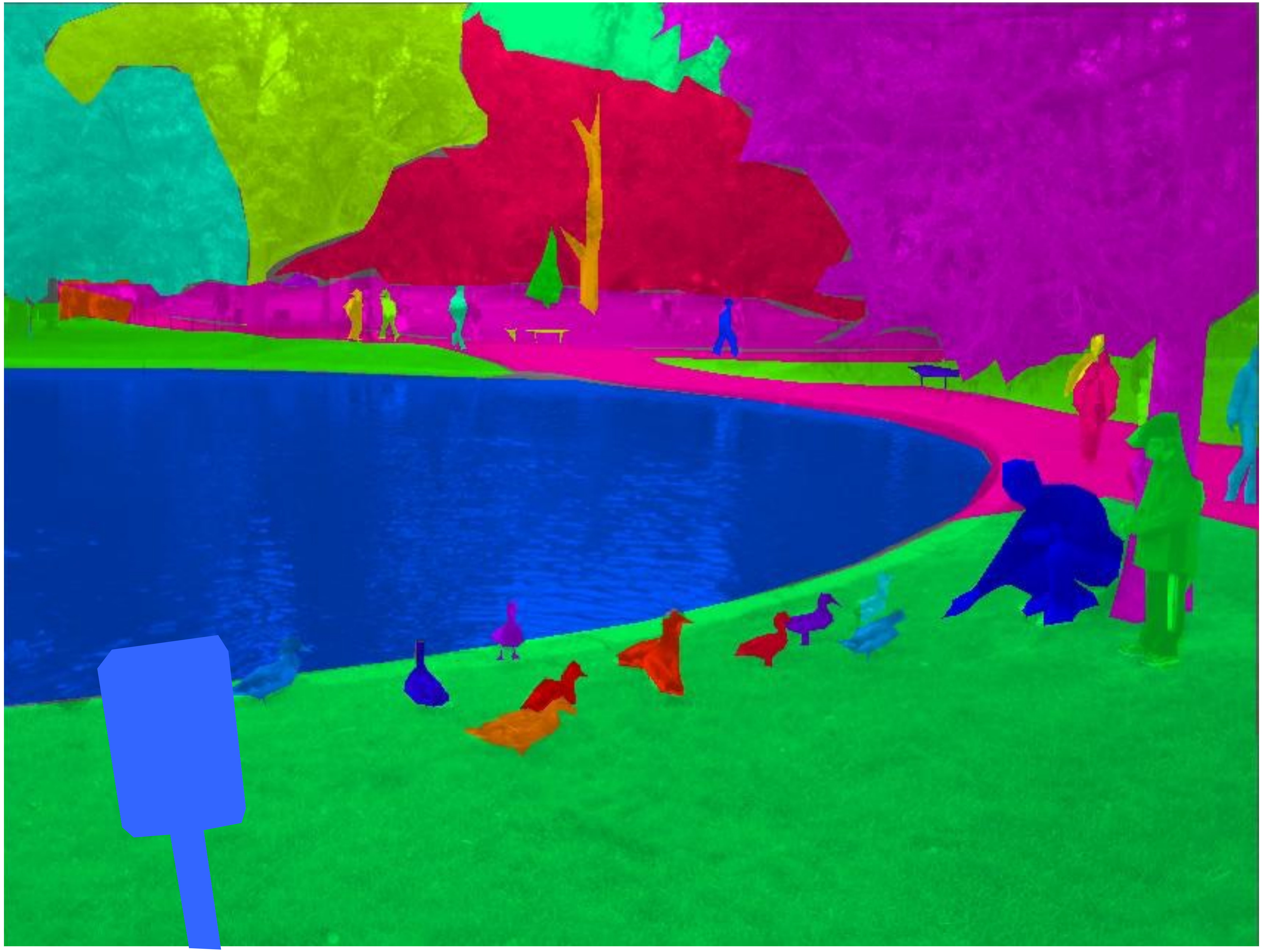
Lecture 10:
Introduction to scene understanding

Announcements

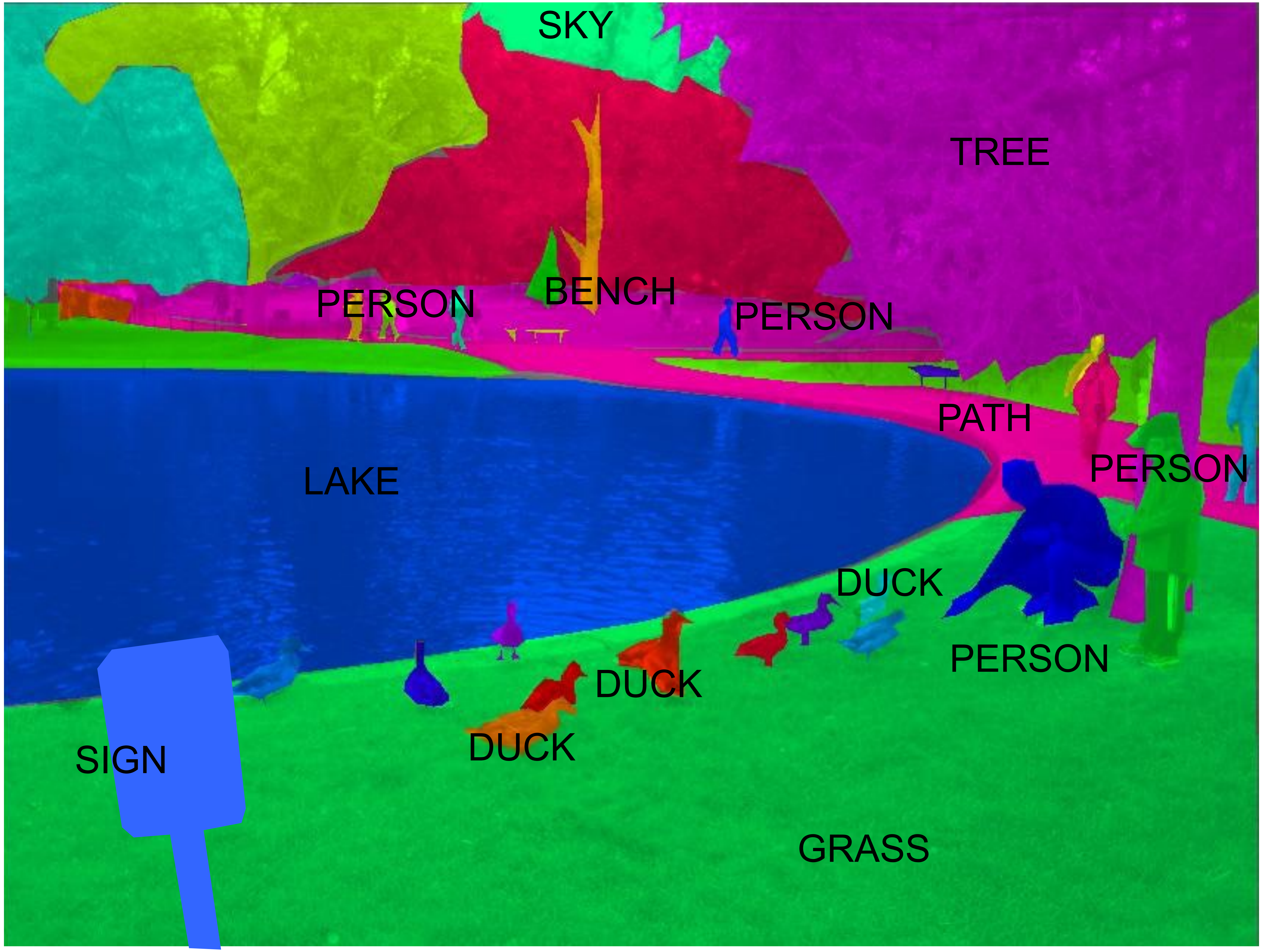
- PS1 grades out
  - Please check your grade!
  - Regrade requests via Gradescope
  - Submit requests by **next Tues**.
- PS5 out
  - More coding than usual!

Image contains Photoshopped sign

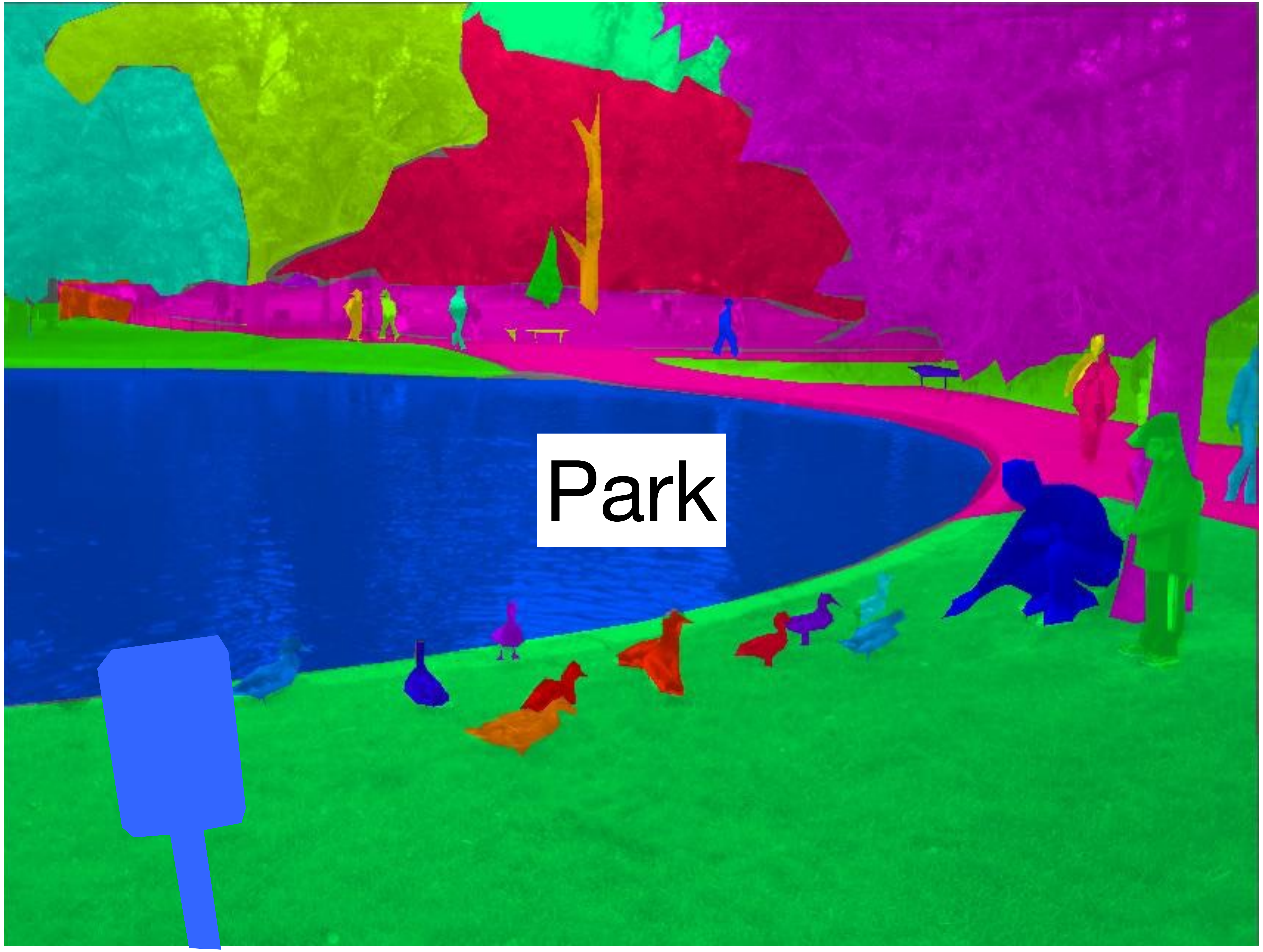
Source: Torralba, Freeman, Isola



Source: Torralba, Freeman, Isola

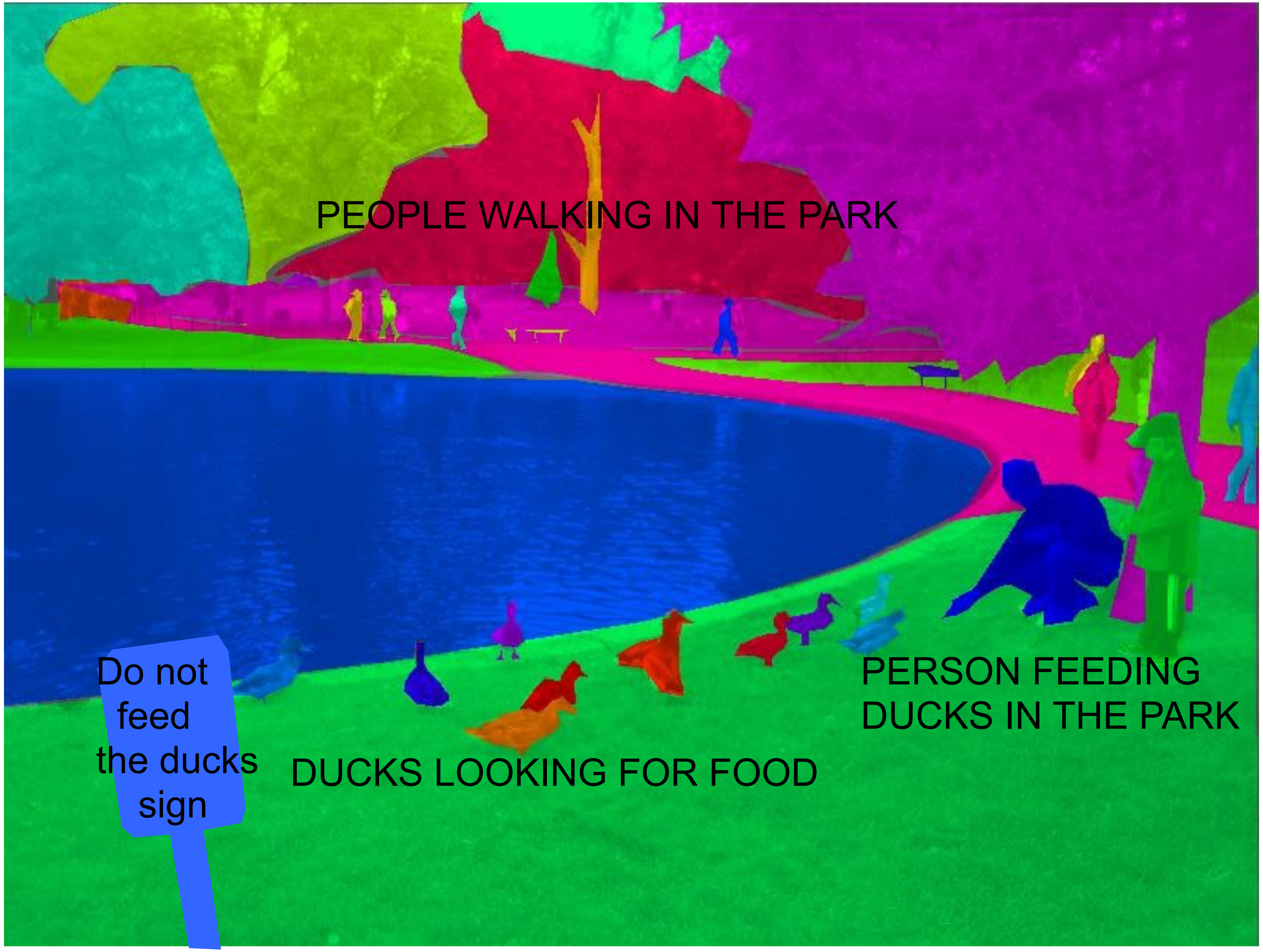


Source: Torralba, Freeman, Isola



Source: Torralba, Freeman, Isola

Source: Torralba, Freeman, Isola



Source: Torralba, Freeman, Isola

Source: Torralba, Freeman, Isola

Today

- History
- Scene recognition
- Pixel labeling problems
- Simple object detection model

Why do we care about recognition?

Perception of function: We can perceive the 3D shape, texture, material properties, without knowing about objects. But, the concept of category encapsulates also information about what can we do with those objects.

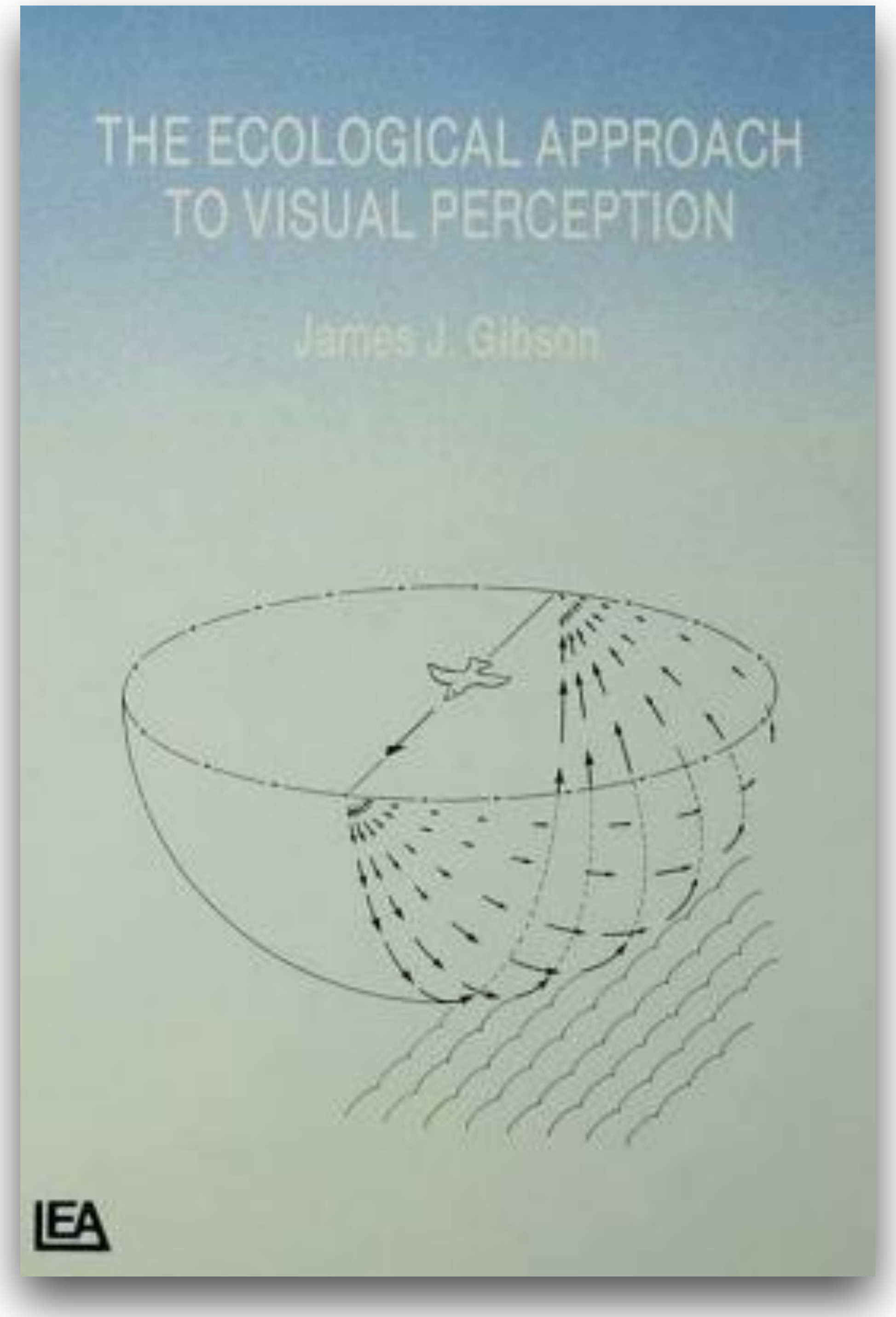
“We therefore include the perception of function as a proper –indeed, crucial- subject for vision science”, *from Vision Science, chapter 9, Palmer*.

Source: Torralba, Freeman, Isola

The perception of function

- Direct perception (affordances): Gibson

Flat surface
Horizontal
Knee-high
…
→ Sittable upon



- Mediated perception (Categorization)

Flat surface
Horizontal
Knee-high
…
→ Chair → Sittable upon

Chair
Chair
Chair?

Source: Torralba, Freeman, Isola

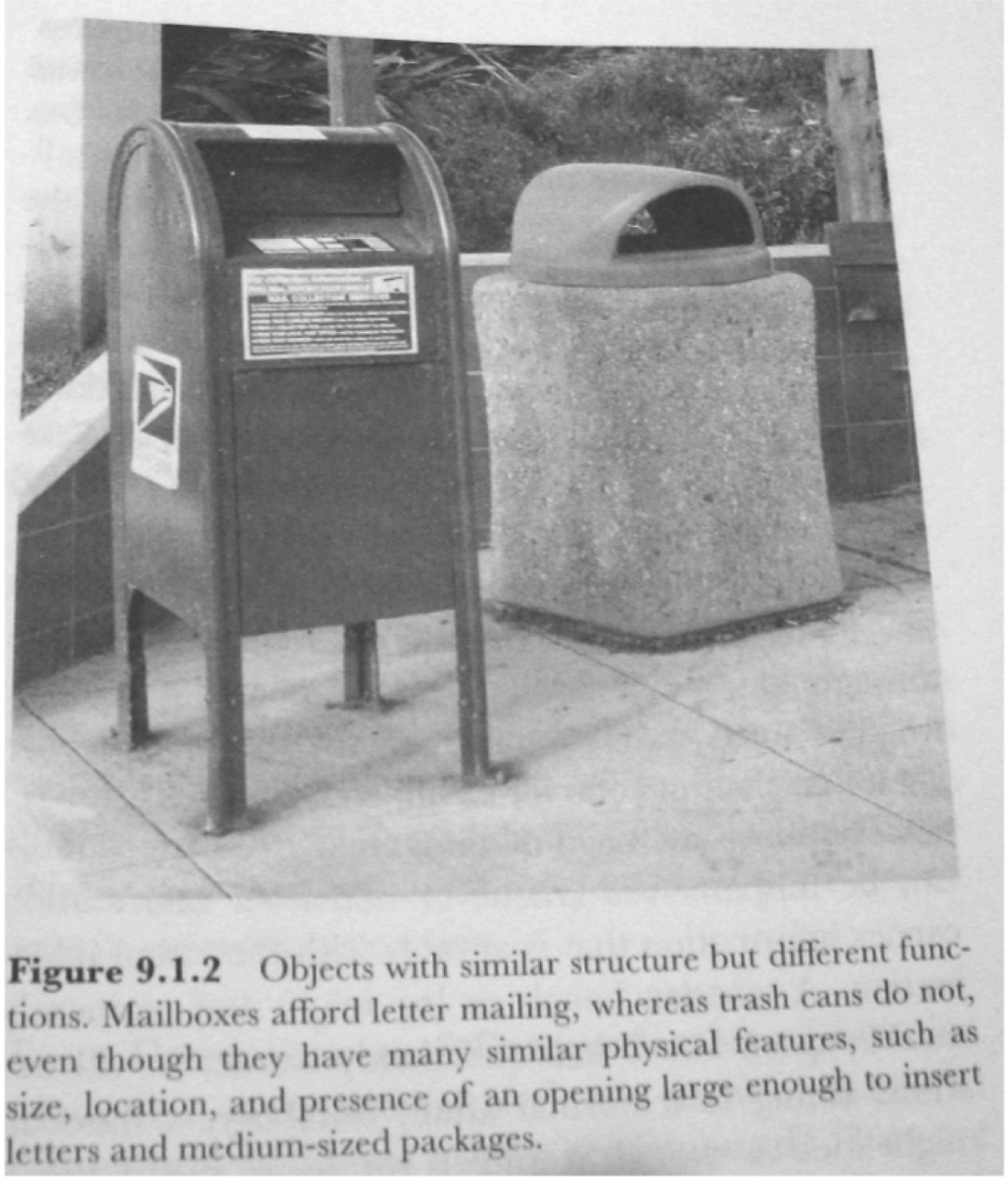
Direct perception

Some aspects of an object function can be perceived directly

Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)

Source: Torralba, Freeman, Isola

Limits to direct perception



Source: Torralba, Freeman, Isola

Object categories aren't everything

Slide by Fei Fei, Fergus & Torralba

Object categories aren't everything

A picture is worth a 1000 words…

Or just 10?

sky

building

flag

face

banner

wall

street lamp

bus

bus

cars

Source: A. Efros

Visual challenges with categories

- A lot of categories are functional

Chair



- Categories are 3D, but images are 2D

car


- World is highly varied


train

Source: A. Efros

What labels? Recognizing exact instances?

