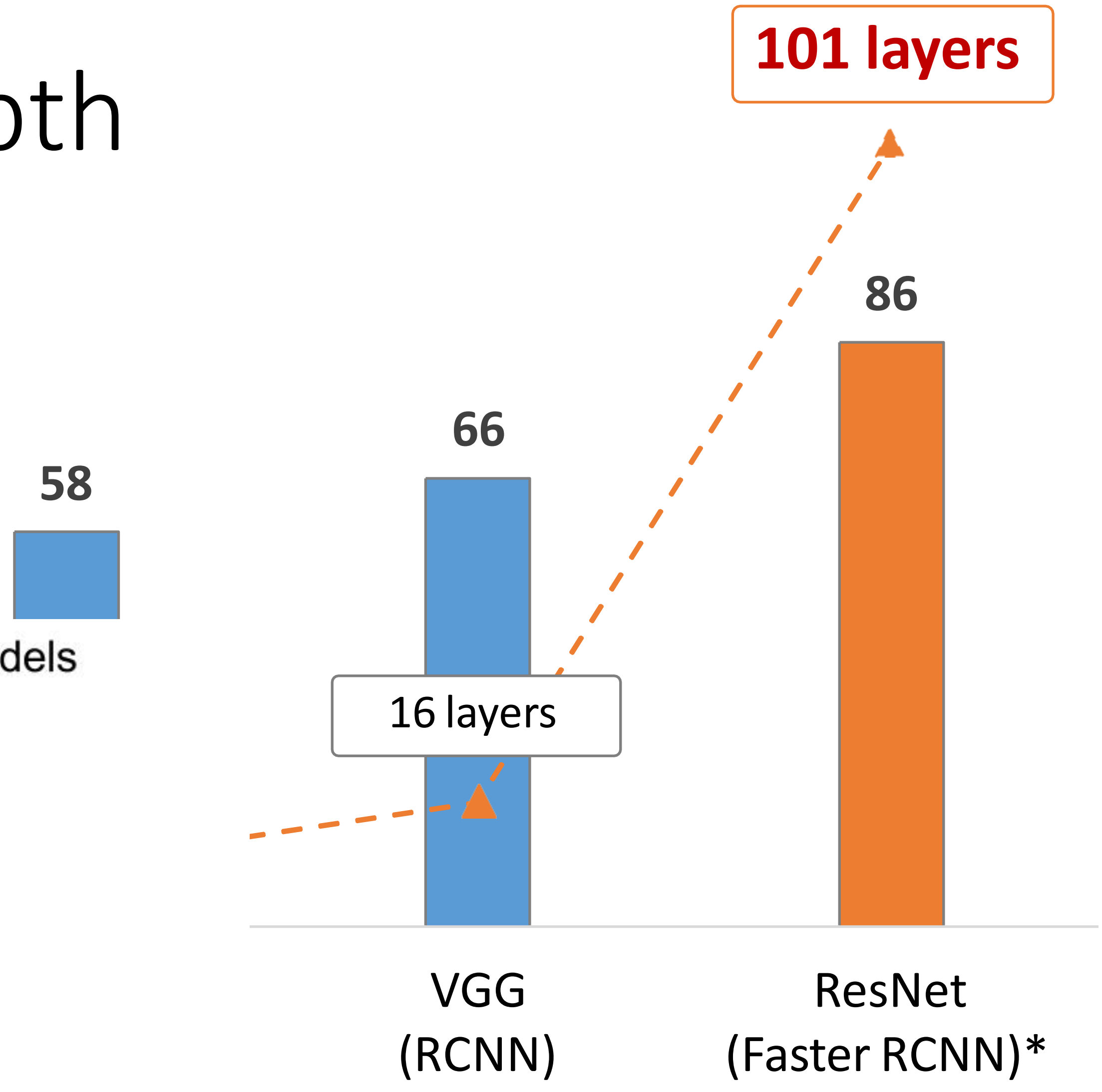
Revolution of Depth

Engines of
visual recognition

Discriminatively trained part-based models



P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," PAMI 2009

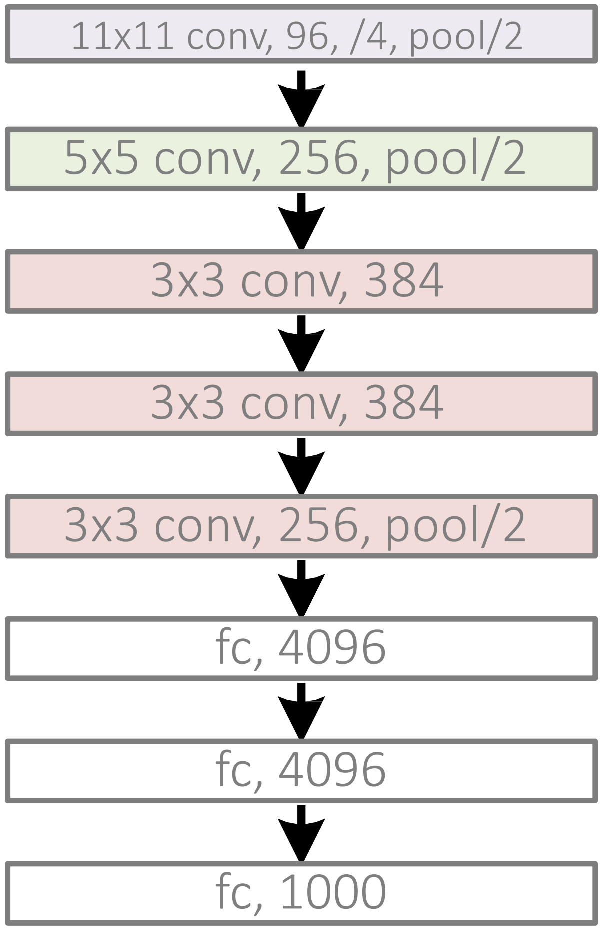


ject Detection mAP (%)

*w/ other improvements & more data

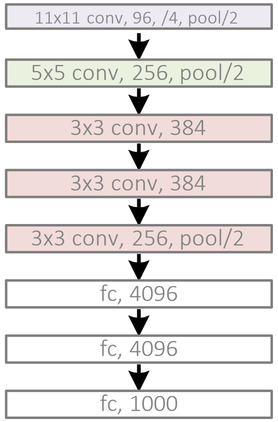
Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)