A Beijing City Transit Bus #17, serial number 43253?

Source: A. Efros

Need more general (useful) information

What can we say the very first time we see this thing?

Functional:

- A large vehicle that may be moving fast, probably to the right, and will kill you if you stand in its way.
- However, at specified places, it will allow you to enter it and transport you quickly over large distances.

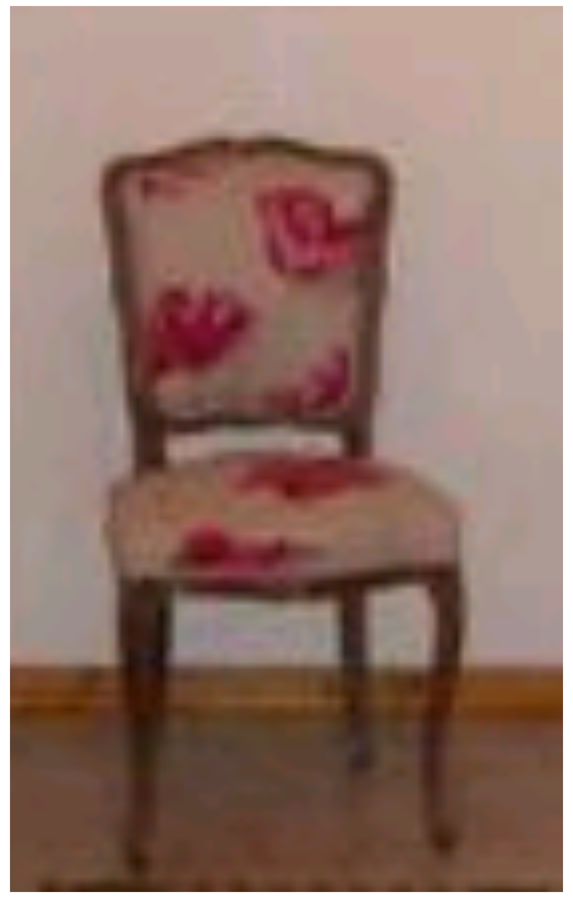
Communicational:

- bus, autobus, λεωφορείο, ônibus, автобус, 公共汽车, etc.

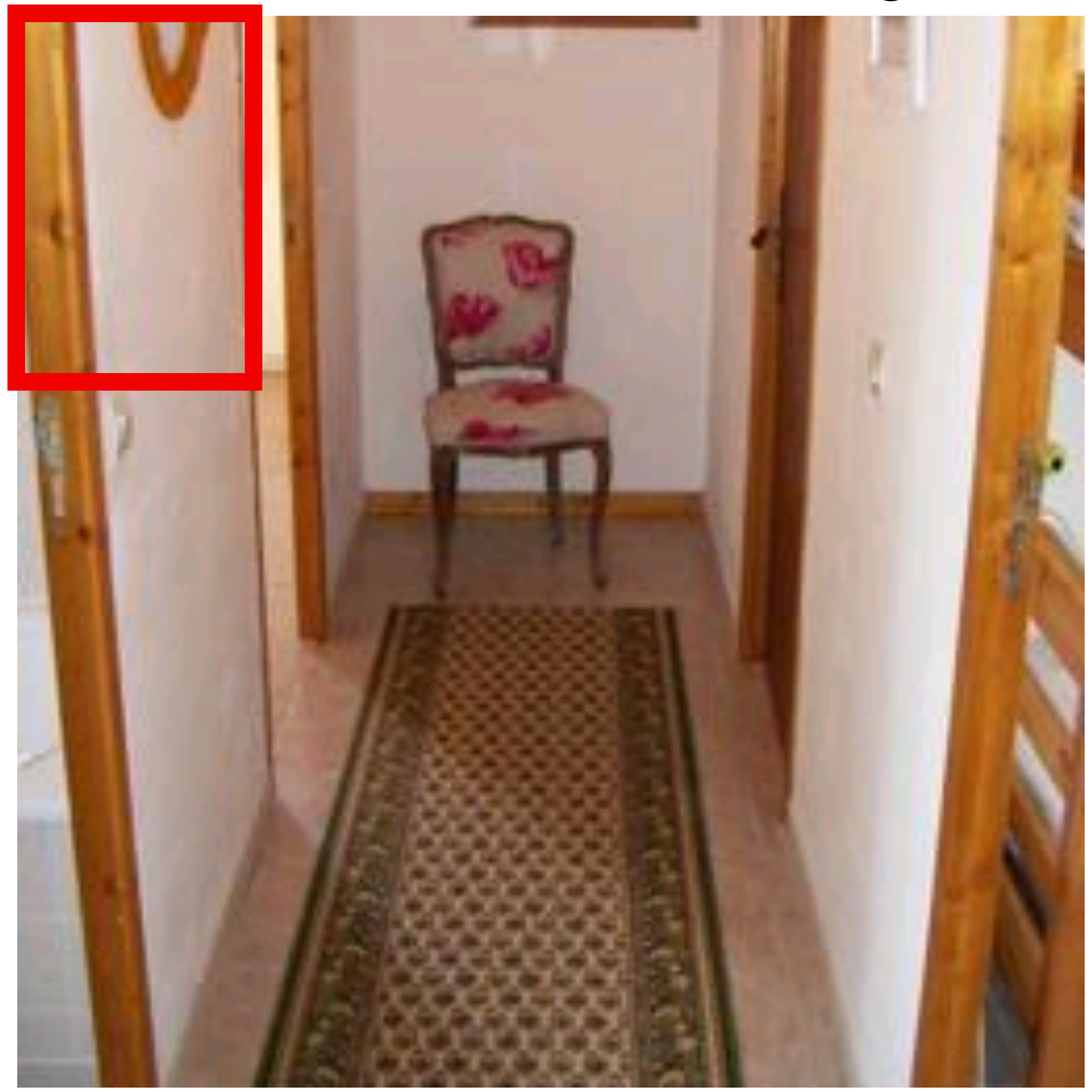
Source: A. Efros

Recognizing objects: is it really so hard?

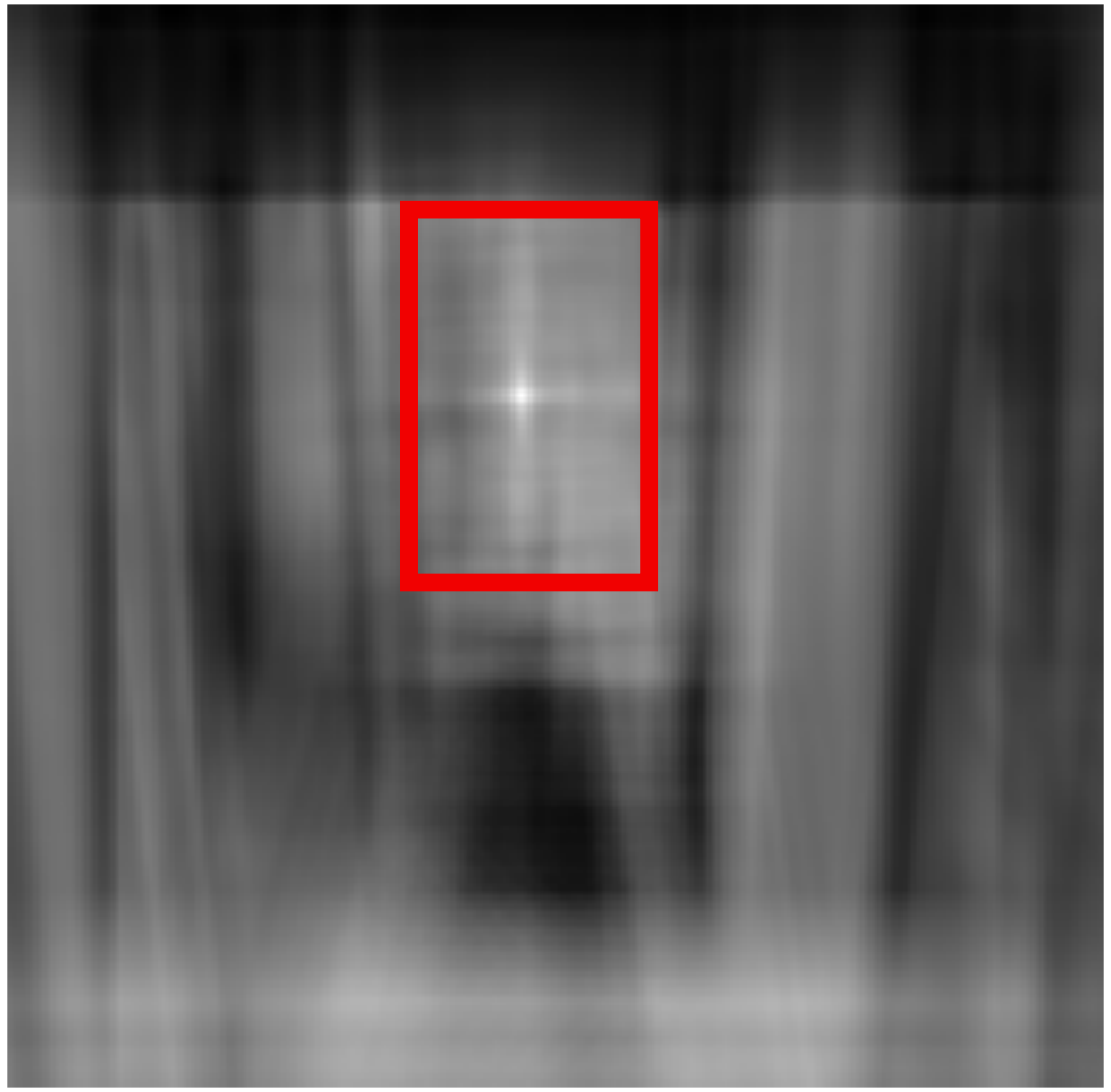
This is a chair



Find the chair in this image

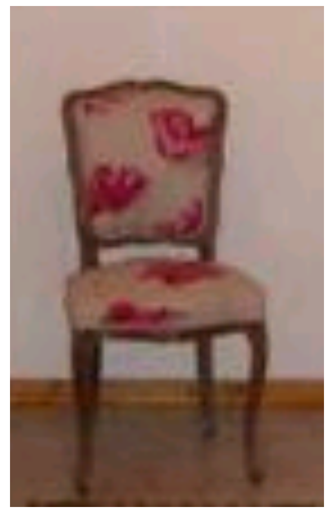


Output of normalized correlation

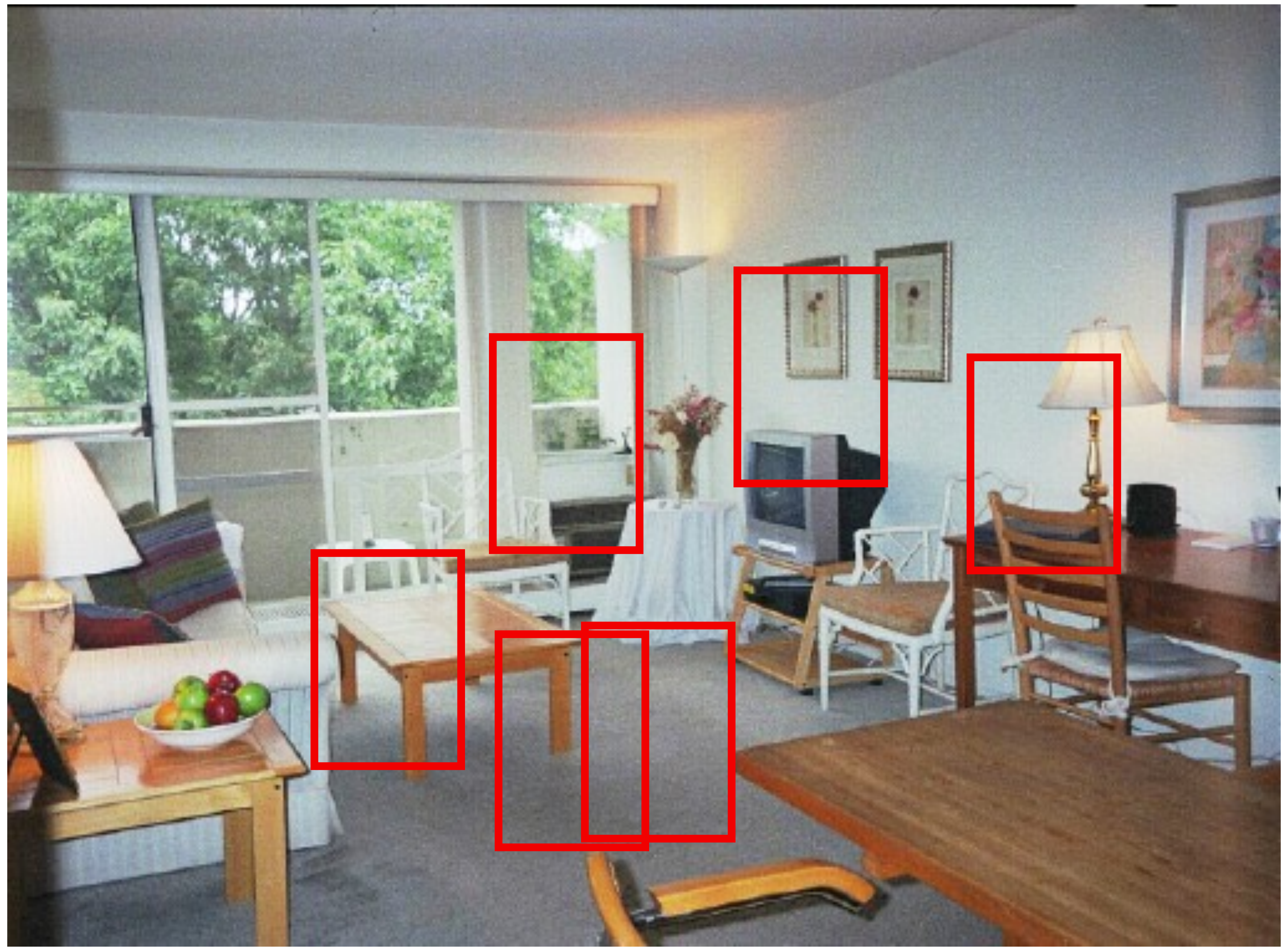


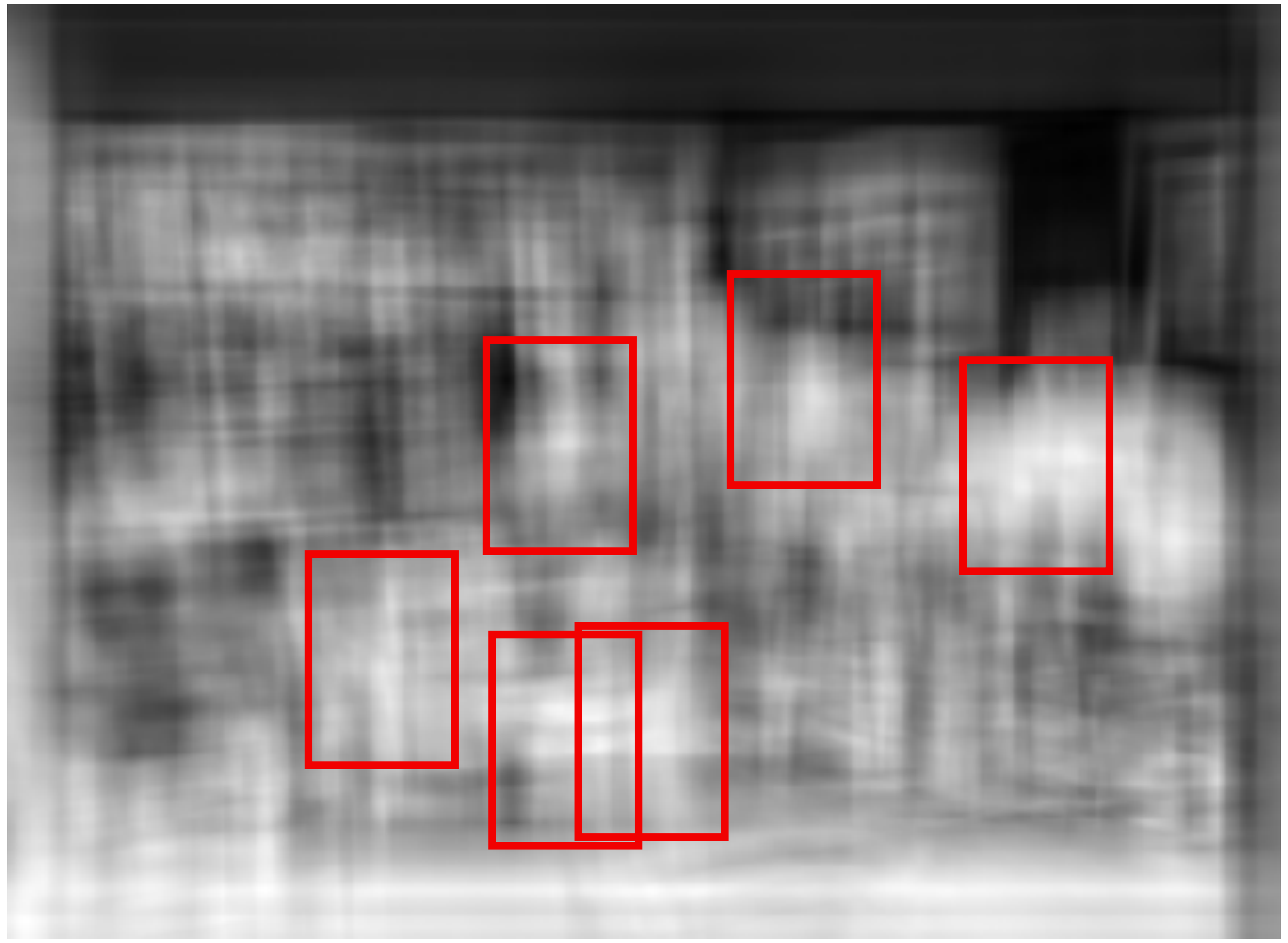
Source: A. Torralba

Recognizing objects: is it really so hard?



Find the chair in this image



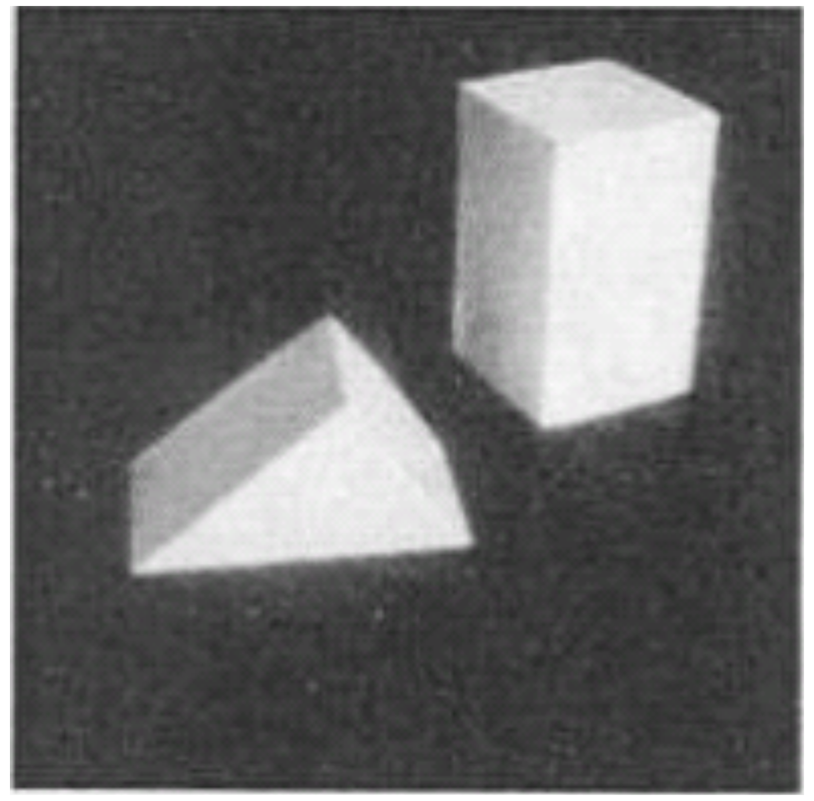


Not so great!

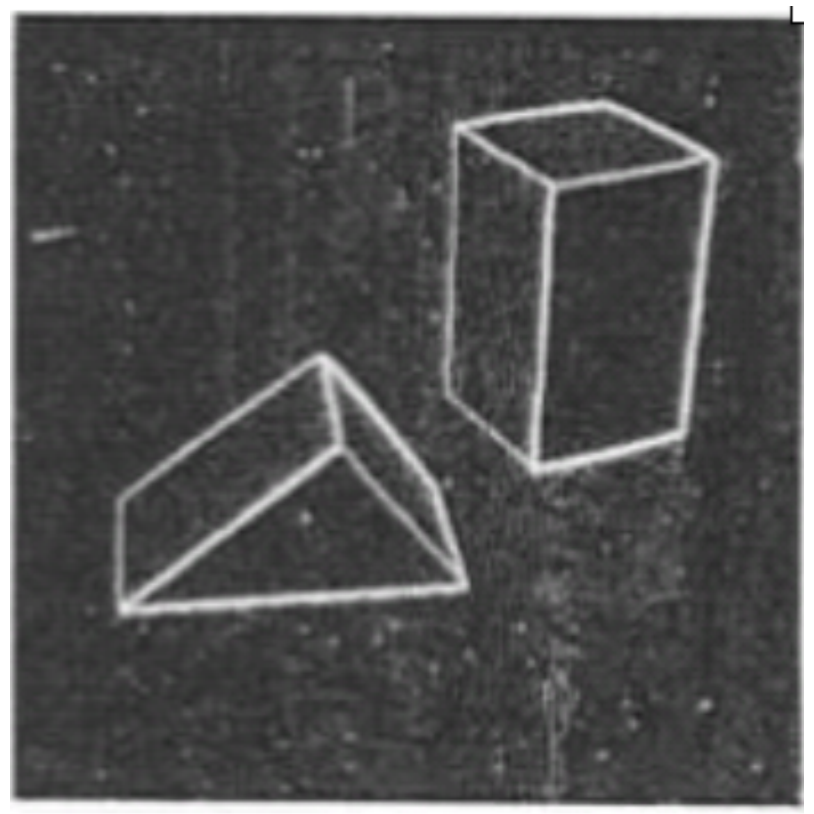
Source: A. Torralba

Simplify the problem: Blocks world

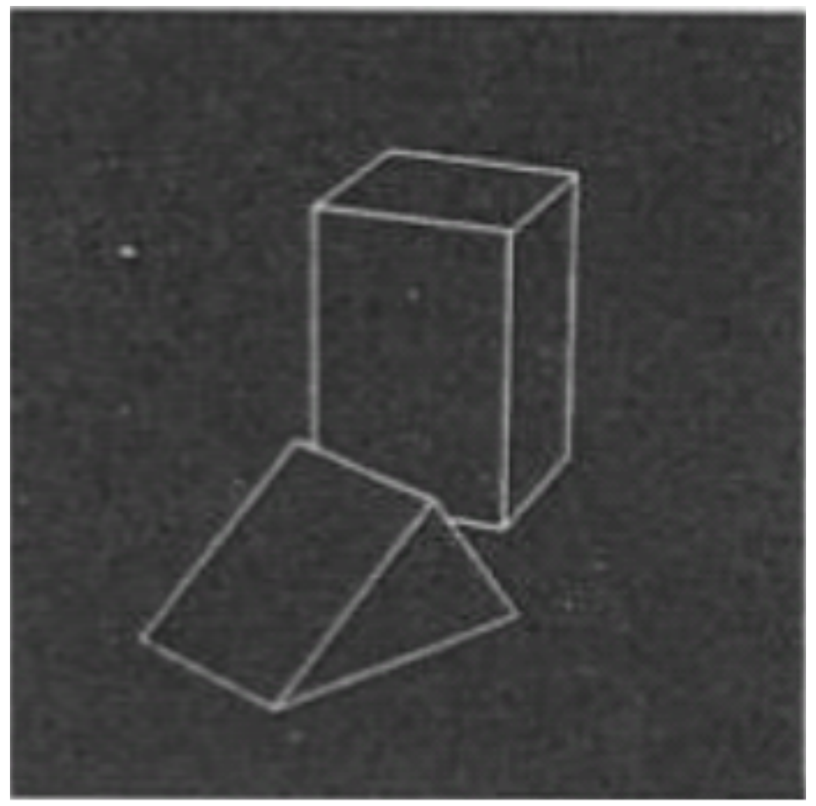
Input

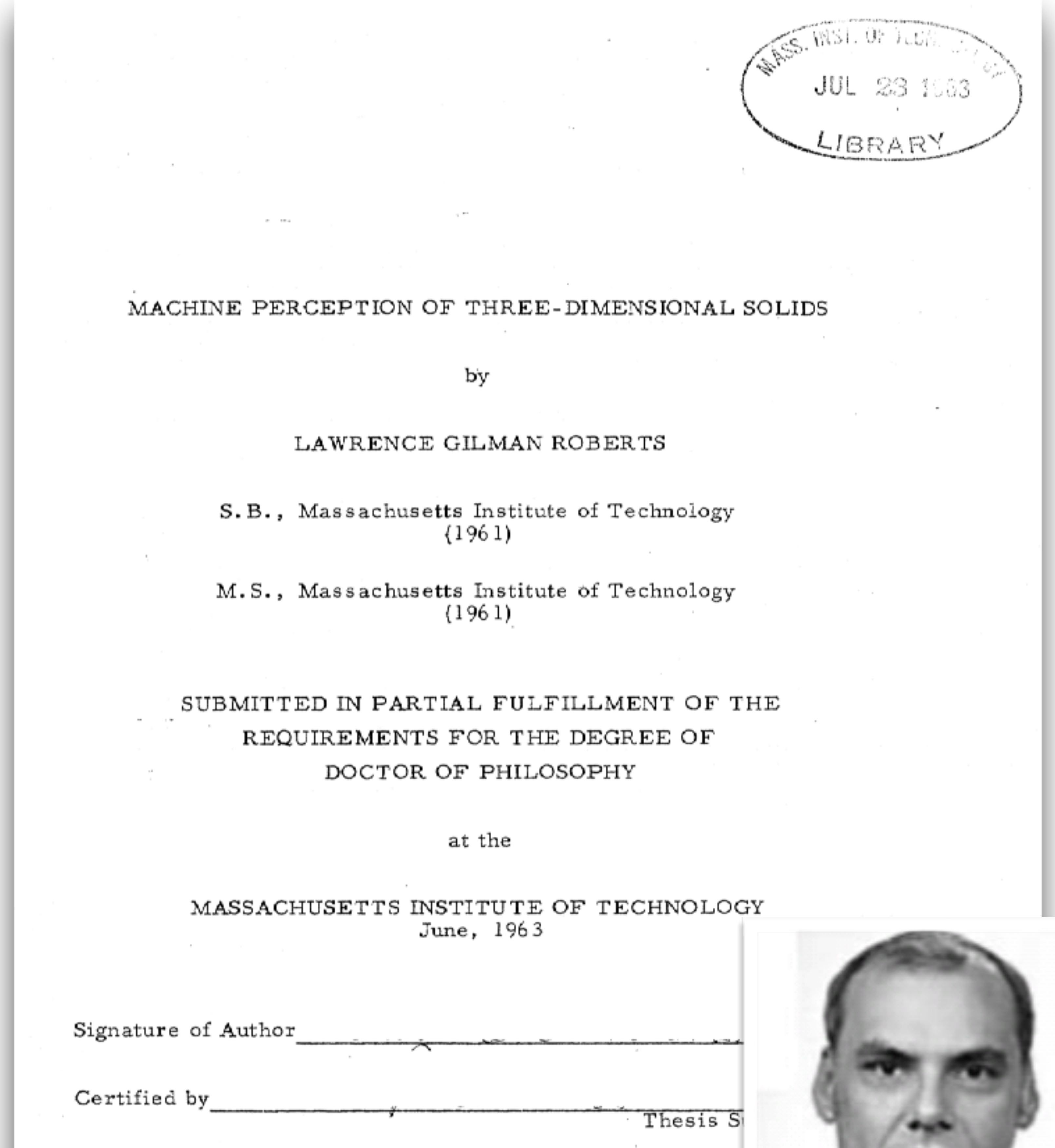


Edges (2x2 gradient)



Reconstructed 3D scene
(new view point)





MACHINE PERCEPTION OF THREE-DIMENSIONAL SOLIDS

by

LAWRENCE GILMAN ROBERTS

S.B., Massachusetts Institute of Technology
(1961)

M.S., Massachusetts Institute of Technology
(1961)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

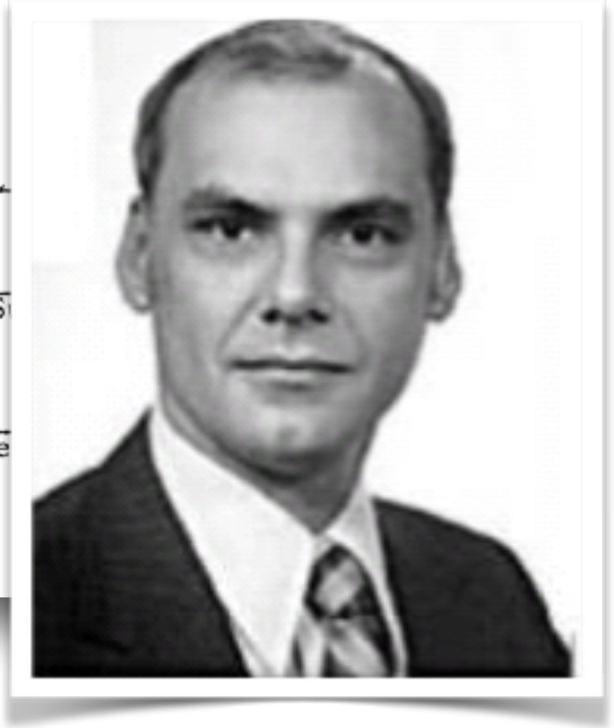
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1963

Signature of Author

Certified by
Thesis S

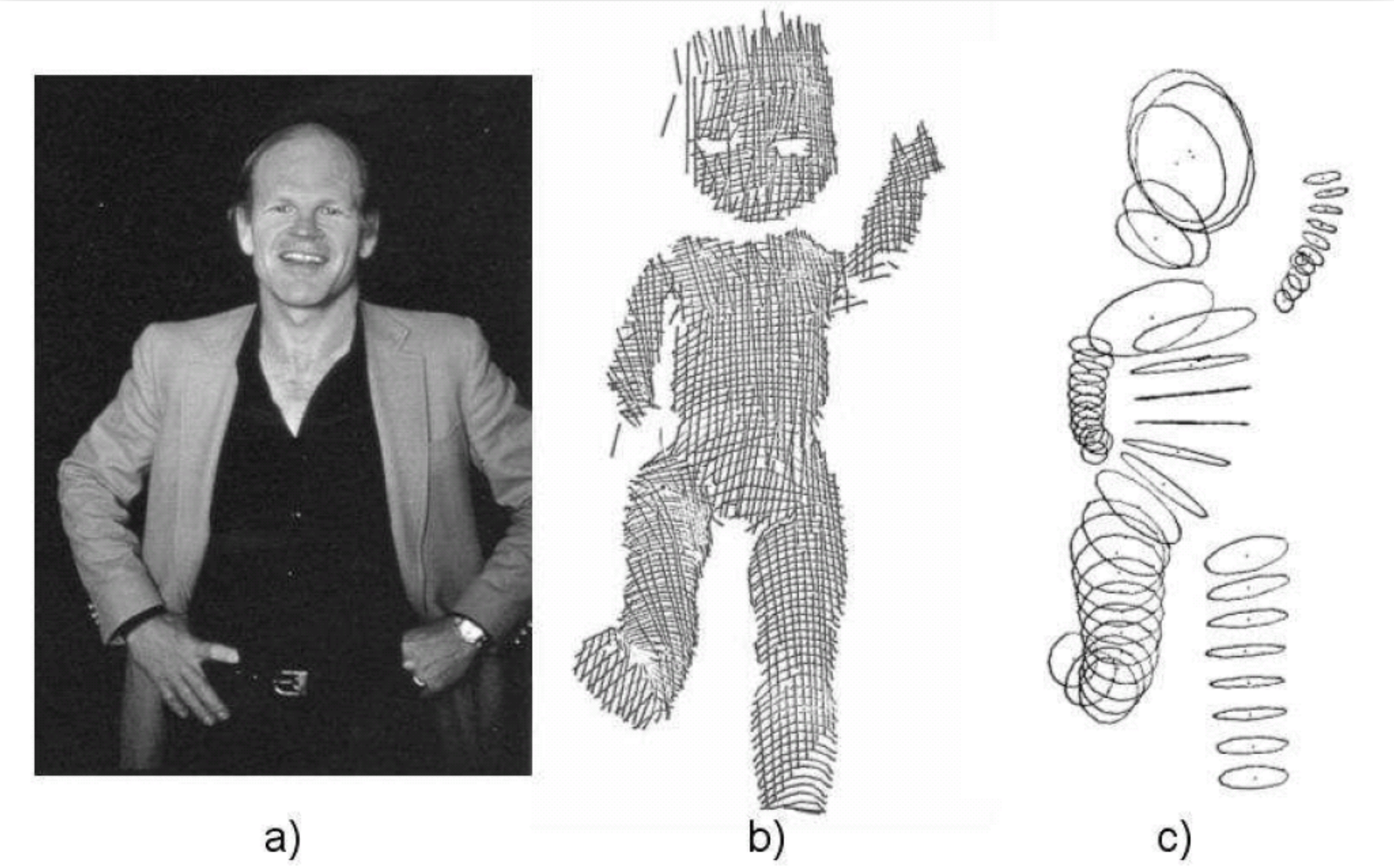
Accepted by
Chairman, Departmental Committee on Graduate



Source: Torralba, Freeman, Isola

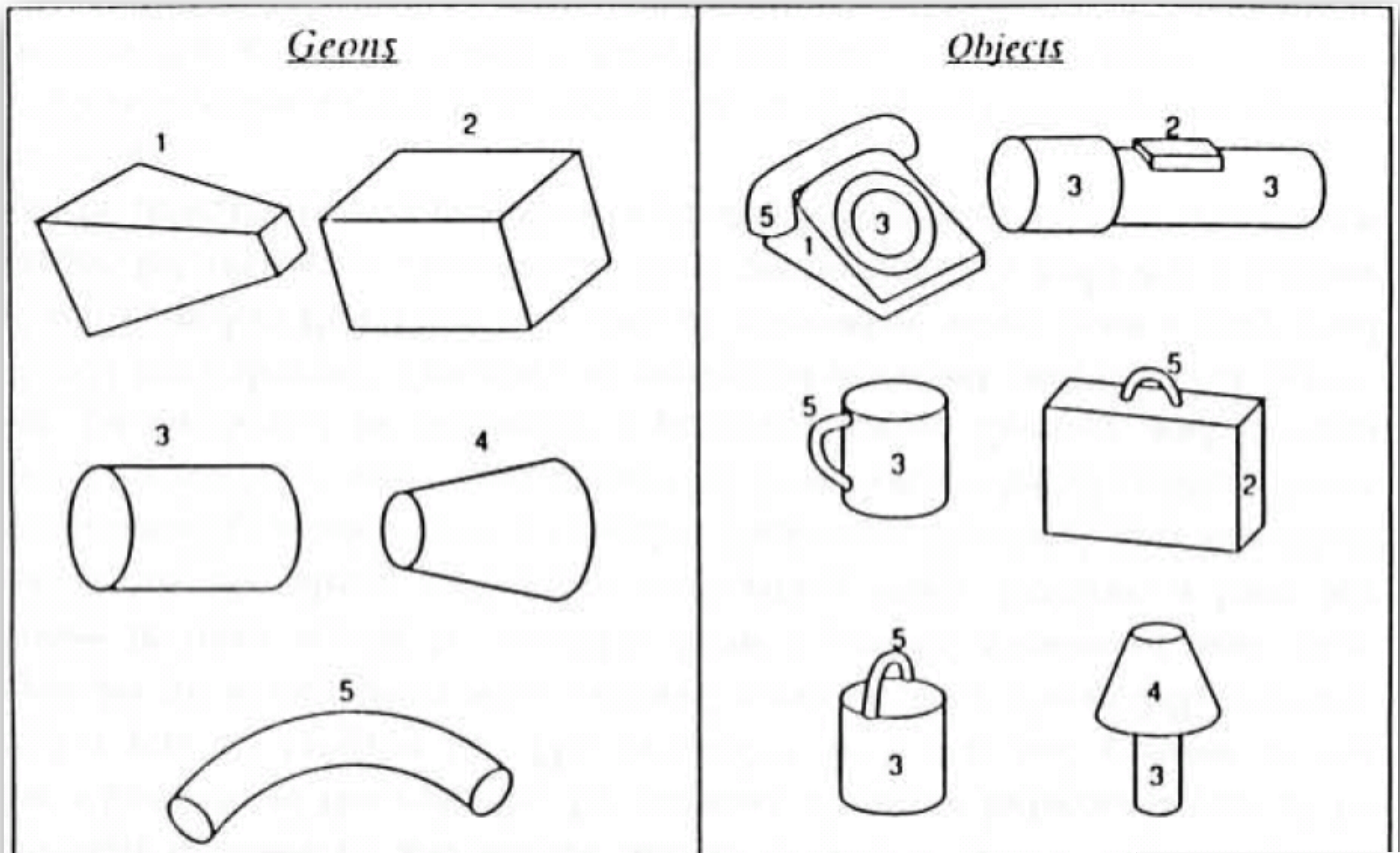
3D, compositional models

Binford and generalized cylinders



Object Recognition in the Geometric Era: a Retrospective. Joseph L. Mundy. 2006

Recognition by components

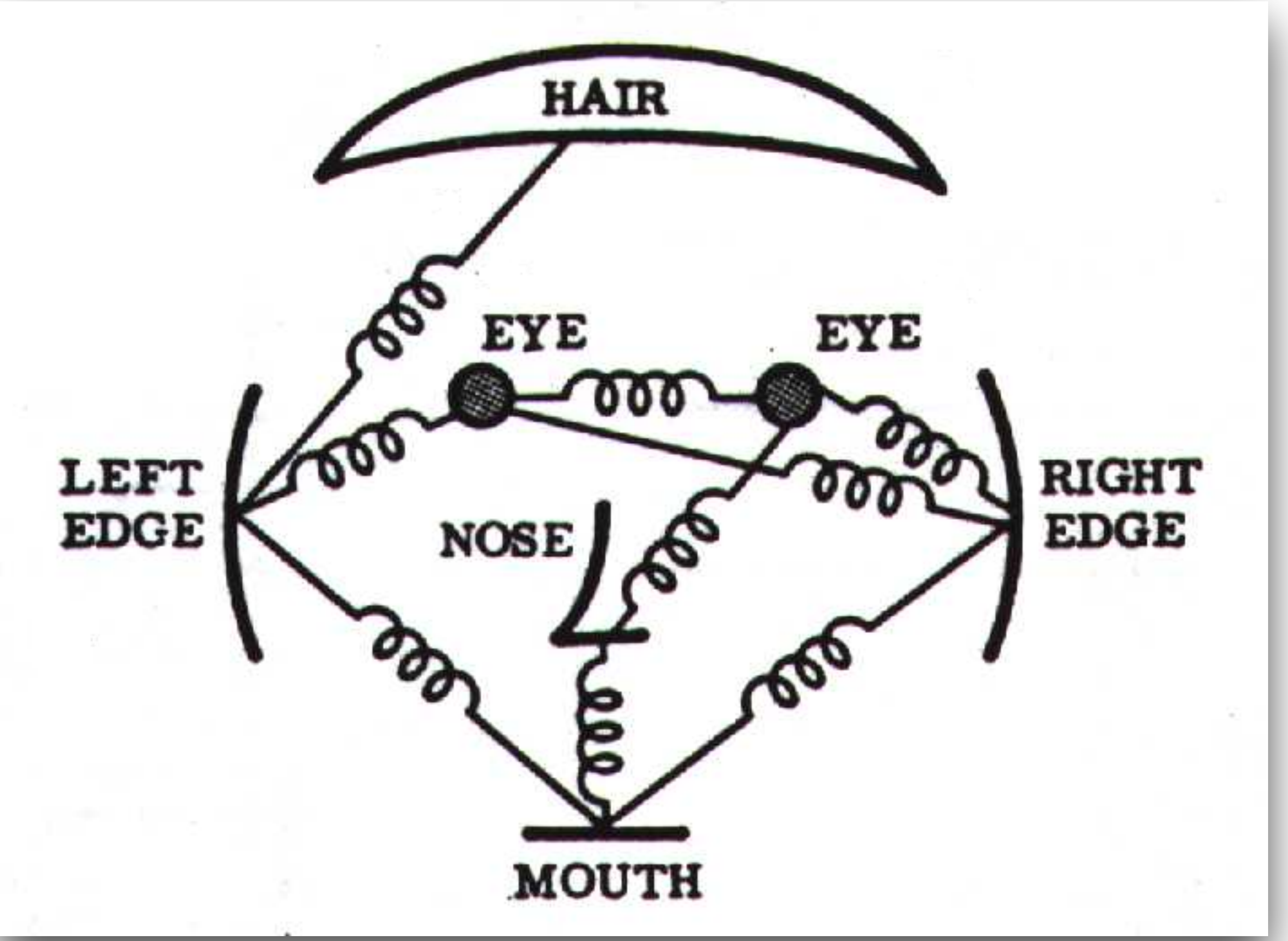


Recognition-by-Components: A Theory of Human Image Understanding. Psychological Review, 1987.

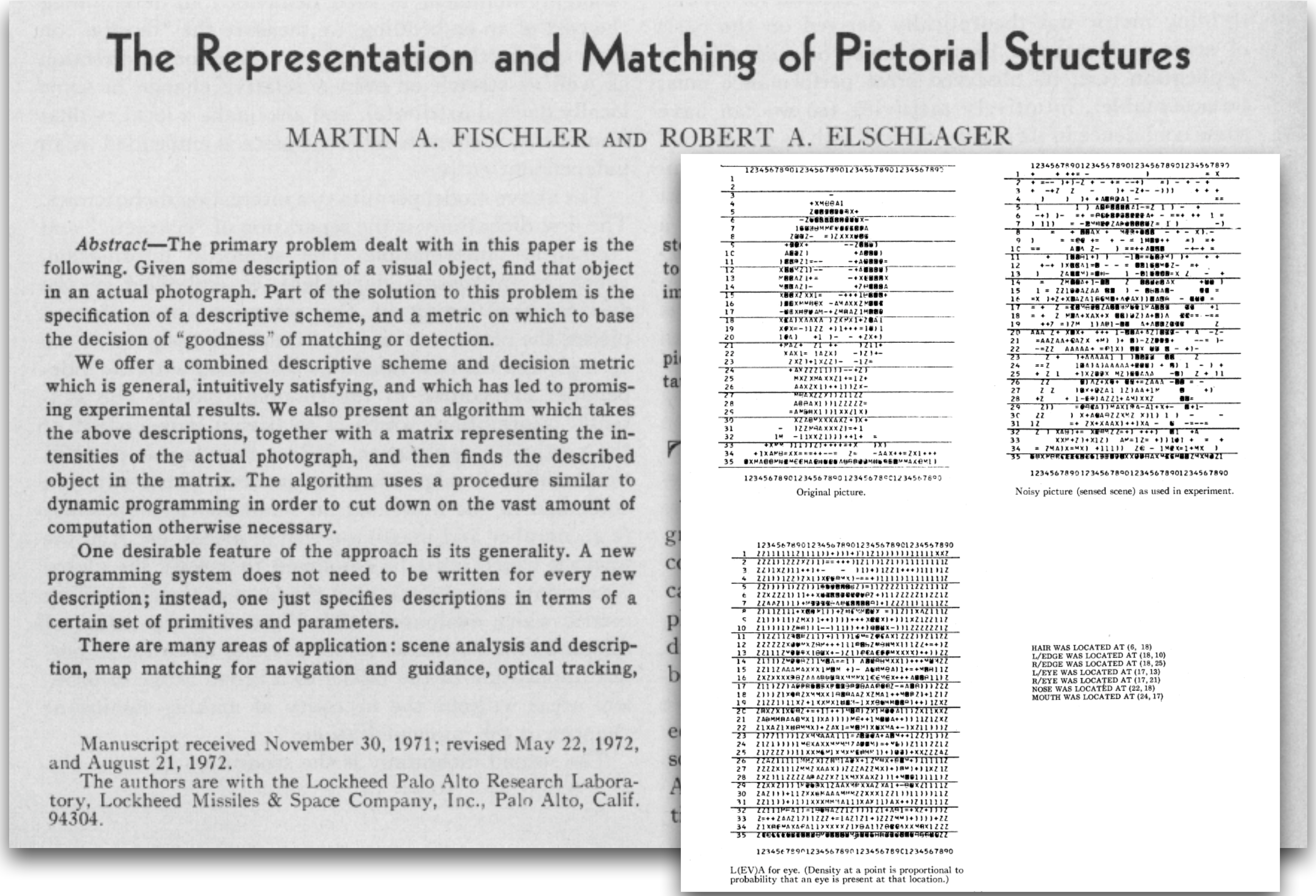
Irving Biederman

Source: Torralba, Freeman, Isola

Part based models



- Object as set of parts
  - Generative representation
- Model:
  - Relative locations between parts
  - Appearance of part
- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter



The Representation and Matching of Pictorial Structures

MARTIN A. FISCHLER AND ROBERT A. ELSCHLAGER

Abstract—The primary problem dealt with in this paper is the following. Given some description of a visual object, find that object in an actual photograph. Part of the solution to this problem is the specification of a descriptive scheme, and a metric on which to base the decision of "goodness" of matching or detection.