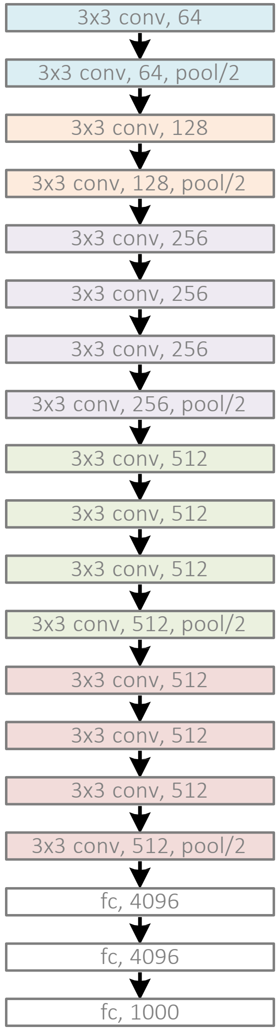


Revolution of Depth

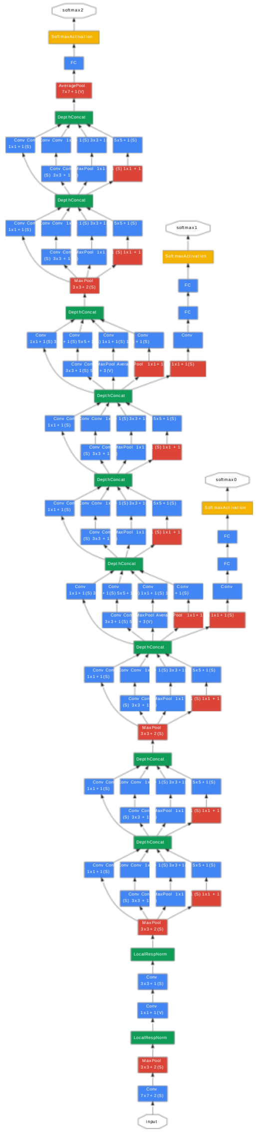
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)

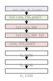


GoogleNet, 22 layers
(ILSVRC 2014)

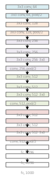


Revolution of Depth

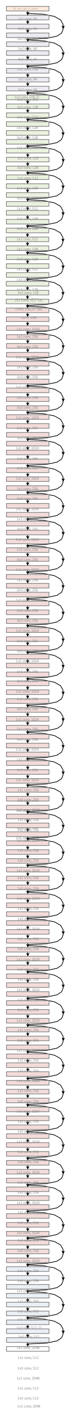
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)

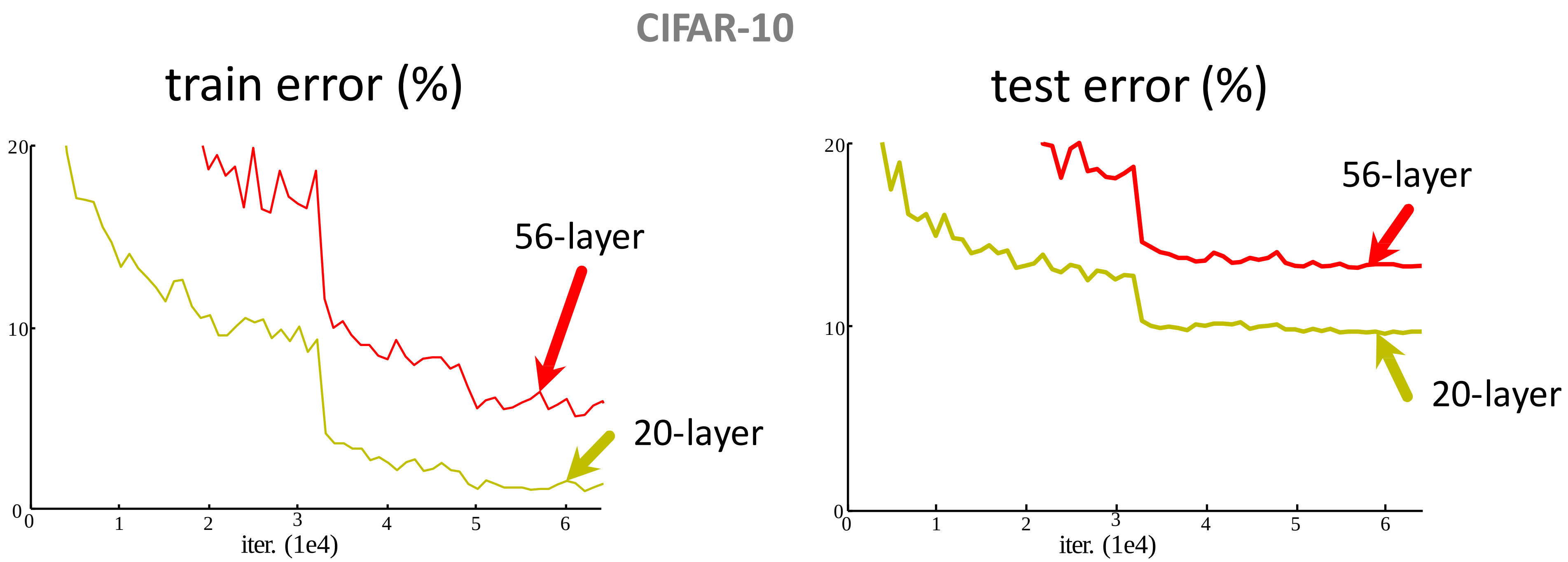


ResNet, 152 layers
(ILSVRC 2015)



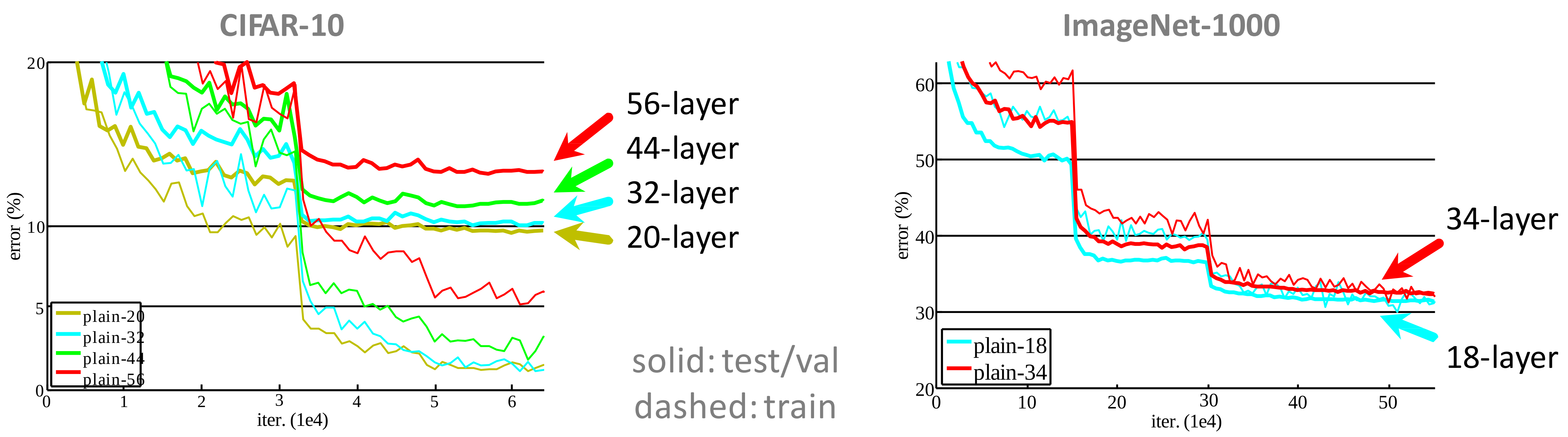
Is learning better networks
as simple as stacking more layers?