We offer a combined descriptive scheme and decision metric which is general, intuitively satisfying, and which has led to promising experimental results. We also present an algorithm which takes the above descriptions, together with a matrix representing the intensities of the actual photograph, and then finds the described object in the matrix. The algorithm uses a procedure similar to dynamic programming in order to cut down on the vast amount of computation otherwise necessary.

One desirable feature of the approach is its generality. A new programming system does not need to be written for every new description; instead, one just specifies descriptions in terms of a certain set of primitives and parameters.

There are many areas of application: scene analysis and description, map matching for navigation and guidance, optical tracking,

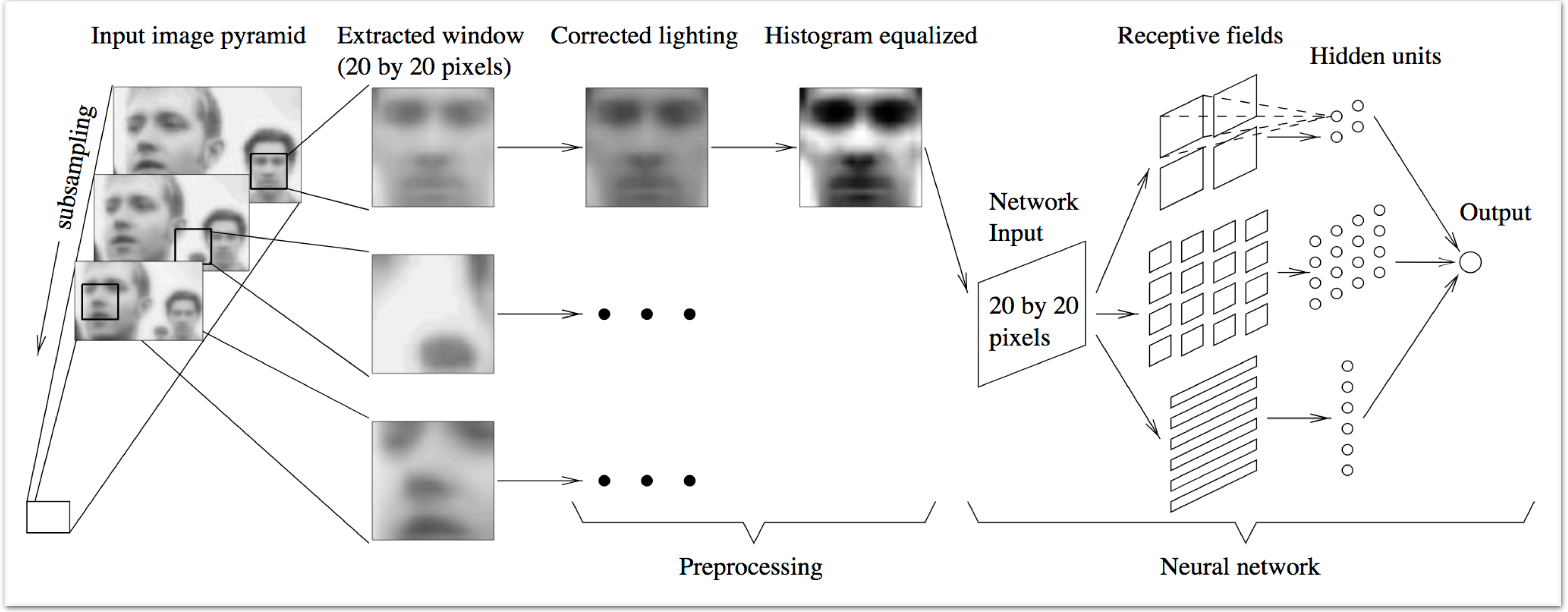
Manuscript received November 30, 1971; revised May 22, 1972, and August 21, 1972.

The authors are with the Lockheed Palo Alto Research Laboratory, Lockheed Missiles & Space Company, Inc., Palo Alto, Calif. 94304.

Source: Torralba, Freeman, Isola

Neural Network-Based Face Detector

Train a set of multilayer perceptrons and arbitrate a decision among all outputs

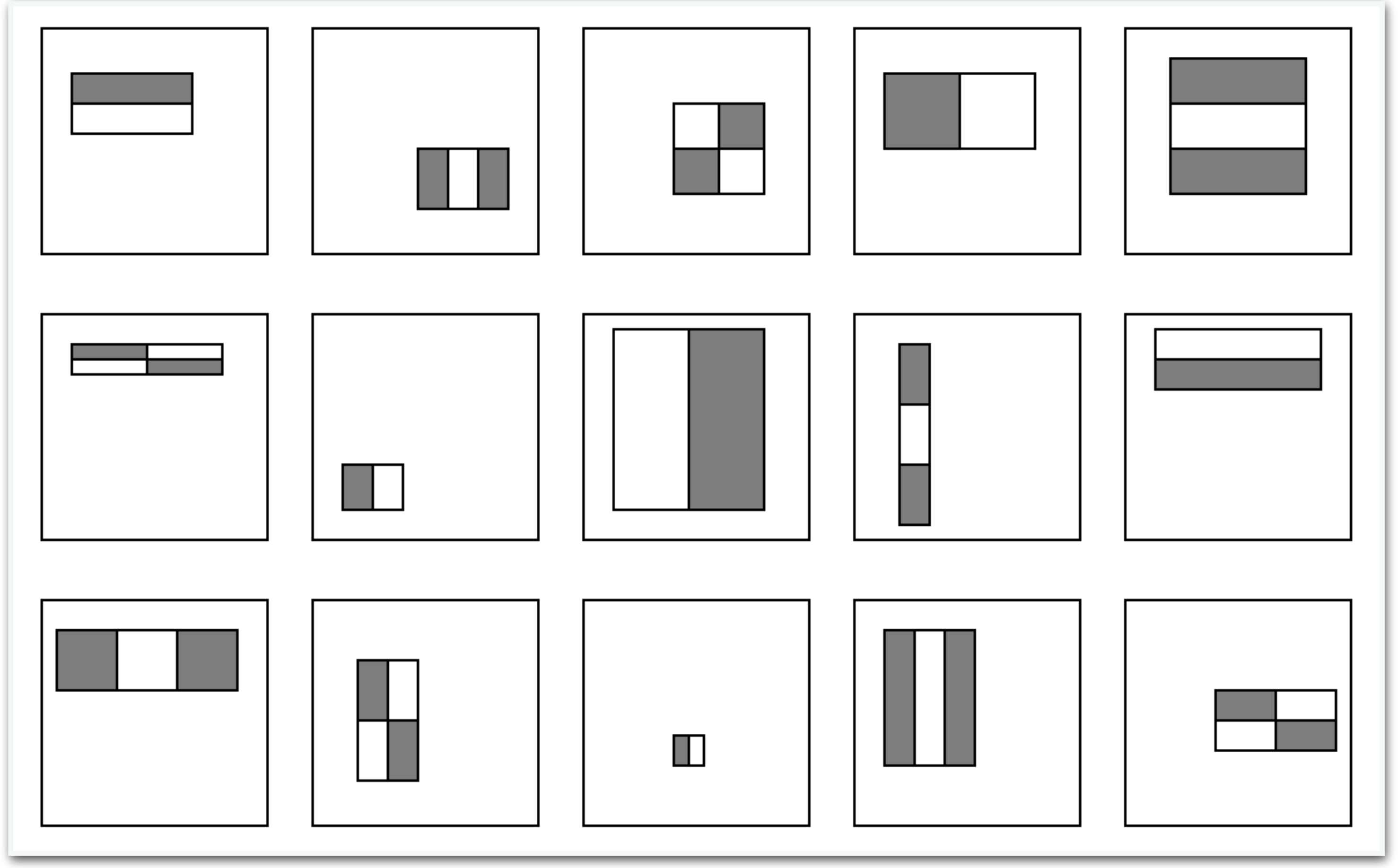


Rowley, Baluja, and Kanade: Neural Network-Based Face Detection (PAMI, January 1998)

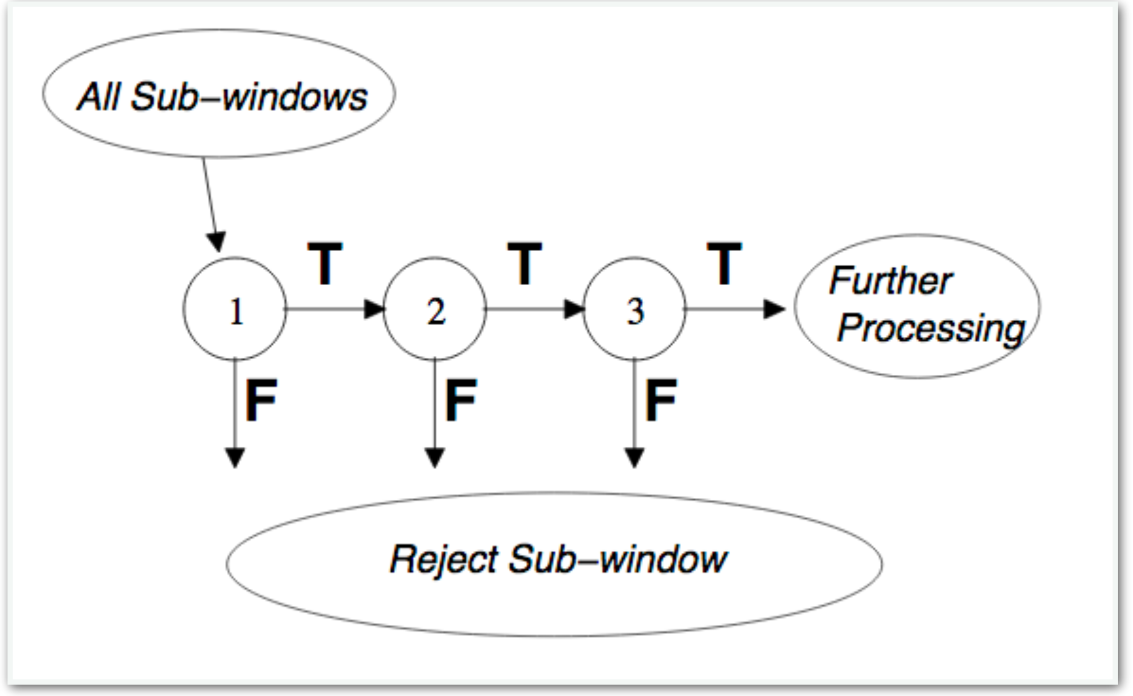
Source: Torralba, Freeman, Isola

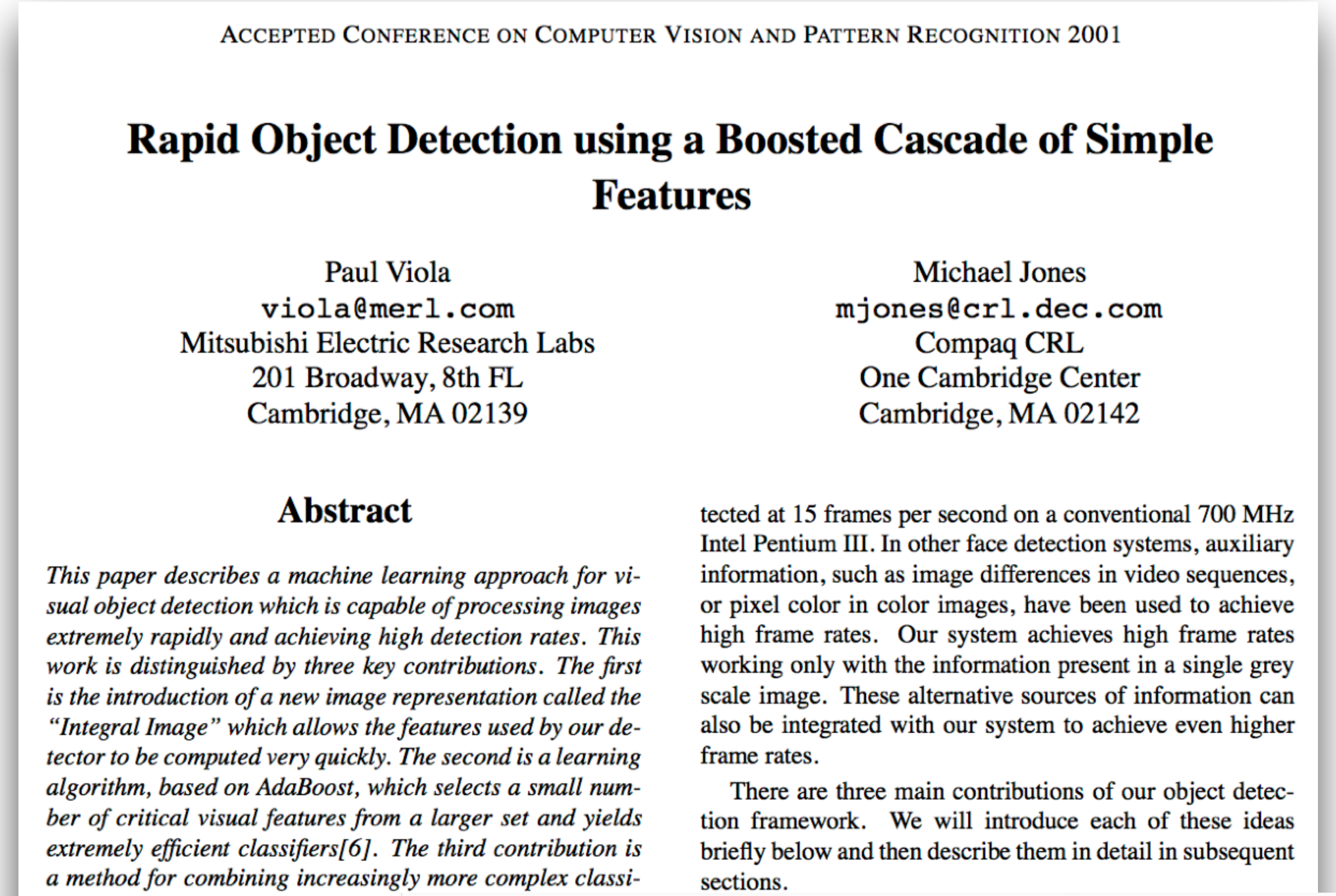
Viola-Jones algorithm

1. Millions of efficient features



2. Boosted feature selection
3. Computational cascade





ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001

Rapid Object Detection using a Boosted Cascade of Simple Features

Paul Viola
viola@merl.com
Mitsubishi Electric Research Labs
201 Broadway, 8th FL
Cambridge, MA 02139

Michael Jones
mjones@crl.dec.com
Compaq CRL
One Cambridge Center
Cambridge, MA 02142

Abstract

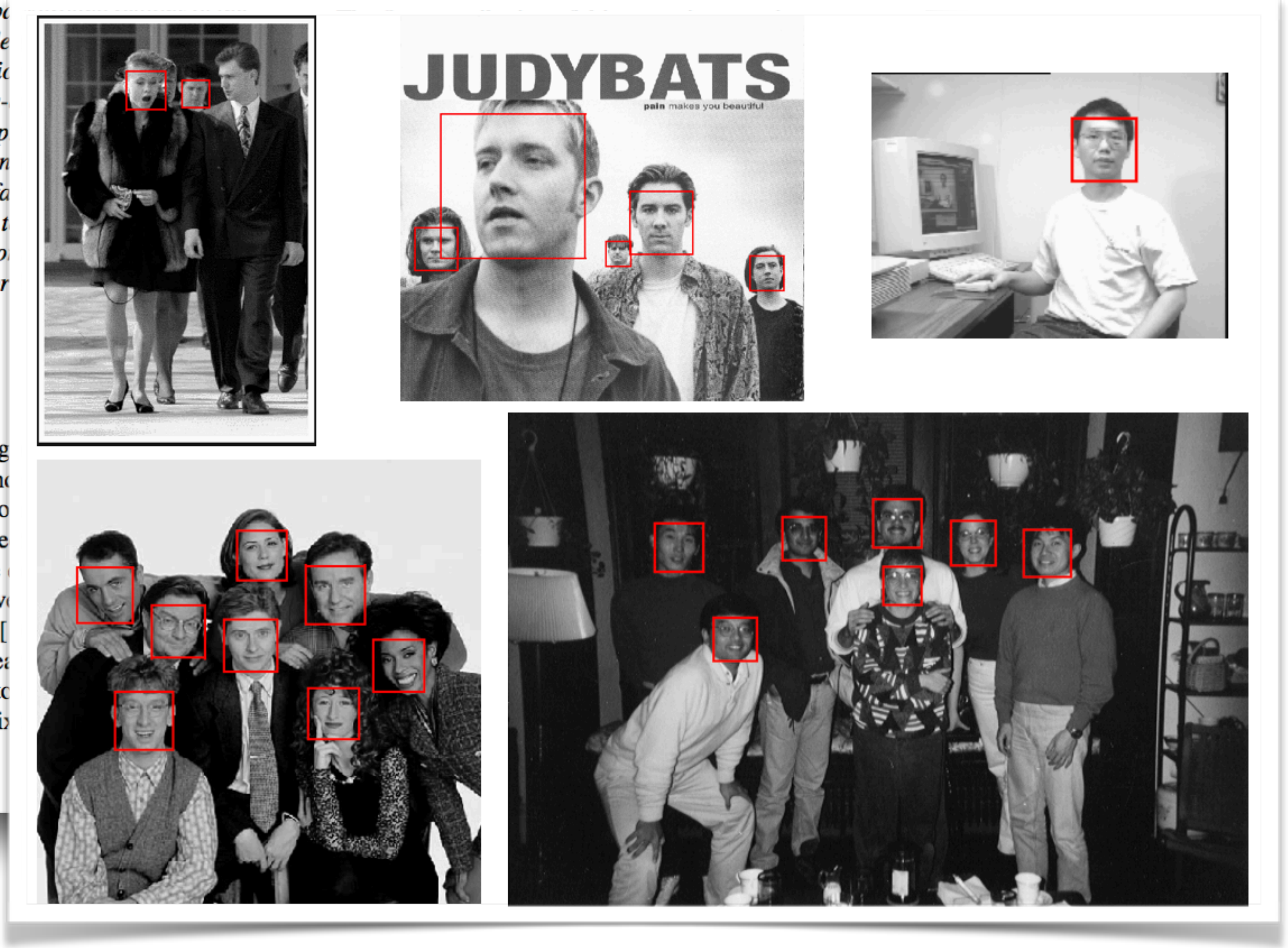
This paper describes a machine learning approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates. This work is distinguished by three key contributions. The first is the introduction of a new image representation called the "Integral Image" which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features from a larger set and yields extremely efficient classifiers[6]. The third contribution is a method for combining increasingly more complex classifiers in a "cascade" which allows b… image to be quickly discarded while… tation on promising object-like regi… viewed as an object specific focus-… which unlike previous approaches p… antees that discarded regions are un… ject of interest. In the domain of fa… yields detection rates comparable t… tems. Used in real-time applicatio… 15 frames per second without resor… ing or skin color detection.

1. Introduction

This paper brings together new alg… construct a framework for robust an… detection. This framework is demo… motivated by, the task of face dete… we have constructed a frontal face … achieves detection and false positiv… alent to the best published results [… face detection system is most clea… previous approaches in its ability t… rapidly. Operating on 384 by 288 pi…

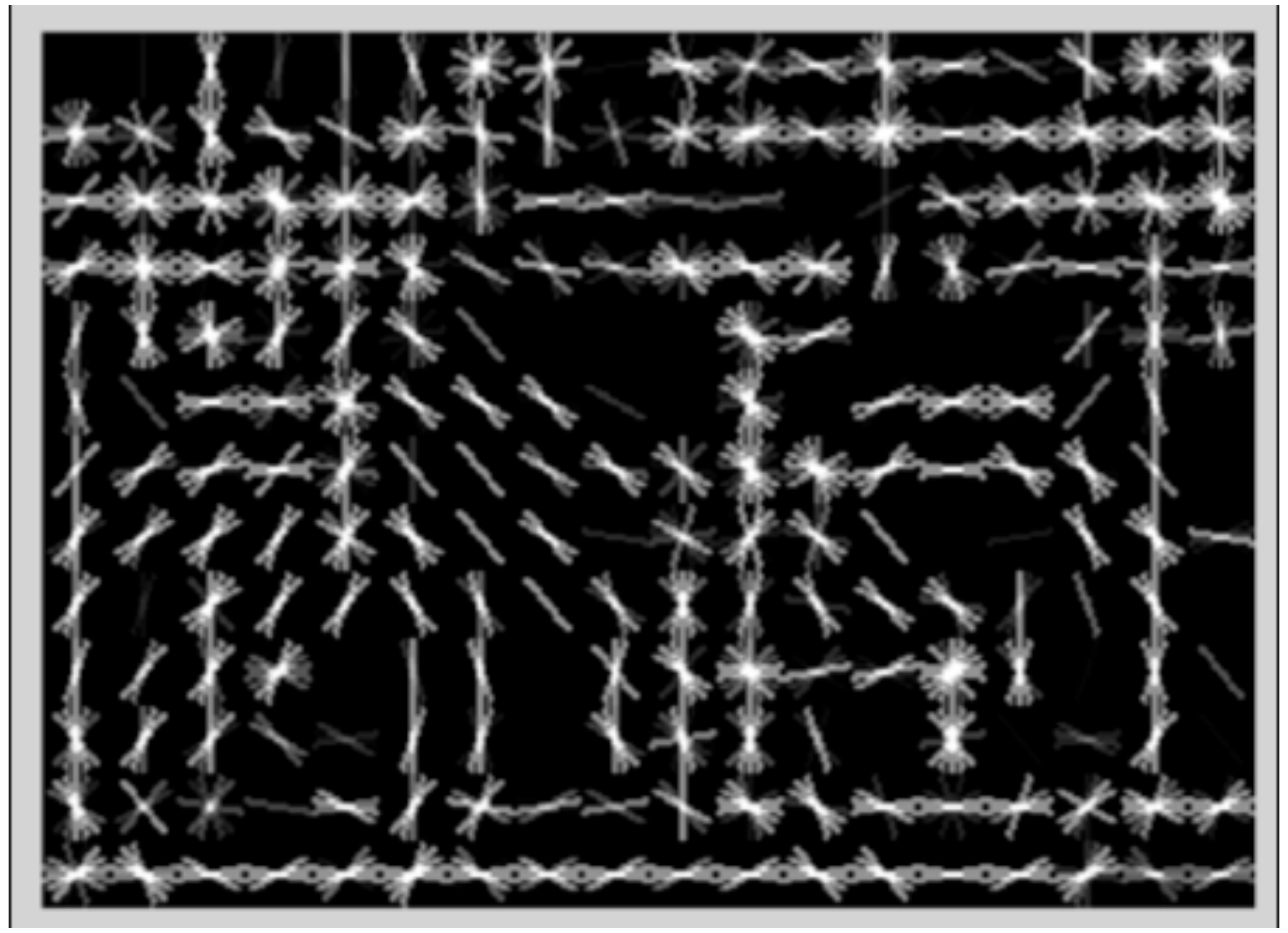
tected at 15 frames per second on a conventional 700 MHz Intel Pentium III. In other face detection systems, auxiliary information, such as image differences in video sequences, or pixel color in color images, have been used to achieve high frame rates. Our system achieves high frame rates working only with the information present in a single grey scale image. These alternative sources of information can also be integrated with our system to achieve even higher frame rates.

There are three main contributions of our object detection framework. We will introduce each of these ideas briefly below and then describe them in detail in subsequent sections.



Source: Torralba, Freeman, Isola

Histograms of oriented gradients (HOG)



Histograms of Oriented Gradients for Human Detection

Navneet Dalal and Bill Triggs
INRIA Rhône-Alps, 655 avenue de l'Europe, Montbonnot 38334, France
{Navneet.Dalal,Bill.Triggs}@inrialpes.fr, http://lear.inrialpes.fr

Abstract

We study the question of feature sets for robust visual object recognition, adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

1 Introduction

Detecting humans in images is a challenging task owing to their variable appearance and the wide range of poses that they can adopt. The first need is a robust feature set that allows the human form to be discriminated cleanly, even in cluttered backgrounds under difficult illumination. We study the issue of feature sets for human detection, showing that locally normalized Histogram of Oriented Gradient (HOG) descriptors provide excellent performance relative to other existing feature sets including wavelets [17,22]. The proposed descriptors are reminiscent of edge orientation histograms [4,5], SIFT descriptors [12] and shape contexts [1], but they are computed on a dense grid of uniformly spaced cells and they use overlapping local contrast normalizations for improved performance. We make a detailed study of the effects of various implementation choices on detector performance, taking "pedestrian detection" (the detection of mostly visible people in more or less upright poses) as a test case. For simplicity and speed, we use linear SVM as a baseline classifier throughout the study. The new detectors give essentially perfect results on the MIT pedestrian test set [18,17], so we have created a more challenging set containing over 1800 pedestrian images with a large range of poses and backgrounds. Ongoing work suggests that our feature set performs equally well for other shape-based object classes.

We briefly discuss previous work on human detection in §2, give an overview of our method §3, describe our data sets in §4 and give a detailed description and experimental evaluation of each stage of the process in §5–6. The main conclusions are summarized in §7.

2 Previous Work

There is an extensive literature on object detection, but here we mention just a few relevant papers on human detection [18,17,22,16,20]. See [6] for a survey. Papageorgiou *et al* [18] describe a pedestrian detector based on a polynomial SVM using rectified Haar wavelets as input descriptors, with a parts (subwindow) based variant in [17]. Depoortere *et al* give an optimized version of this [2]. Gavrila & Philomen [8] take a more direct approach, extracting edge images and matching them to a set of learned exemplars using chamfer distance. This has been used in a practical real-time pedestrian detection system [7]. Viola *et al* [22] build an efficient moving person detector, using AdaBoost to train a chain of progressively more complex region rejection rules based on Haar-like wavelets and space-time differences. Ronfard *et al* [19] build an articulated body detector by incorporating SVM based limb classifiers over 1st and 2nd order Gaussian filters in a dynamic programming framework similar to those of Felzenszwalb & Huttenlocher [3] and Ioffe & Forsyth [9]. Mikolajczyk *et al* [16] use combinations of orientation-position histograms with binary-thresholded gradient magnitudes to build a parts based method containing detectors for faces, heads, and front and side profiles of upper and lower body parts. In contrast, our detector uses a simpler architecture with a single detection window, but appears to give significantly higher performance on pedestrian images.

3 Overview of the Method

This section gives an overview of our feature extraction chain, which is summarized in fig. 1. Implementation details are postponed until §6. The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. Similar features have seen increasing use over the past decade [4,5,12,15]. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or

1

https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

1. Bin gradients from 8x8 pixel neighborhoods into 9 orientations
2. Linear SVM

Source: Torralba, Freeman, Isola

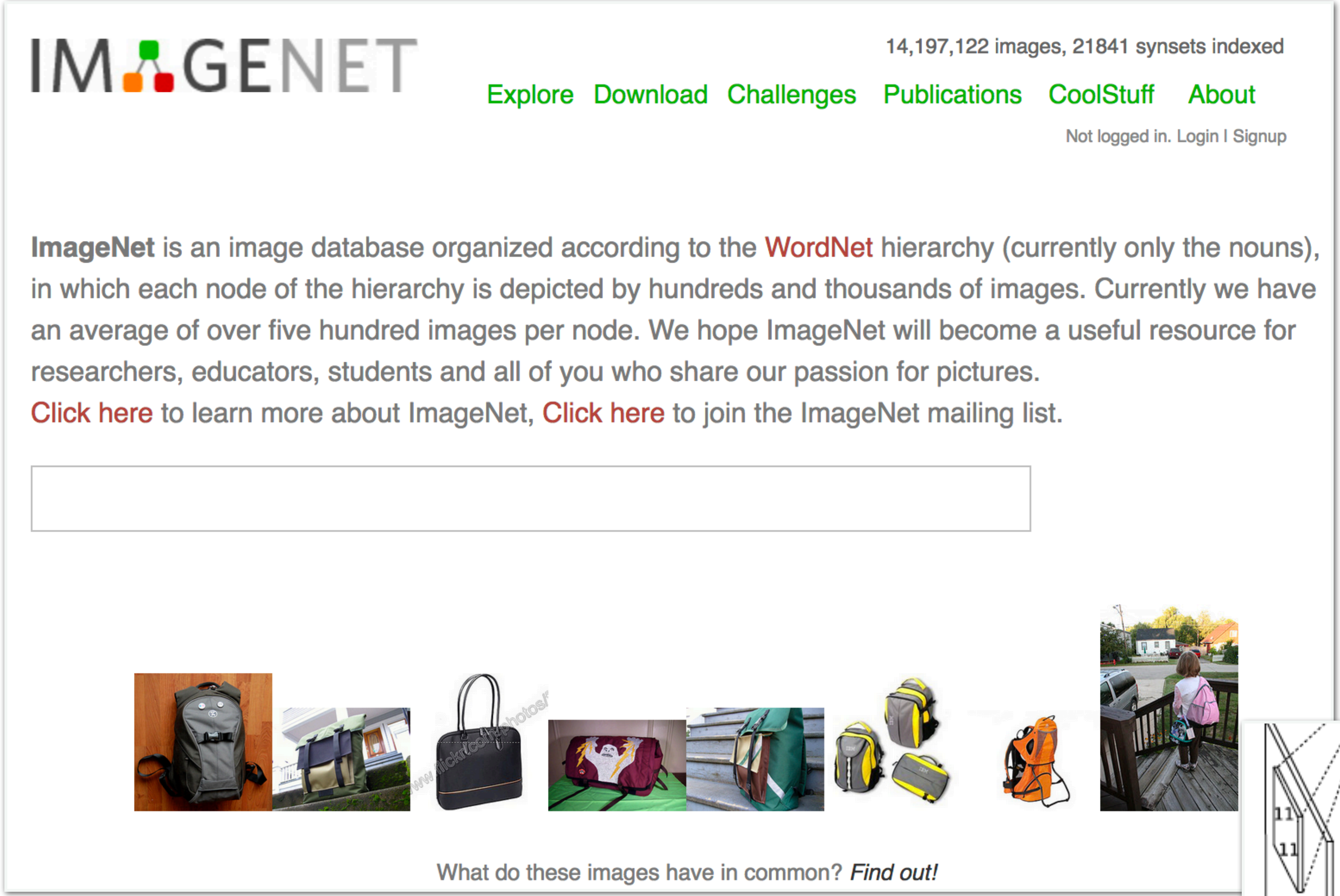
["HOGgles", Vondrick et al. , ICCV 2013]

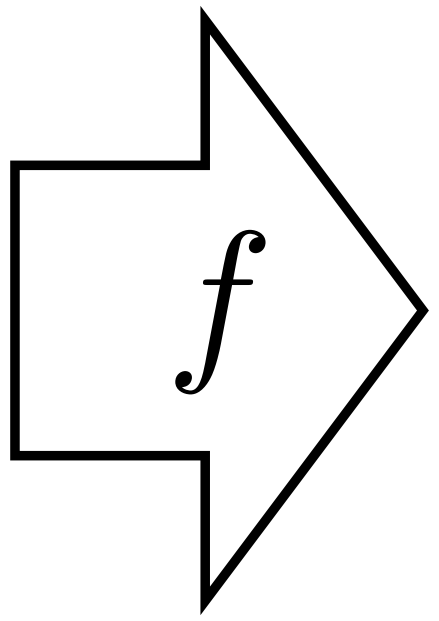
Source: Torralba, Freeman, Isola

["HOGgles", Vondrick et al. , ICCV 2013]

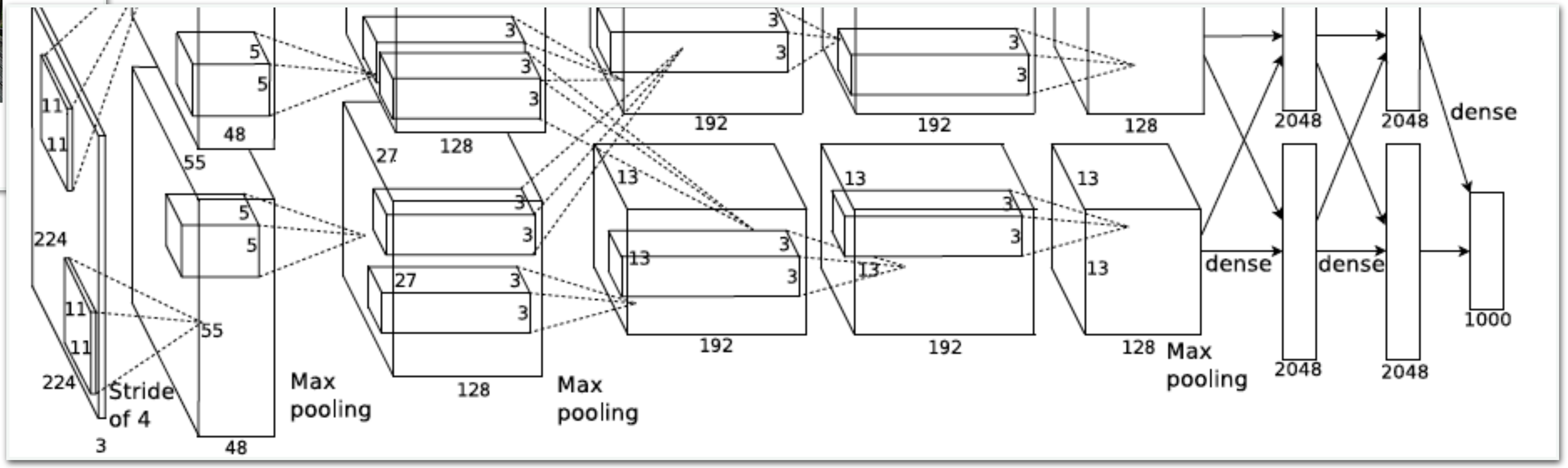
Source: Torralba, Freeman, Isola

ImageNet classification and Neural nets





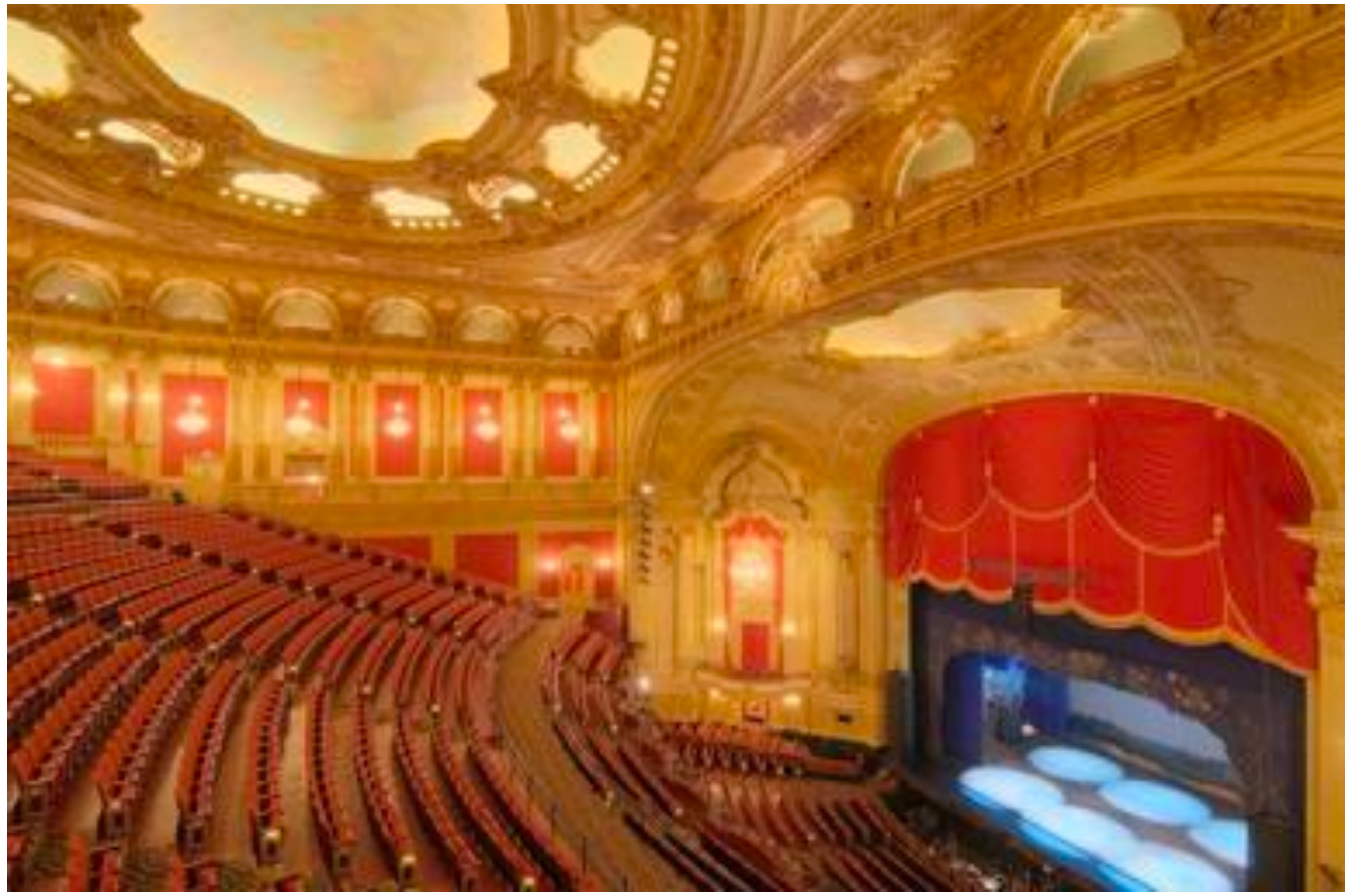
"Birds"



Source: Torralba, Freeman, Isola

Scene recognition

Scene recognition problem



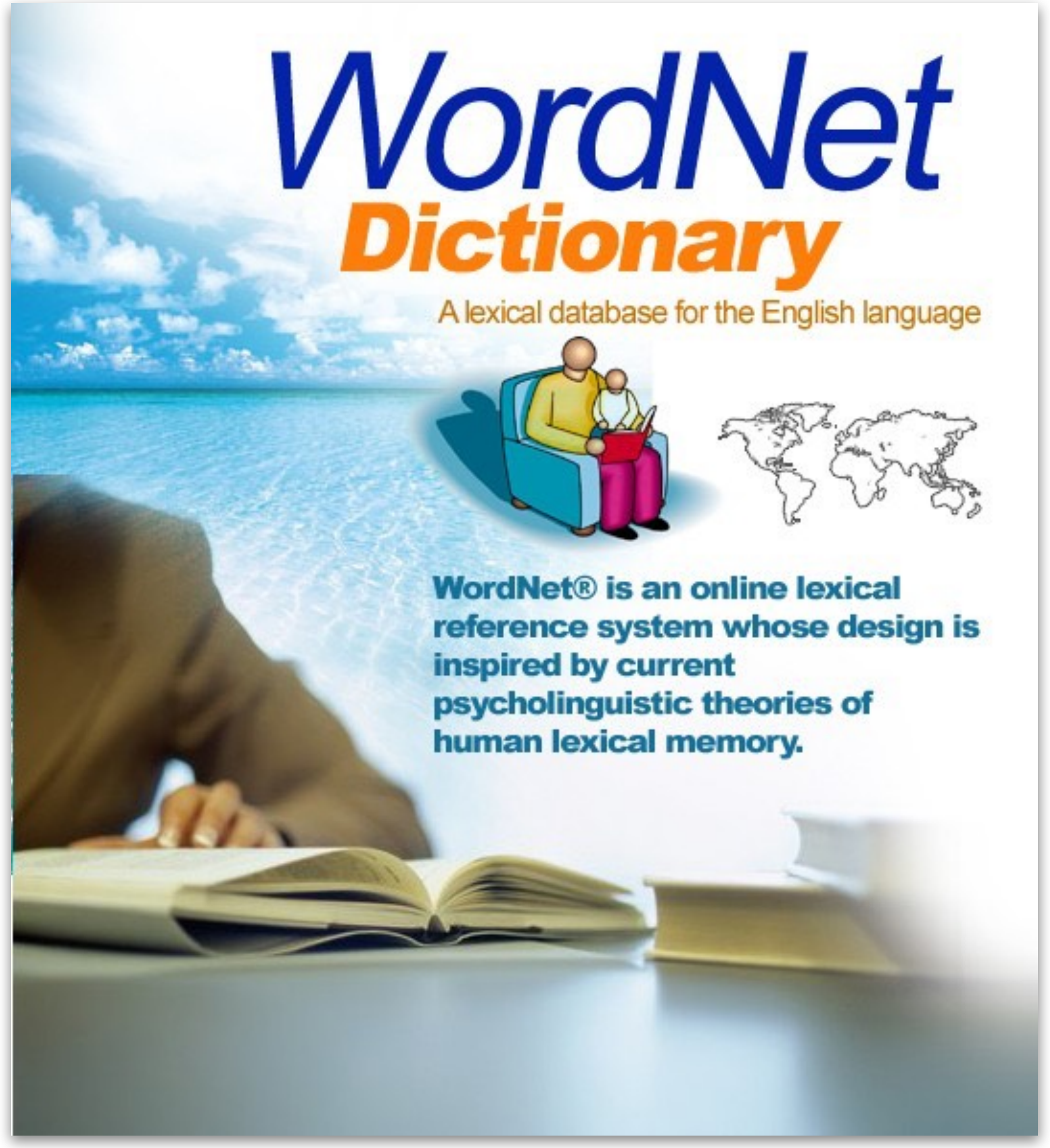
f

“Auditorium”

Your next problem set!

places

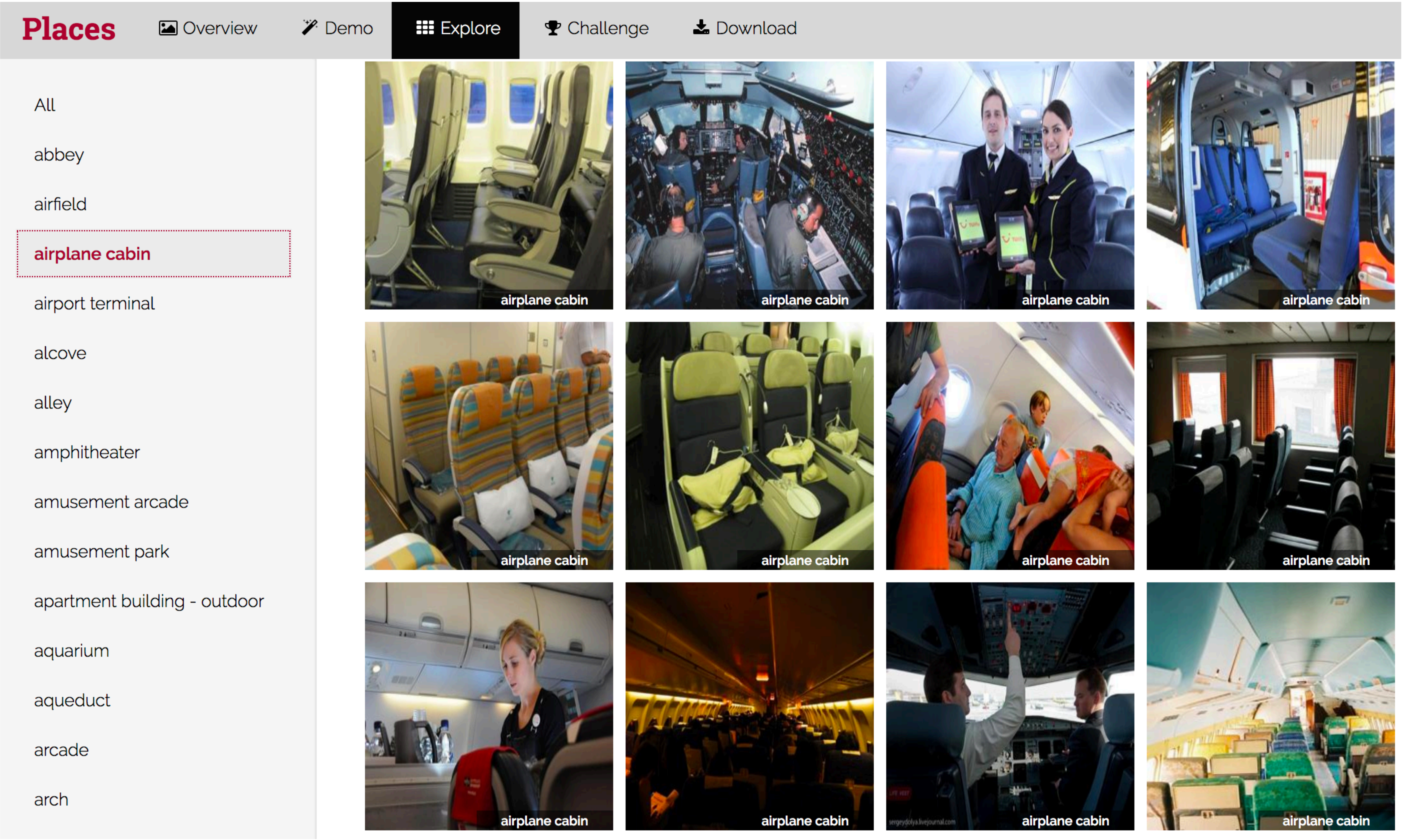
1. Take all scene words from a dictionary