Simply stacking layers?



- *Plain* nets: stacking 3x3 conv layers…
- 56-layer net has **higher training error** and test error than 20-layer net

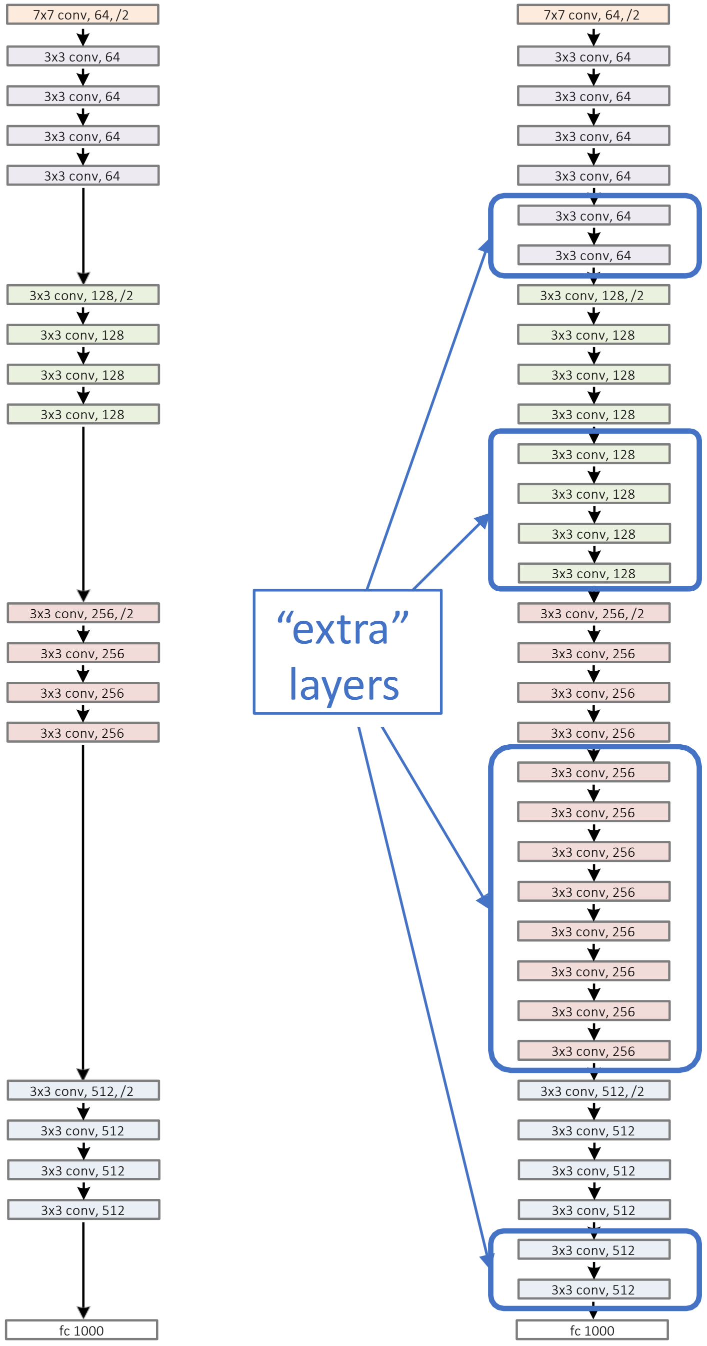
Simply stacking layers?



- "Overly deep" plain nets have **higher training error**
- A general phenomenon, observed in many datasets

a shallower model (18 layers)

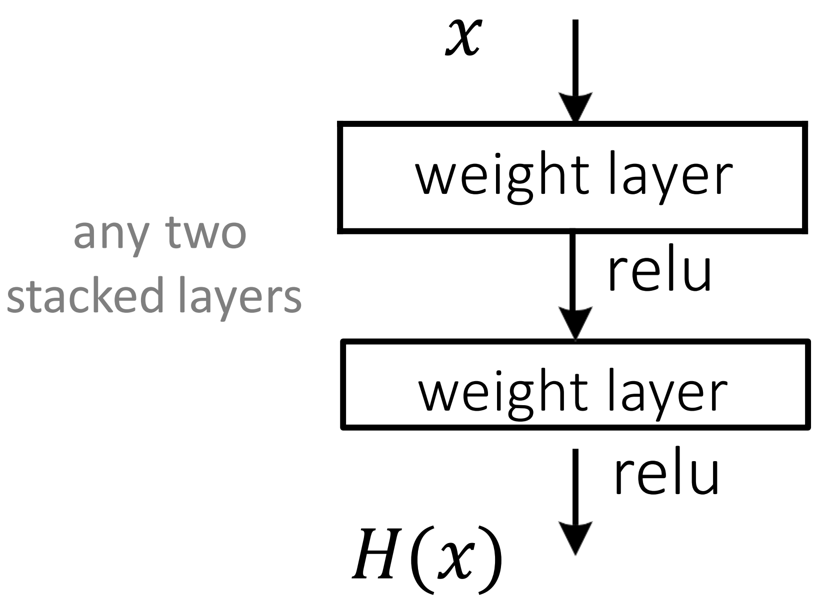
a deeper counterpart (34 layers)



- Richer solution space
- A deeper model should not have **higher training error**
- A solution *by construction*:
    - original layers: copied from a learned shallower model
    - extra layers: set as identity
    - at least the same training error
- Optimization difficulties: solvers cannot find the solution when going deeper…

Deep Residual Learning

- Plain net

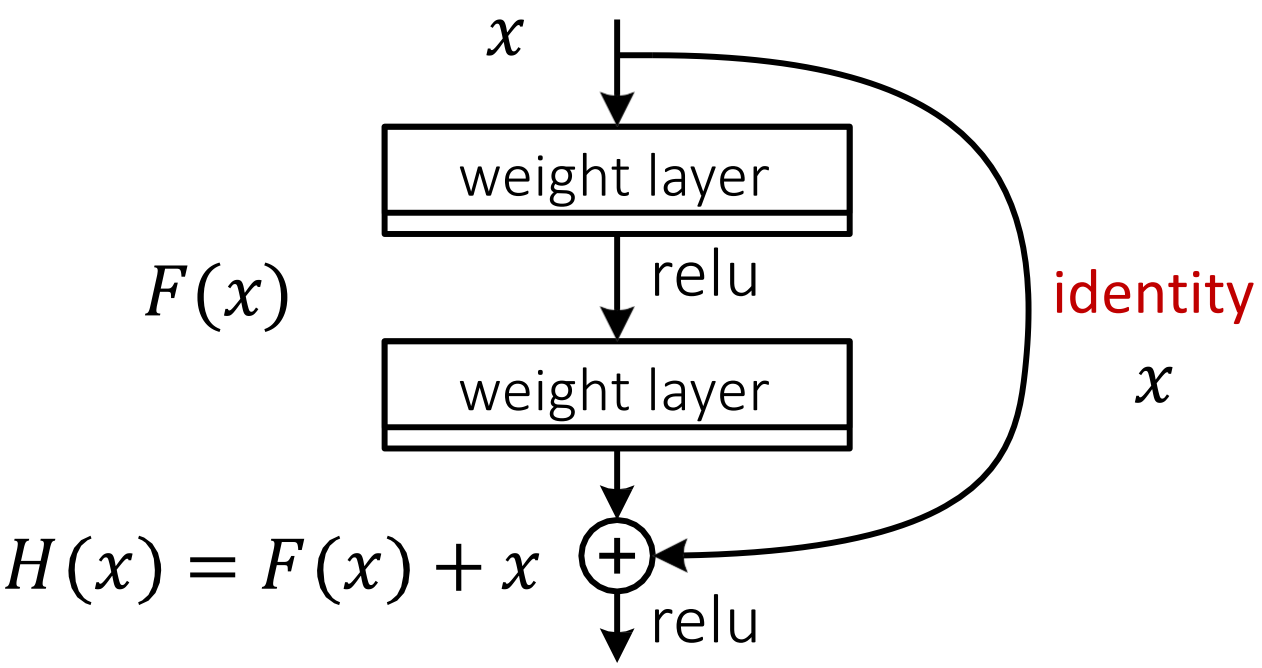


$H(x)$ is any desired mapping,

hope the 2 weight layers fit $H(x)$

Deep Residual Learning

- Residual net



$H(x)$ is any desired mapping,

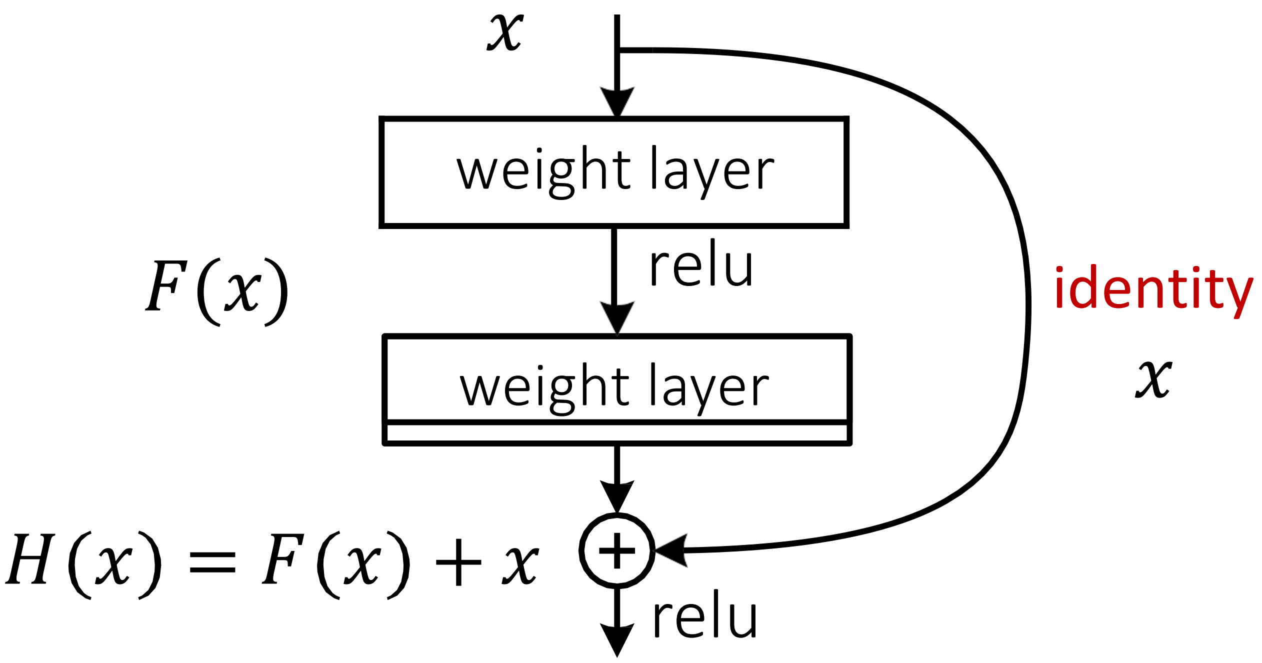
~~hope the 2 weight layers fit $H(x)$~~

hope the 2 weight layers fit $F(x)$