2. Download images and clean the categories

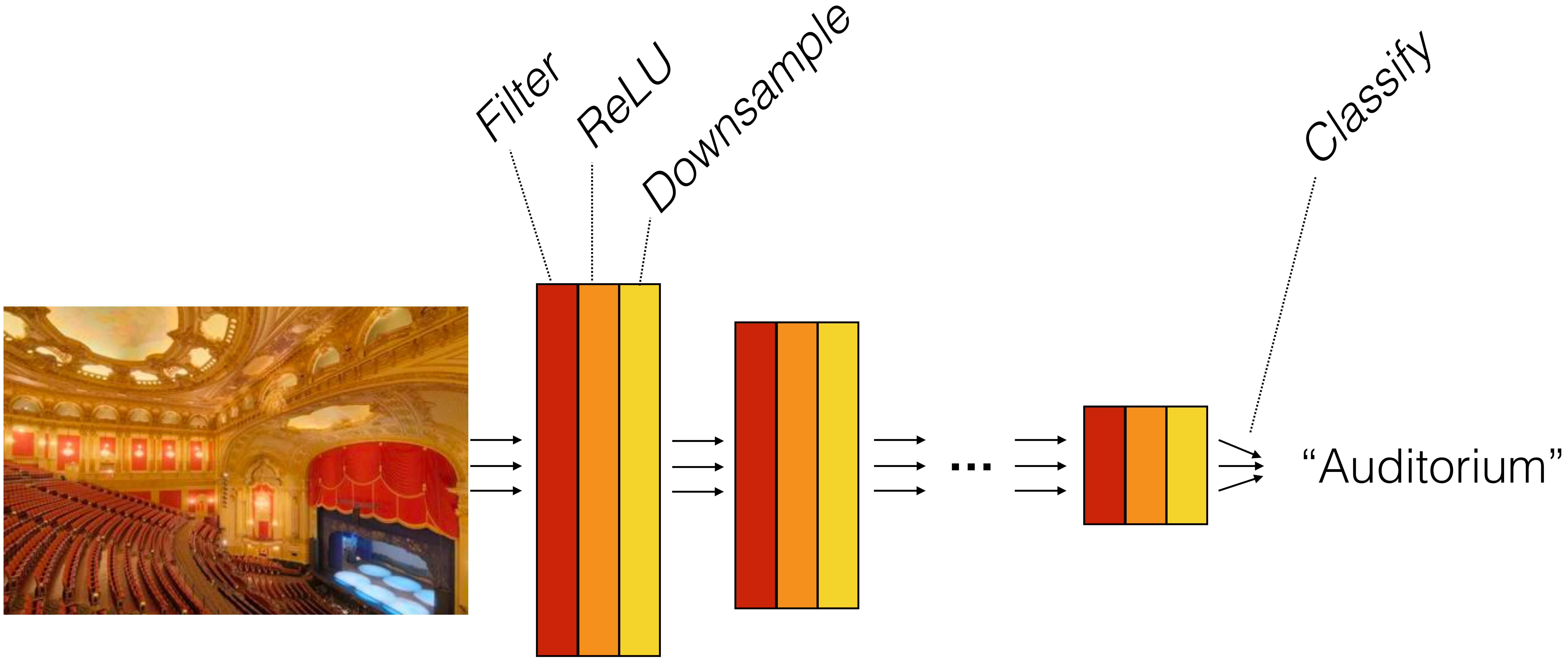
Source: A. Oliva

Zhou, Lapedriza, Xiao, Oliva, Torralba (NIPS 2014)



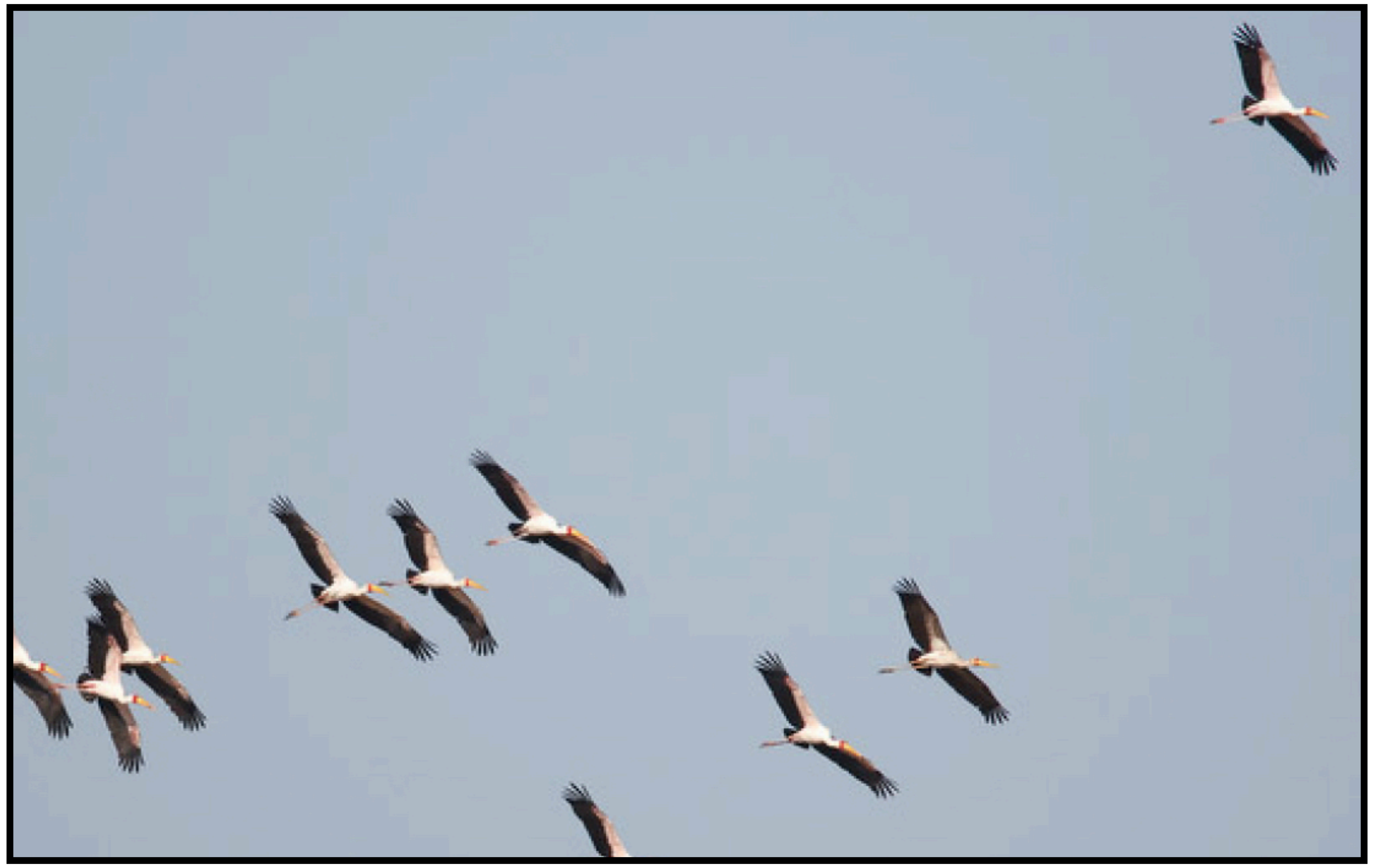
Source: Torralba, Freeman, Isola

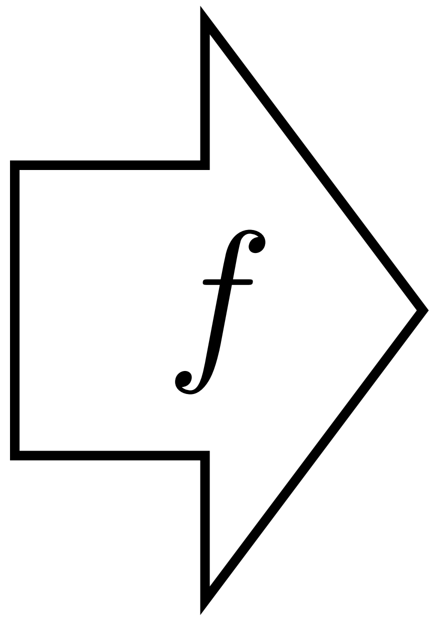
Scene recognition with CNN



PS5: implement this in PyTorch

Object recognition: what objects are in the image?

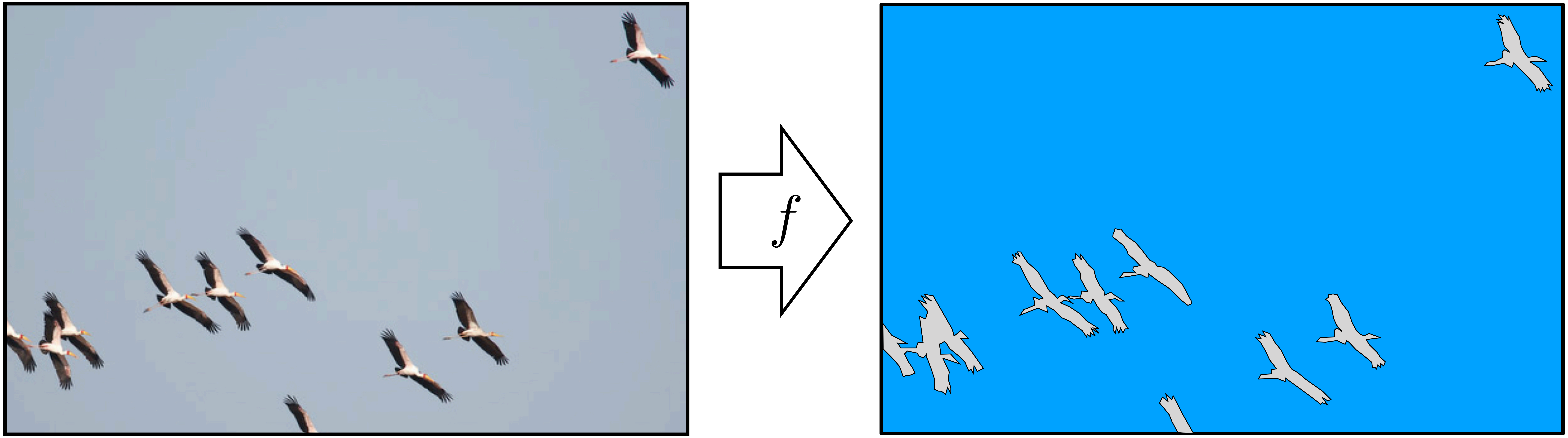




“Birds”

Source: Torralba, Freeman, Isola

Semantic segmentation



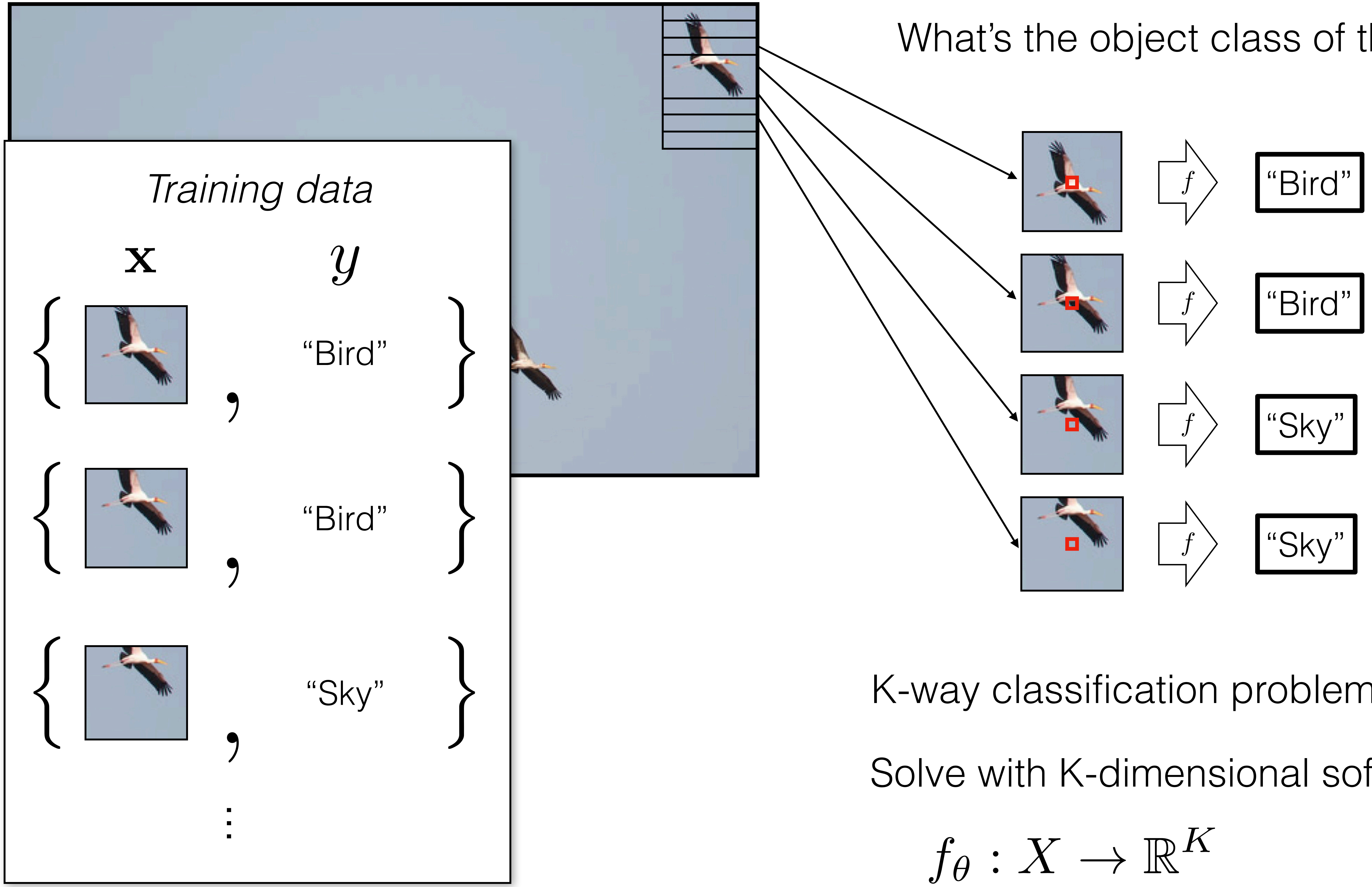
(Colors represent categories)

General technique: predict something at every pixel!

Source: Torralba, Freeman, Isola

Idea #1: Independently classify windows

What's the object class of the center pixel?



K-way classification problem

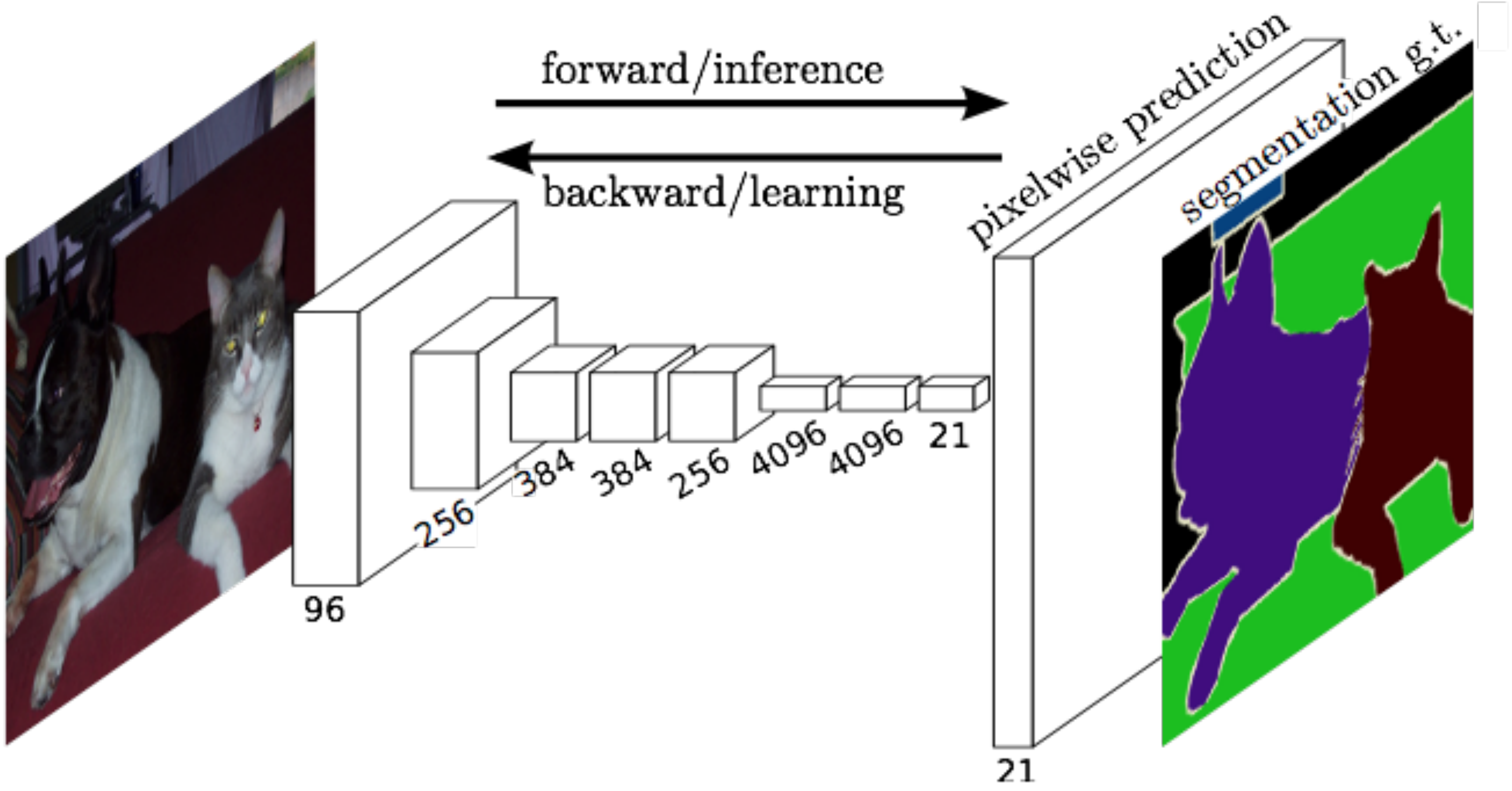
Solve with K-dimensional softmax regression:

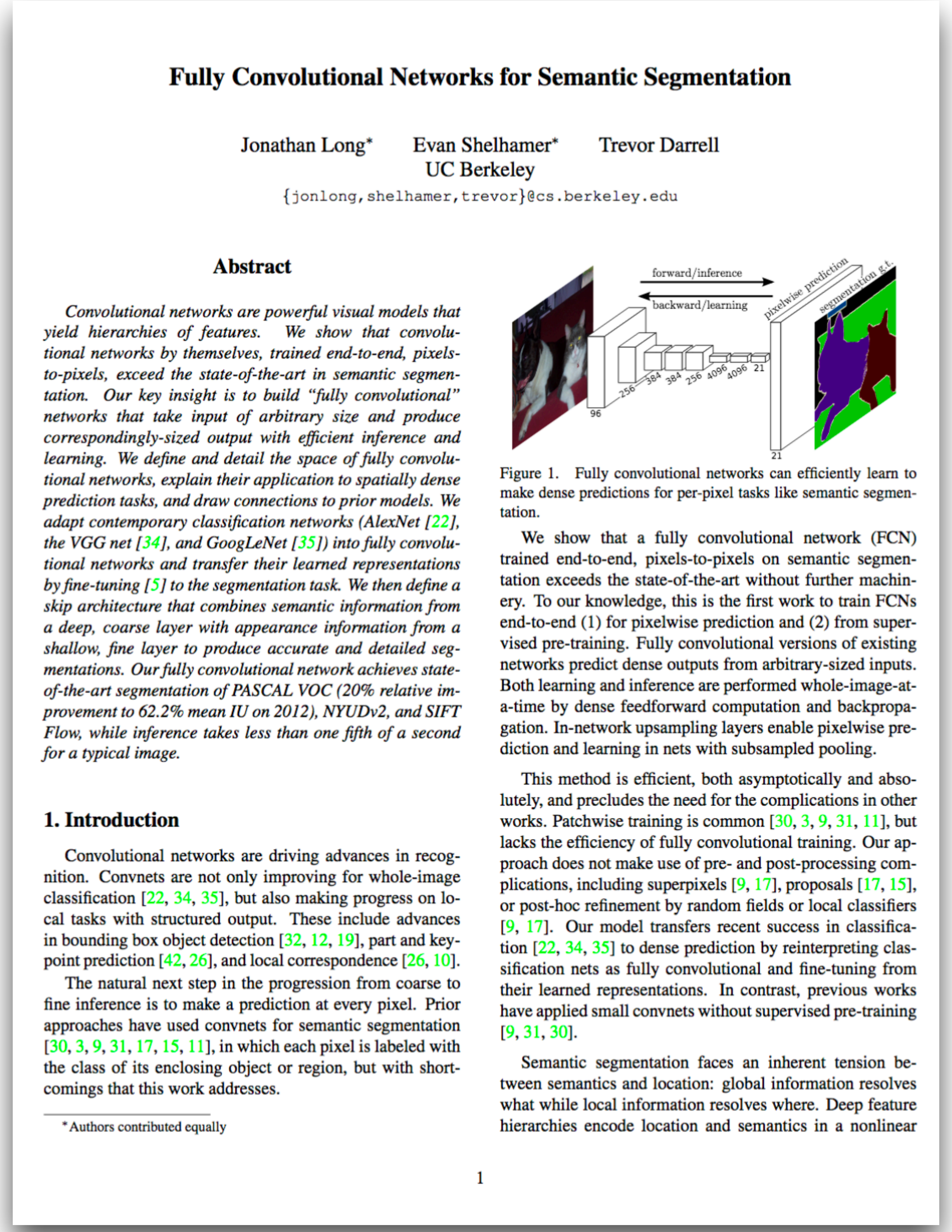
$$f_\theta : X \to \mathbb{R}^K$$

Source: Torralba, Freeman, Isola

Idea #2: Fully convolutional networks

Fully Convolutional Networks





Fully Convolutional Networks for Semantic Segmentation

Jonathan Long* Evan Shelhamer* Trevor Darrell
UC Berkeley
{jonlong,shelhamer,trevor}@cs.berkeley.edu

Abstract

Convolutional networks are powerful visual models that yield hierarchies of features. We show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, exceed the state-of-the-art in semantic segmentation. Our key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. We define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. We adapt contemporary classification networks (AlexNet [22], the VGG net [34], and GoogLeNet [35]) into fully convolutional networks and transfer their learned representations by fine-tuning [5] to the segmentation task. We then define a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Our fully convolutional network achieves state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes less than one fifth of a second for a typical image.

1. Introduction

Convolutional networks are driving advances in recognition. Convnets are not only improving for whole-image classification [22, 34, 35], but also making progress on local tasks with structured output. These include advances in bounding box object detection [32, 12, 19], part and keypoint prediction [42, 26], and local correspondence [26, 10].

The natural next step in the progression from coarse to fine inference is to make a prediction at every pixel. Prior approaches have used convnets for semantic segmentation [30, 3, 9, 31, 17, 15, 11], in which each pixel is labeled with the class of its enclosing object or region, but with shortcomings that this work addresses.

*Authors contributed equally

Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

We show that a fully convolutional network (FCN) trained end-to-end, pixels-to-pixels on semantic segmentation exceeds the state-of-the-art without further machinery. To our knowledge, this is the first work to train FCNs end-to-end (1) for pixelwise prediction and (2) from supervised pre-training. Fully convolutional versions of existing networks predict dense outputs from arbitrary-sized inputs. Both learning and inference are performed whole-image-at-a-time by dense feedforward computation and backpropagation. In-network upsampling layers enable pixelwise prediction and learning in nets with subsampled pooling.