let $H(x) = F(x) + x$

Deep Residual Learning

- $F(x)$ is a residual mapping w.r.t. identity

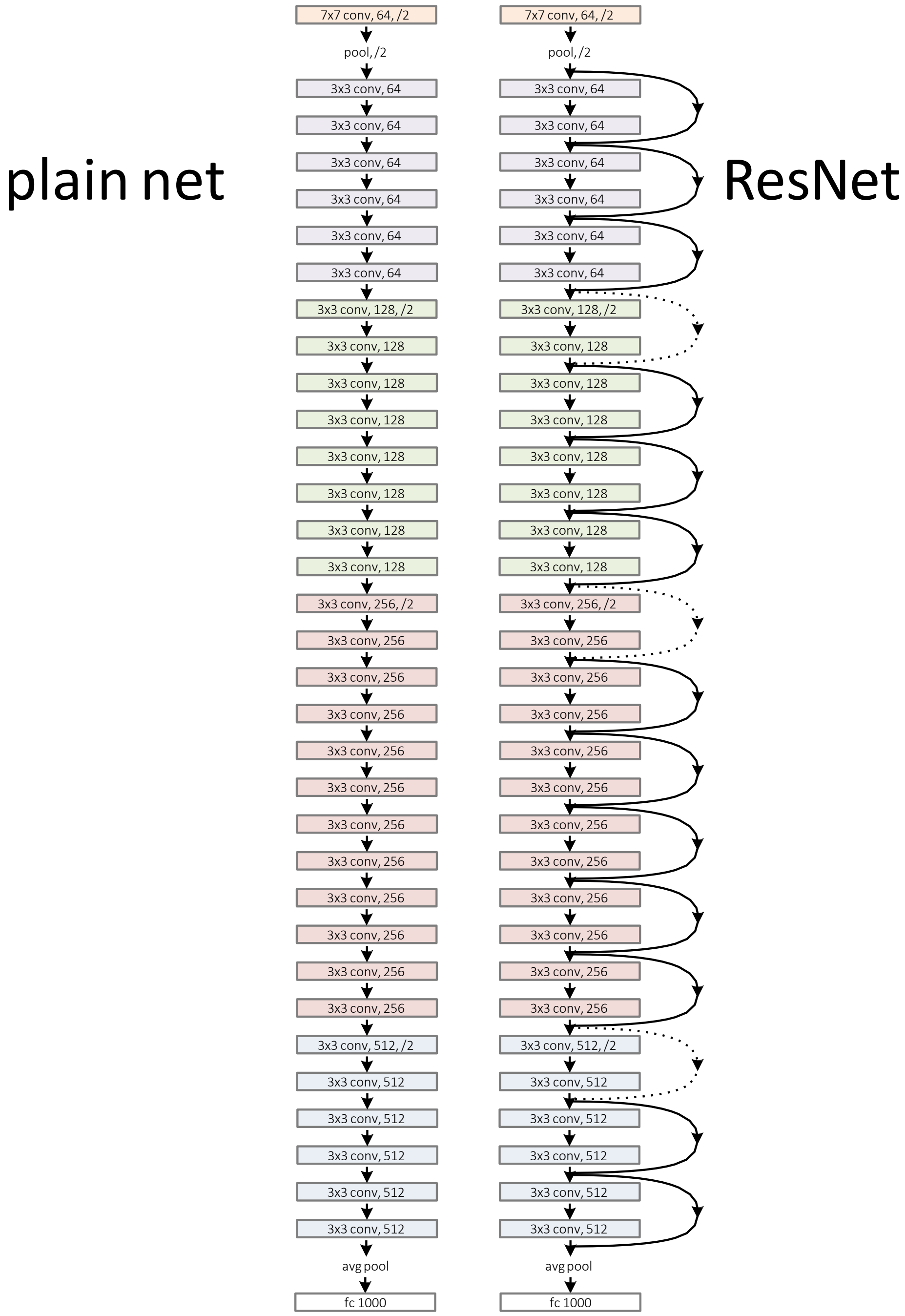


- If identity were optimal, easy to set weights as 0
- If optimal mapping is closer to identity, easier to find small fluctuations

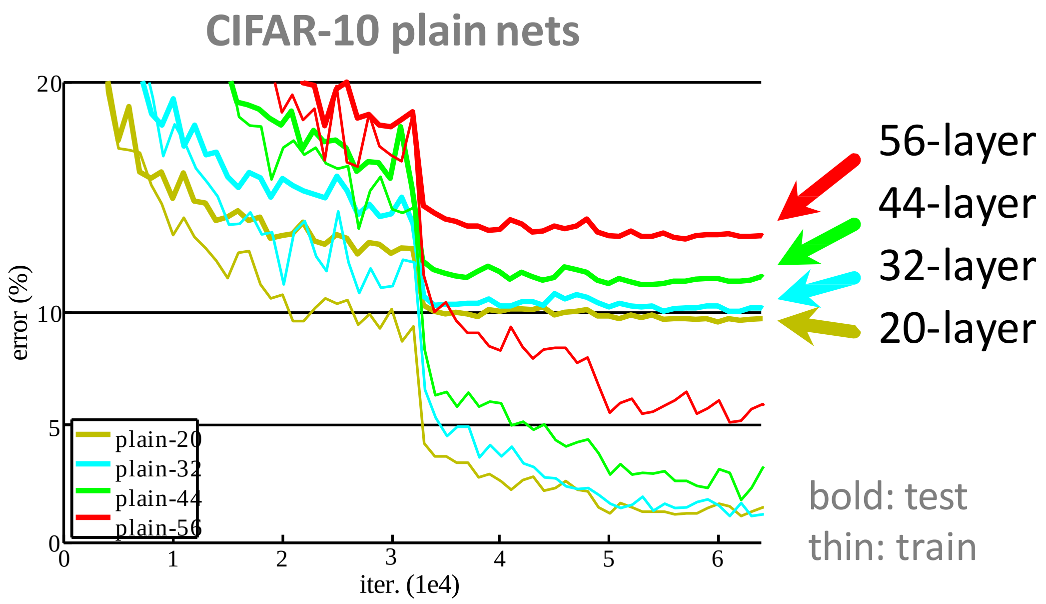
Network “Design”

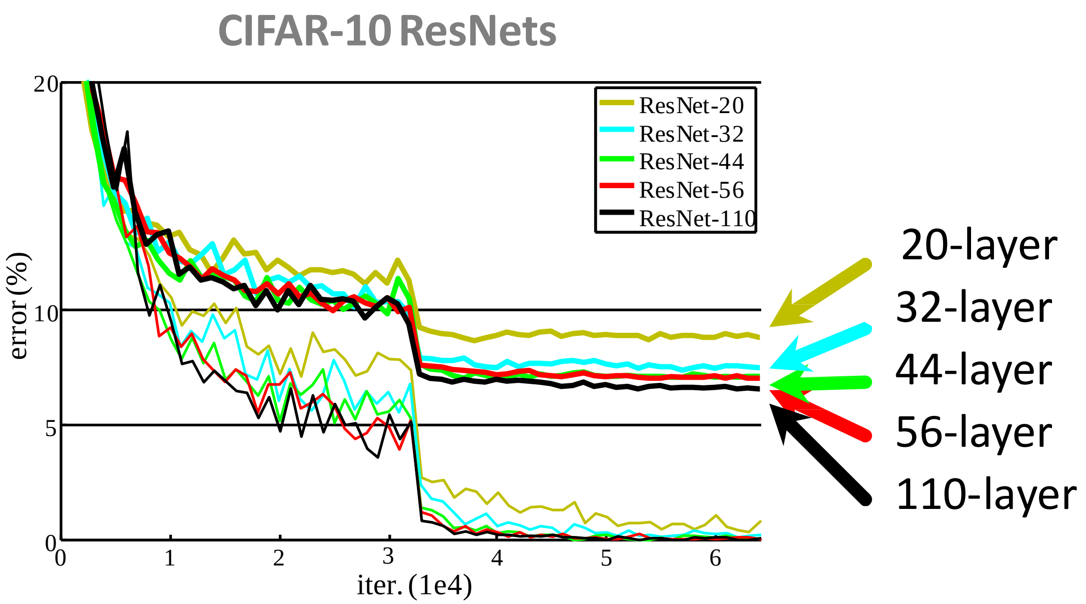
- Keep it simple

- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2
  - Simple design; just deep!



CIFAR-10 experiments

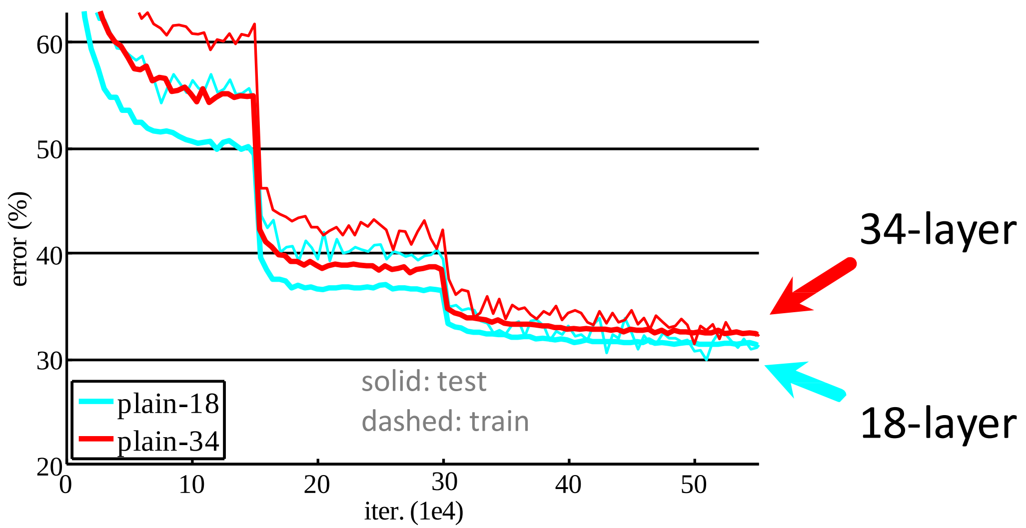




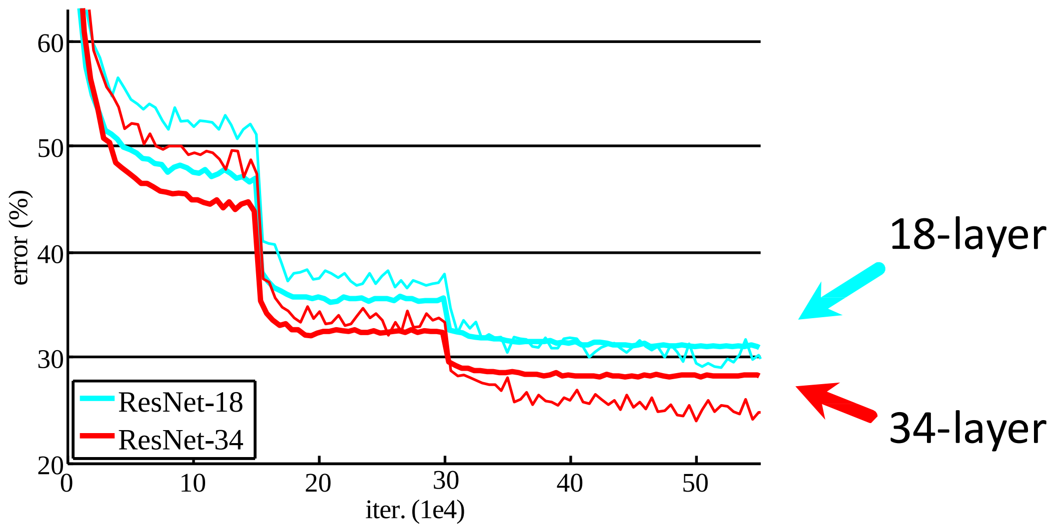
- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