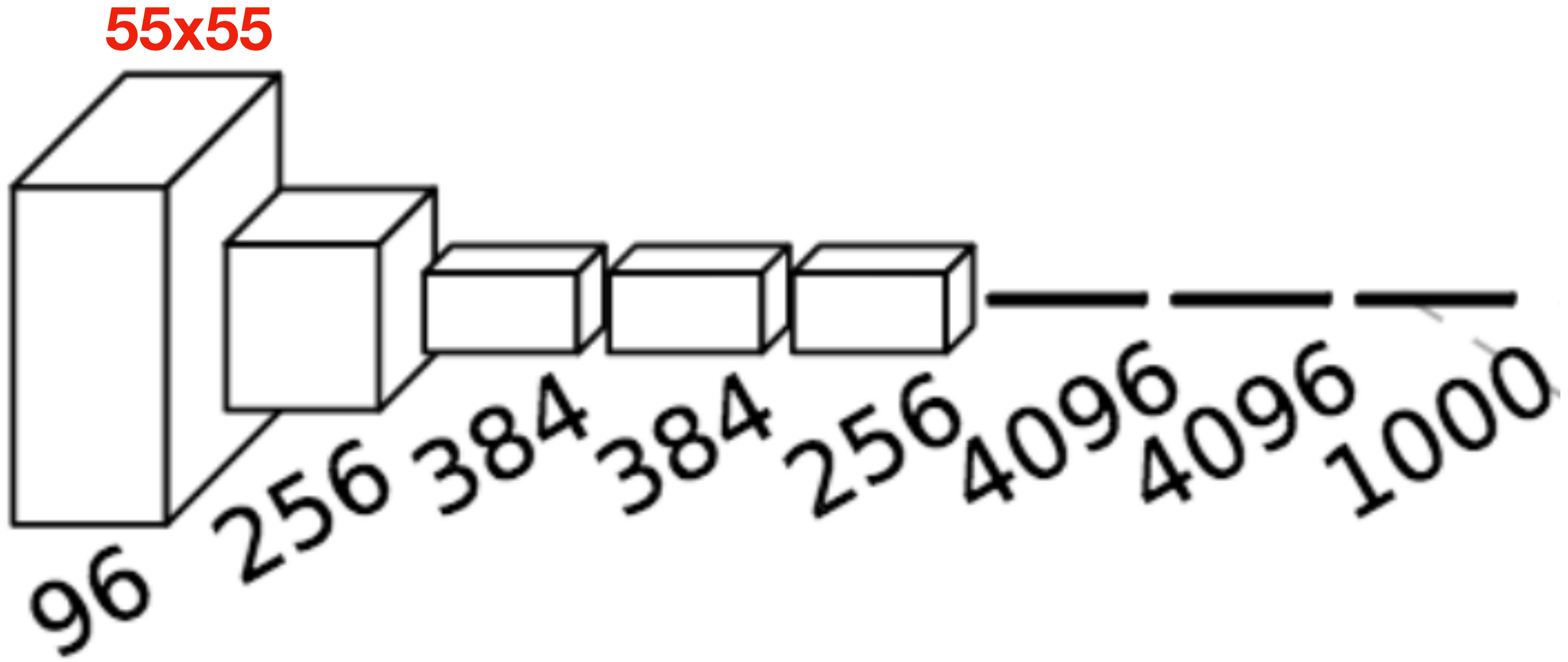
This method is efficient, both asymptotically and absolutely, and precludes the need for the complications in other works. Patchwise training is common [30, 3, 9, 31, 11], but lacks the efficiency of fully convolutional training. Our approach does not make use of pre- and post-processing complications, including superpixels [9, 17], proposals [17, 15], or post-hoc refinement by random fields or local classifiers [9, 17]. Our model transfers recent success in classification [22, 34, 35] to dense prediction by reinterpreting classification nets as fully convolutional and fine-tuning from their learned representations. In contrast, previous works have applied small convnets without supervised pre-training [9, 31, 30].

Semantic segmentation faces an inherent tension between semantics and location: global information resolves what while local information resolves where. Deep feature hierarchies encode location and semantics in a nonlinear

1

Source: Torralba, Freeman, Isola

Fully Convolutional Networks

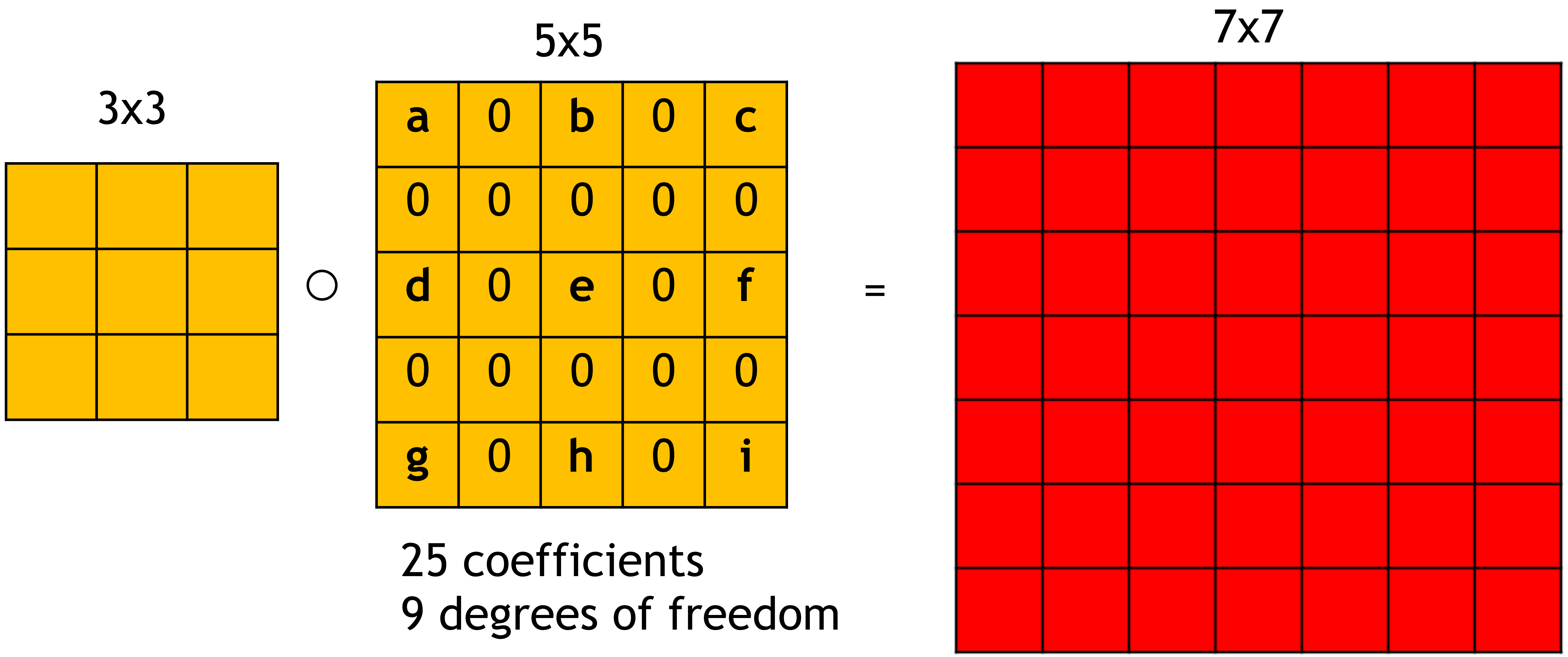


Fully Convolutional Networks

Reuse features across windows. Less computation!

Idea #3: Dilated convolutions

Dilated convolutions



25 coefficients
9 degrees of freedom

49 coefficients
18 degrees of freedom

[Yu and Koltun 2016, https://arxiv.org/pdf/1511.07122.pdf]

Source: Isola, Torralba, Freeman

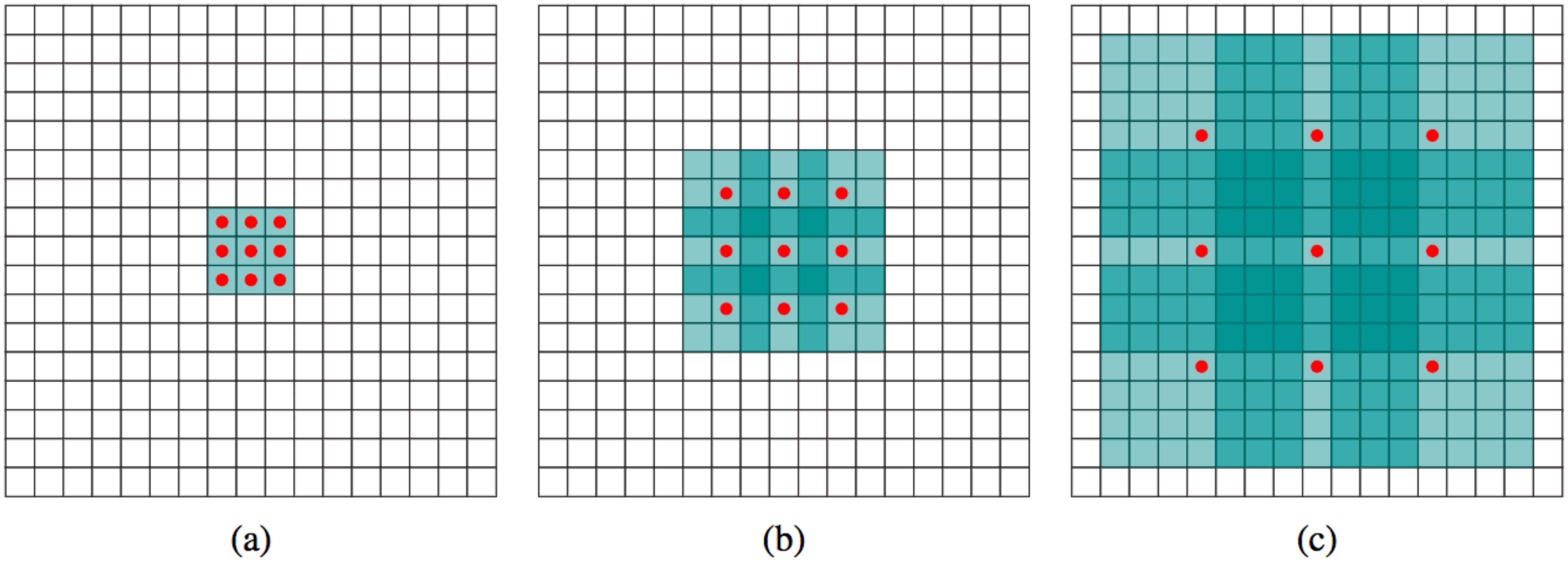
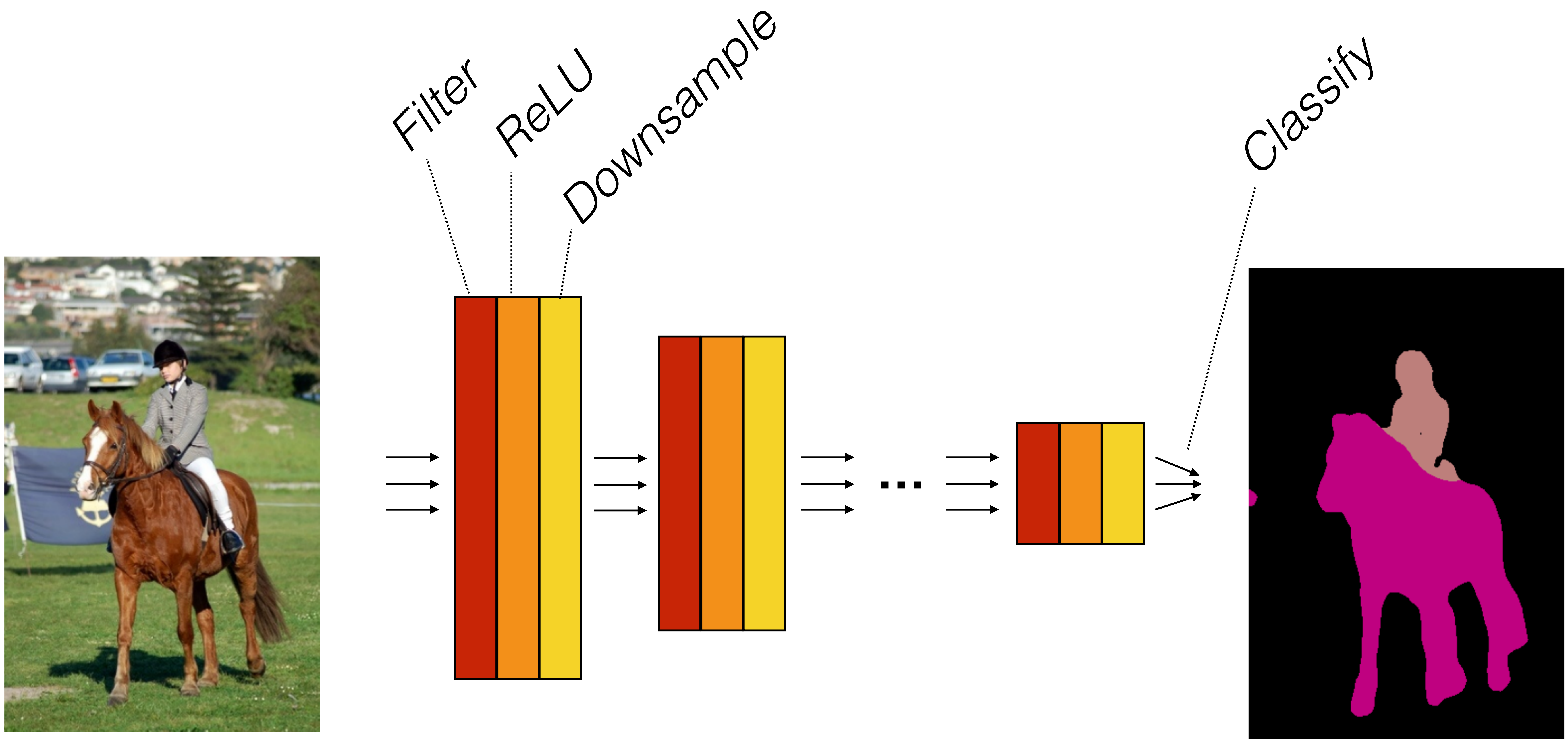


Figure 1: Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage. (a) F_1 is produced from F_0 by a 1-dilated convolution; each element in F_1 has a receptive field of 3×3. (b) F_2 is produced from F_1 by a 2-dilated convolution; each element in F_2 has a receptive field of 7×7. (c) F_3 is produced from F_2 by a 4-dilated convolution; each element in F_3 has a receptive field of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

[Yu and Koltun 2016, https://arxiv.org/pdf/1511.07122.pdf]