ImageNet experiments

ImageNet plain nets

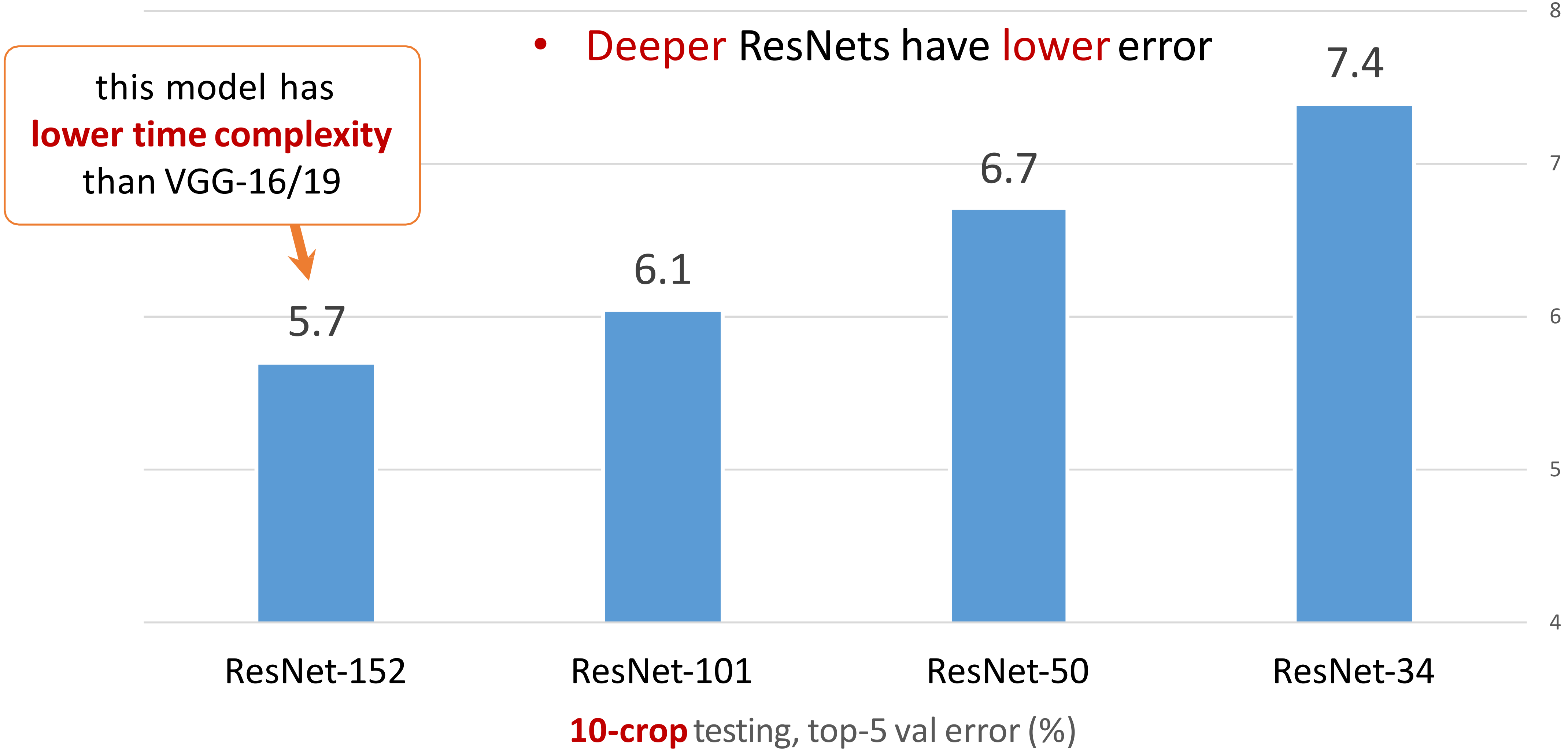


ImageNet ResNets



- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

ImageNet experiments



Beyond classification

A treasure from ImageNet is on learning features.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

"Features matter." (quote [Girshick et al. 2014], the R-CNN paper)

task	2nd-place winner	ResNets	margin (relative)
ImageNet Localization (top-5 error)	12.0	9.0	**27%**
ImageNet Detection (mAP@.5)	53.6	62.1	**16%**
COCO Detection (mAP@.5:.95)	33.5	37.3	**11%**
COCO Segmentation (mAP@.5:.95)	25.1	28.2	**12%**

absolute 8.5% better!

- Our results are all based on ResNet-101
- Our features are well transferrable

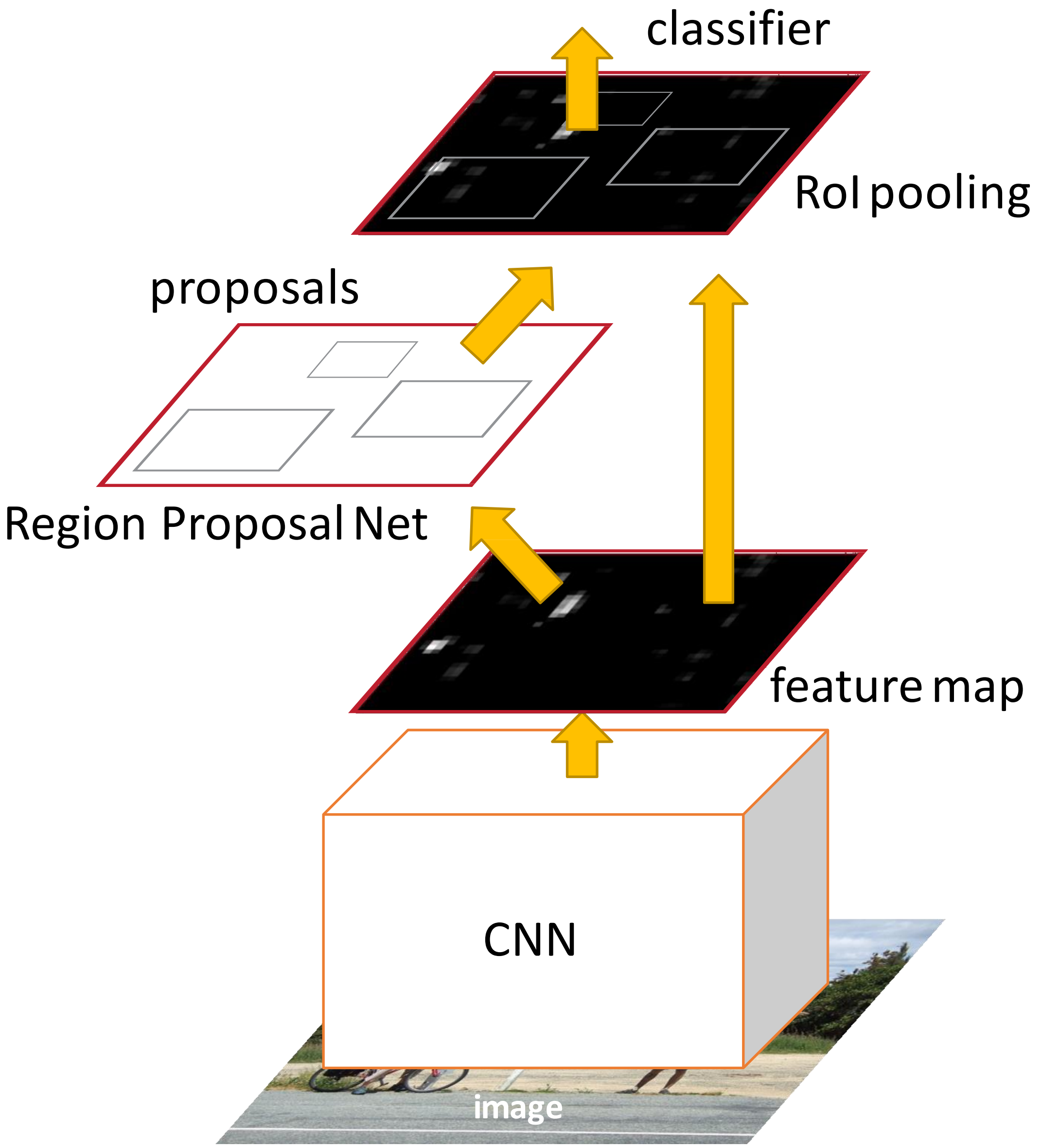
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.

Object Detection (brief)

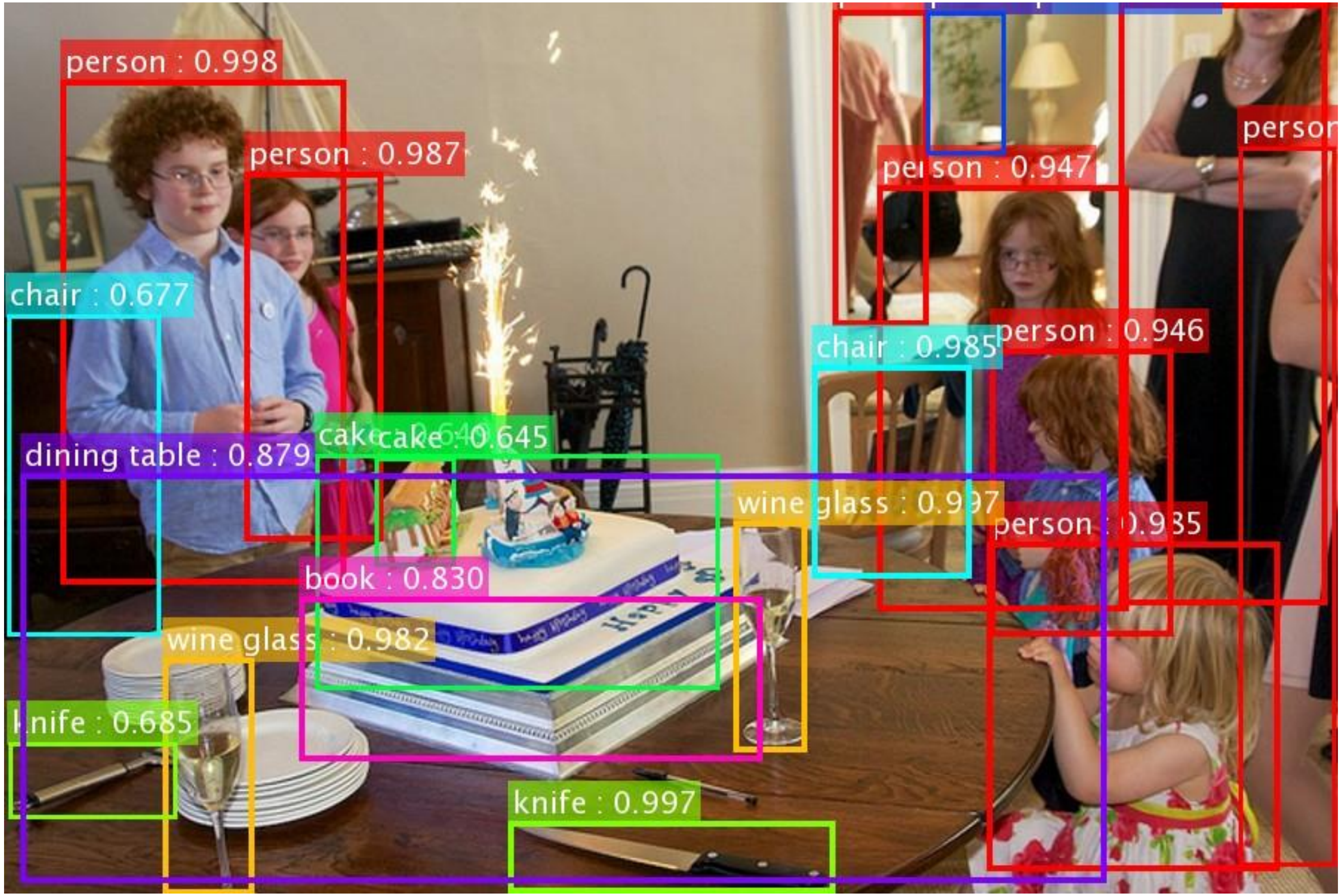
- Simply "Faster R-CNN + ResNet"

Faster R-CNN baseline	mAP@.5	mAP@.5:.95
VGG-16	41.5	21.5
ResNet-101	**48.4**	**27.2**

COCO detection results
(ResNet has 28% relative gain)



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.



Our results on MS COCO

*the original image is from the COCO dataset

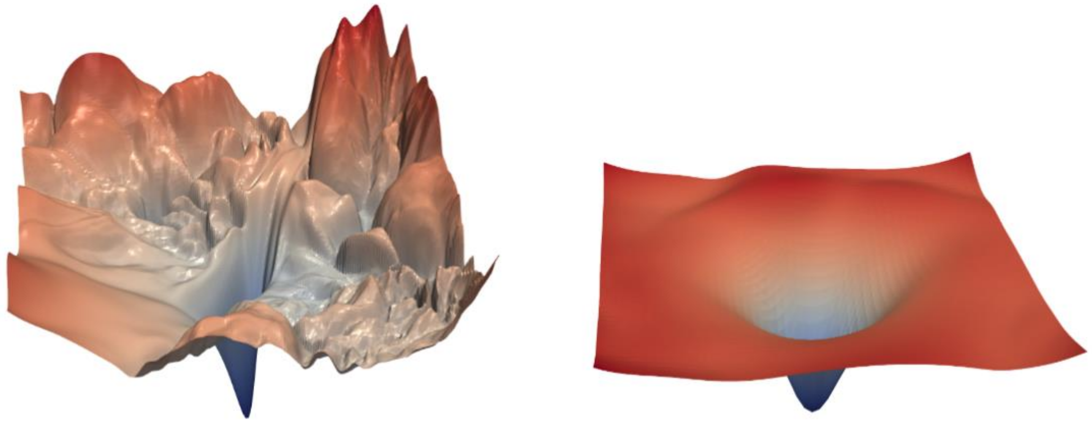
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.

Why does ResNet work so well?

- The architecture is somehow easier to optimize.
- The authors argue it probably isn’t because it solves the “vanishing gradient” problem.
- While the gradients might not be “vanishing” in “plain” nets, they don’t seem as stable and trustworthy, according to follow up work, e.g.

Visualizing the Loss Landscape of Neural Nets. Hao Li, Zheng Xu , Gavin Taylor, Christoph Studer, Tom Goldstein. NeurIPS 2018.

We argue that this optimization difficulty is *unlikely* to be caused by vanishing gradients. These plain networks are trained with BN [16], which ensures forward propagated signals to have non-zero variances. We also verify that the backward propagated gradients exhibit healthy norms with BN. So neither forward nor backward signals vanish. In



(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.