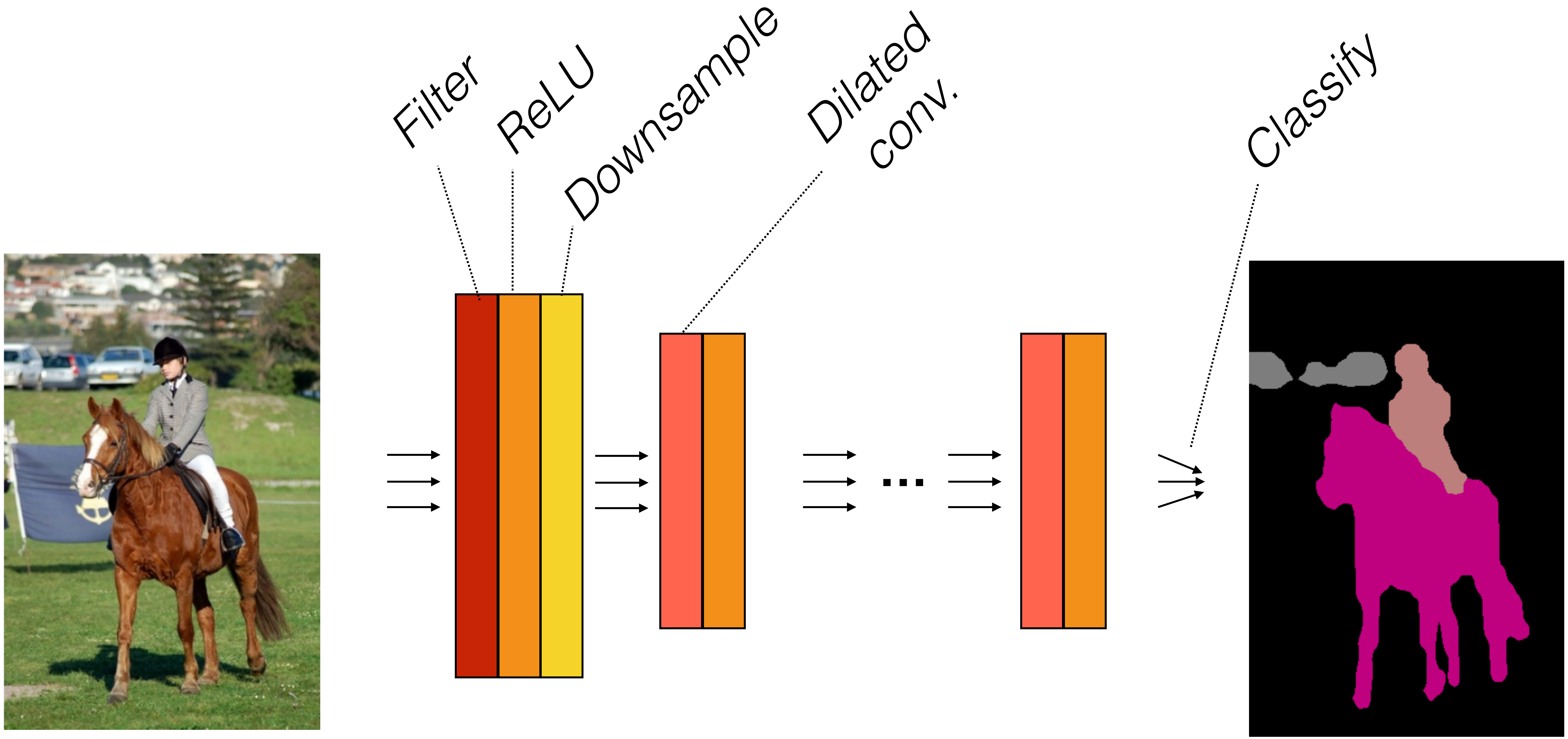
Source: Isola, Torralba, Freeman

Fully convolutional network



Apply CNN convolutionally

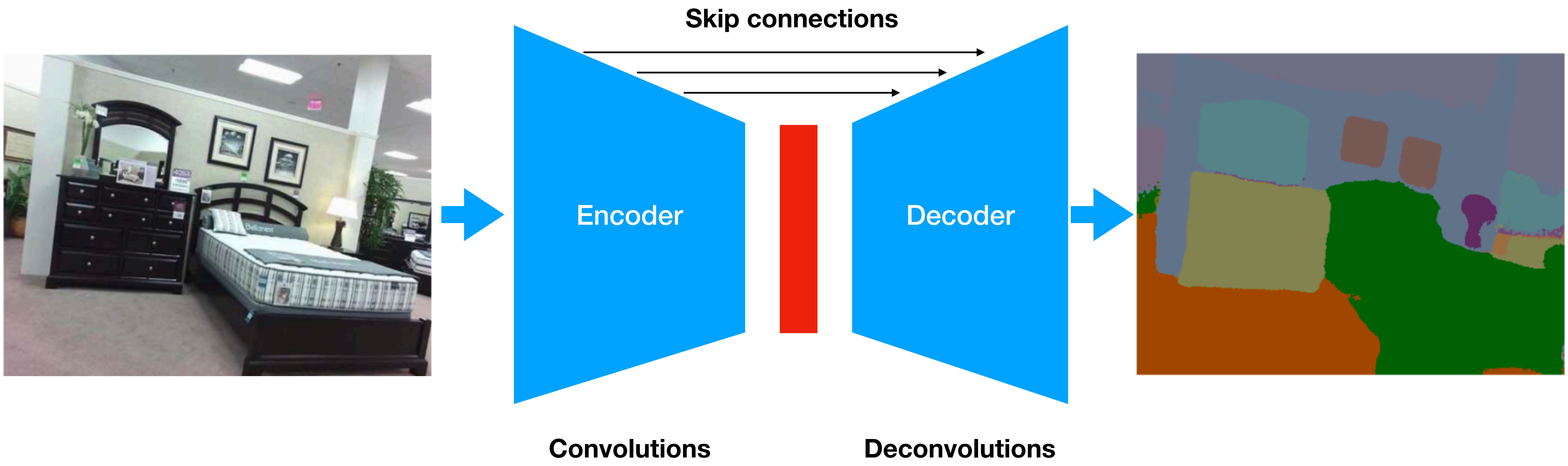
Fully convolutional network



Output still usually not given at full-resolution

Idea #4: Skip connections

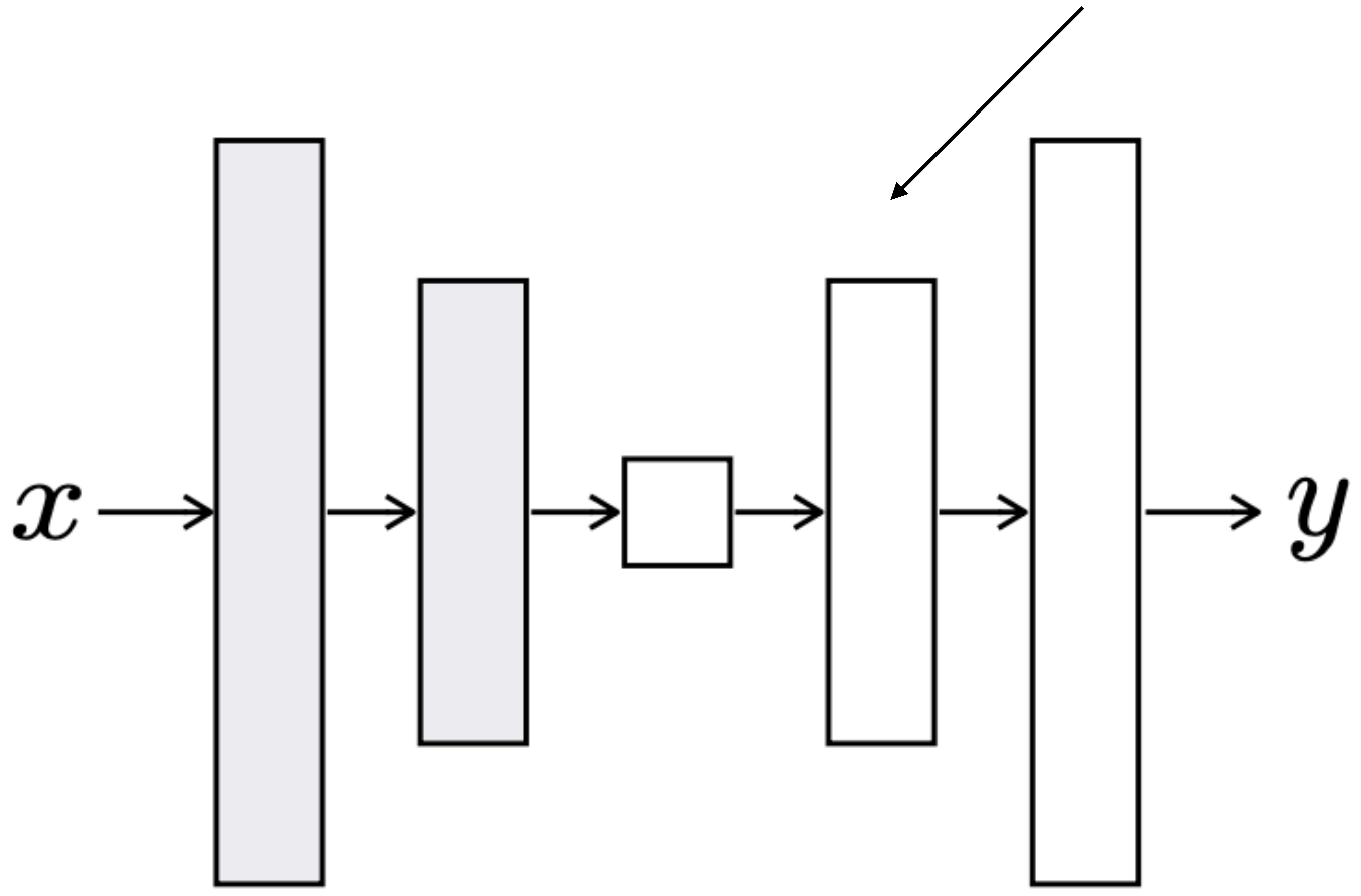
Encoder-decoder architectures



Source: Torralba, Freeman, Isola

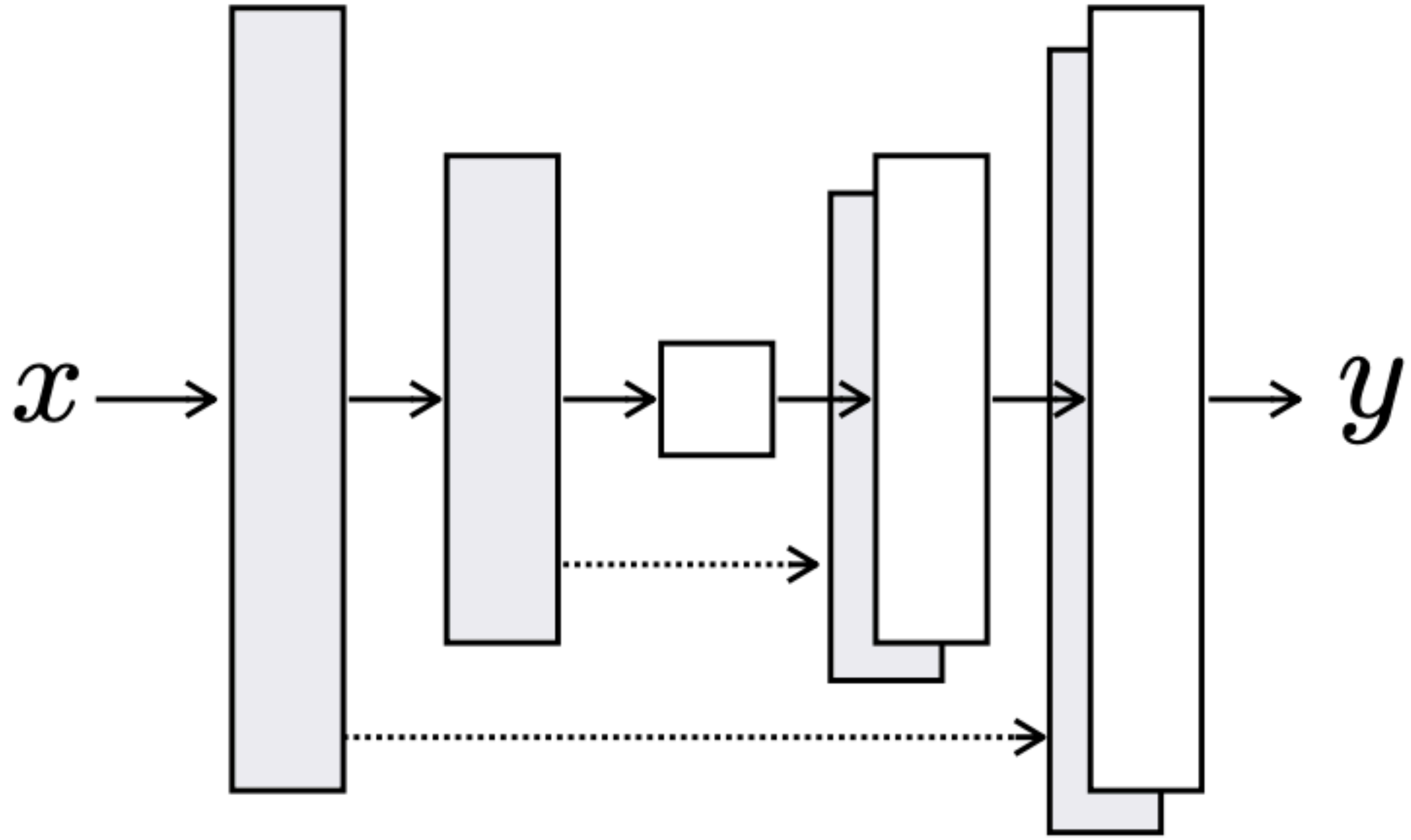
Encoder-decoder architectures

Deconvolution (a.k.a. “upconvolution” or “transposed convolution”)



“Vanilla” encoder-decoder architecture

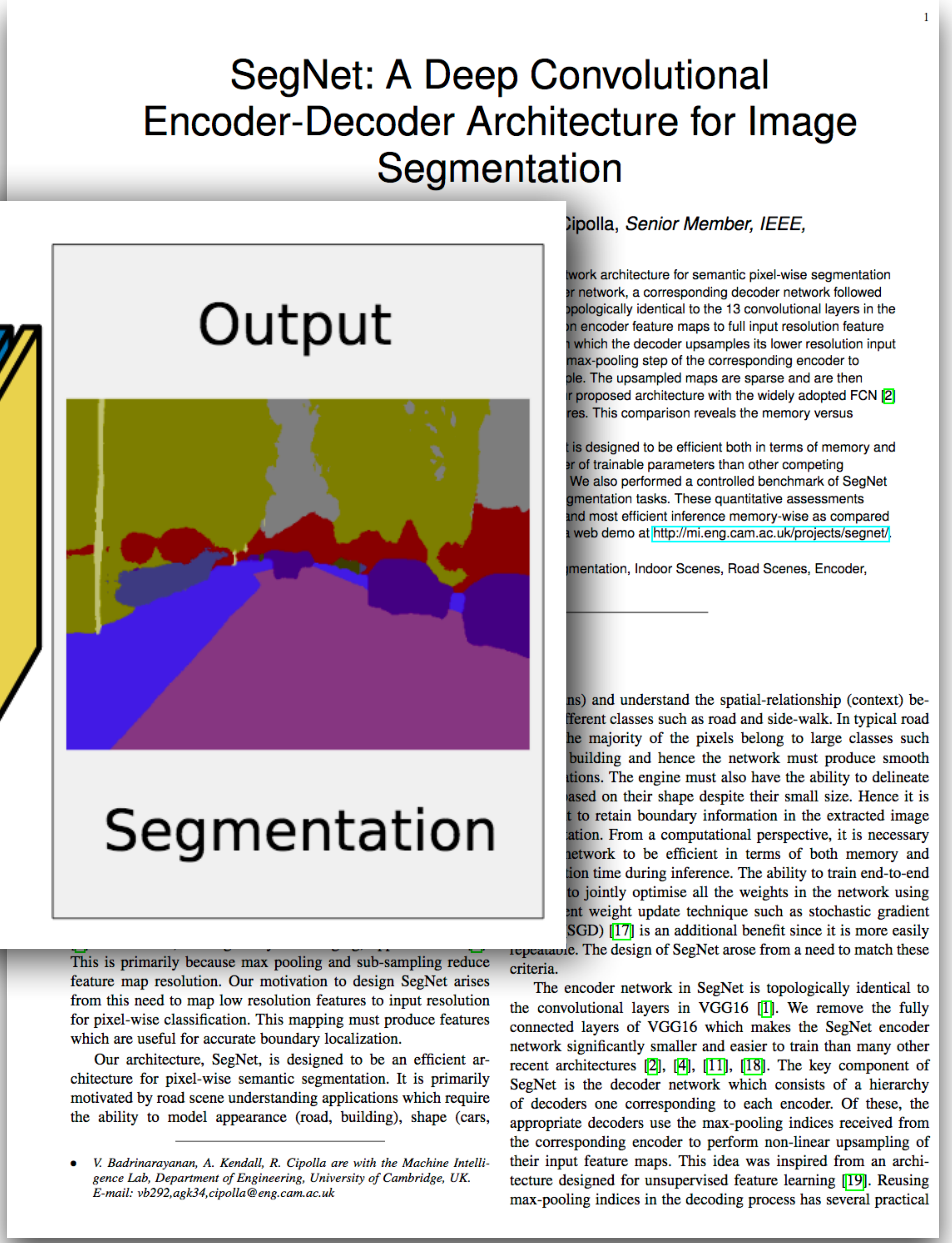
Early layers and late layers have same shape. Concatenate them!



U-Net

Figures from [Isola et al., “Image-to-Image Translation with Conditional Adversarial Networks”, 2017]

Encoder-decoder architectures


SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

...ipolla, *Senior Member, IEEE*,

...work architecture for semantic pixel-wise segmentation ...r network, a corresponding decoder network followed ...pologically identical to the 13 convolutional layers in the ...n encoder feature maps to full input resolution feature ... which the decoder upsamples its lower resolution input ...max-pooling step of the corresponding encoder to ...ble. The upsampled maps are sparse and are then ...r proposed architecture with the widely adopted FCN [2] ...res. This comparison reveals the memory versus

... is designed to be efficient both in terms of memory and ...r of trainable parameters than other competing ... We also performed a controlled benchmark of SegNet ...gmentation tasks. These quantitative assessments ...nd most efficient inference memory-wise as compared ... web demo at http://mi.eng.cam.ac.uk/projects/segnet/.

...mentation, Indoor Scenes, Road Scenes, Encoder,

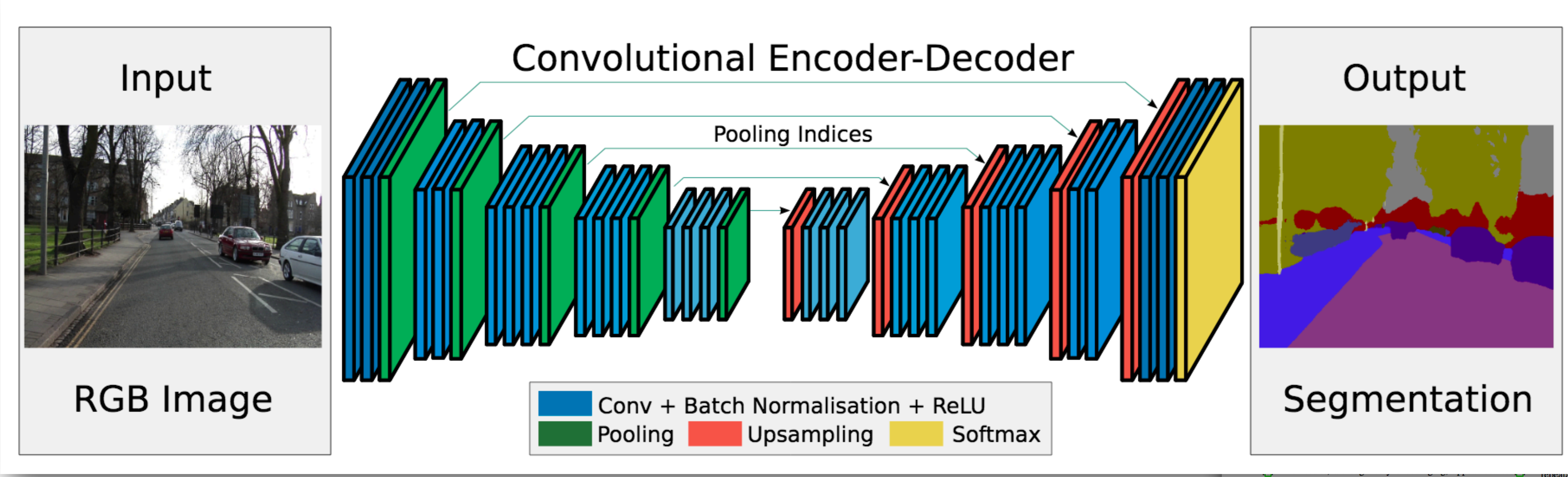
This is primarily because max pooling and sub-sampling reduce feature map resolution. Our motivation to design SegNet arises from this need to map low resolution features to input resolution for pixel-wise classification. This mapping must produce features which are useful for accurate boundary localization.

Our architecture, SegNet, is designed to be an efficient architecture for pixel-wise semantic segmentation. It is primarily motivated by road scene understanding applications which require the ability to model appearance (road, building), shape (cars,

...ns) and understand the spatial-relationship (context) be...fferent classes such as road and side-walk. In typical road ...he majority of the pixels belong to large classes such ...building and hence the network must produce smooth ...tions. The engine must also have the ability to delineate ...ased on their shape despite their small size. Hence it is ...t to retain boundary information in the extracted image ...ation. From a computational perspective, it is necessary ...etwork to be efficient in terms of both memory and ...ion time during inference. The ability to train end-to-end ...to jointly optimise all the weights in the network using ...nt weight update technique such as stochastic gradient ...SGD) [17] is an additional benefit since it is more easily repeatable. The design of SegNet arose from a need to match these criteria.

The encoder network in SegNet is topologically identical to the convolutional layers in VGG16 [1]. We remove the fully connected layers of VGG16 which makes the SegNet encoder network significantly smaller and easier to train than many other recent architectures [2], [4], [11], [18]. The key component of SegNet is the decoder network which consists of a hierarchy of decoders one corresponding to each encoder. Of these, the appropriate decoders use the max-pooling indices received from the corresponding encoder to perform non-linear upsampling of their input feature maps. This idea was inspired from an architecture designed for unsupervised feature learning [19]. Reusing max-pooling indices in the decoding process has several practical

- *V. Badrinarayanan, A. Kendall, R. Cipolla are with the Machine Intelligence Lab, Department of Engineering, University of Cambridge, UK. E-mail: vb292,agk34,cipolla@eng.cam.ac.uk*

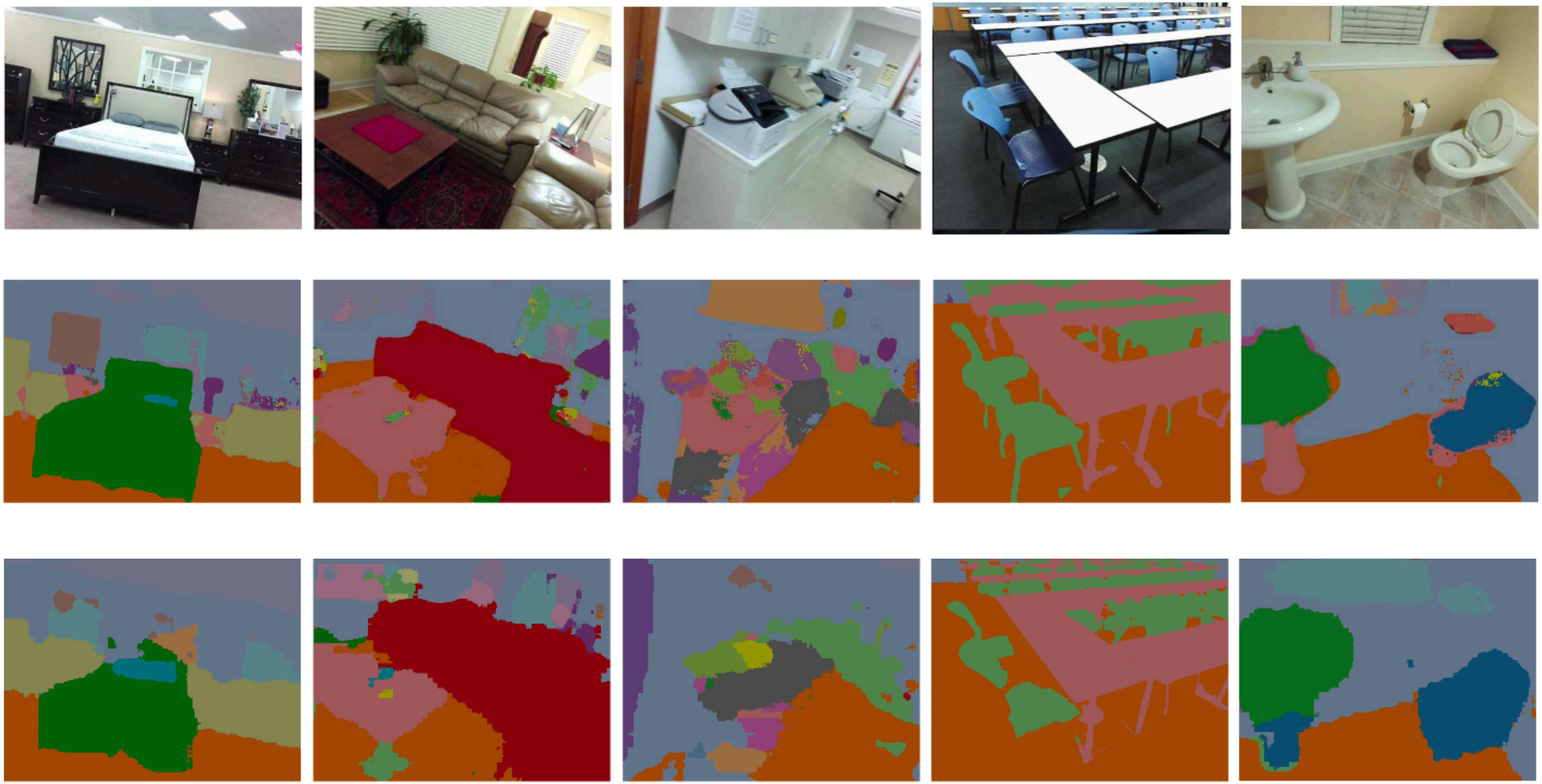


Source: Torralba, Freeman, Isola

Input

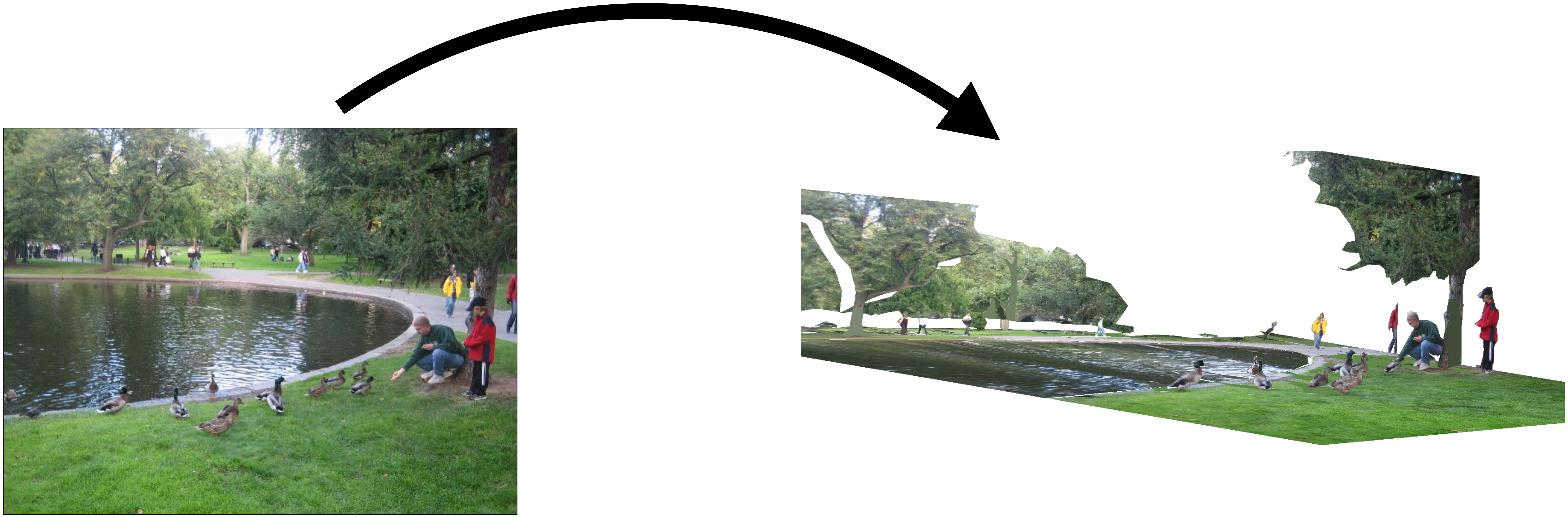
Segnet

FCN



Source: Torralba, Freeman, Isola

Depth perception



Vision systems

One camera

Two cameras

N cameras

Source: Torralba, Freeman, Isola

1 eye

Source: Torralba, Freeman, Isola

Shadows

Source: Steve Marschner

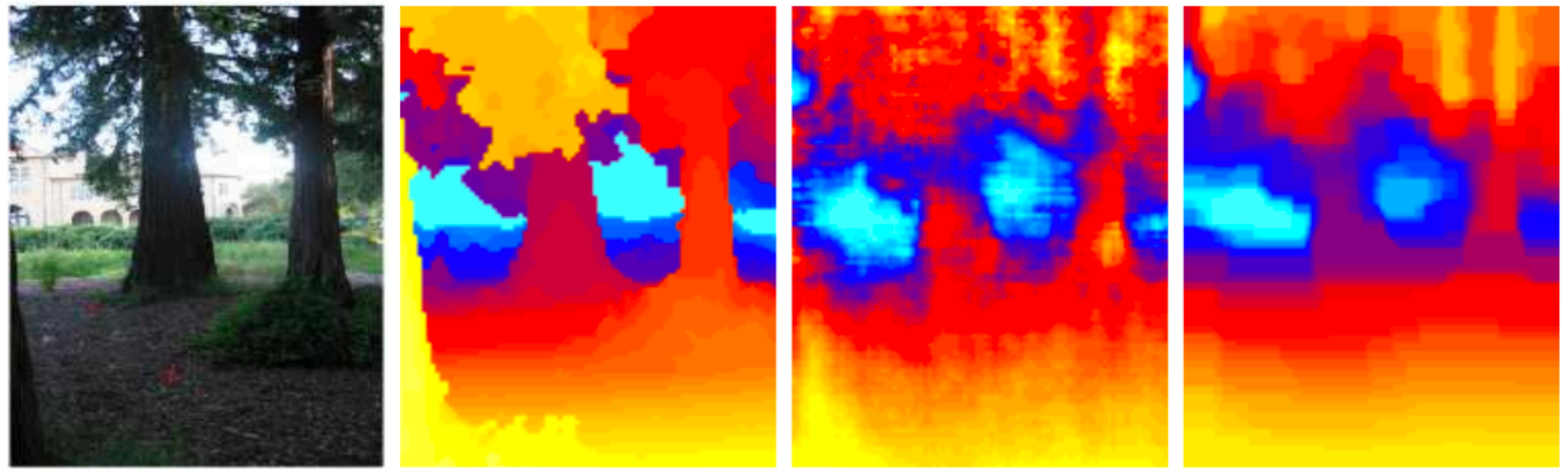
Learning based models

D. Hoiem, A.A. Efros, and M. Hebert, SIGGRAPH 2005.

A. Saxena, M. Sun, A. Y. Ng. 2007.

Make3D

Ashutosh Saxena, Sung H. Chung, Andrew Y. Ng. NeurIPS *18*, 2005.

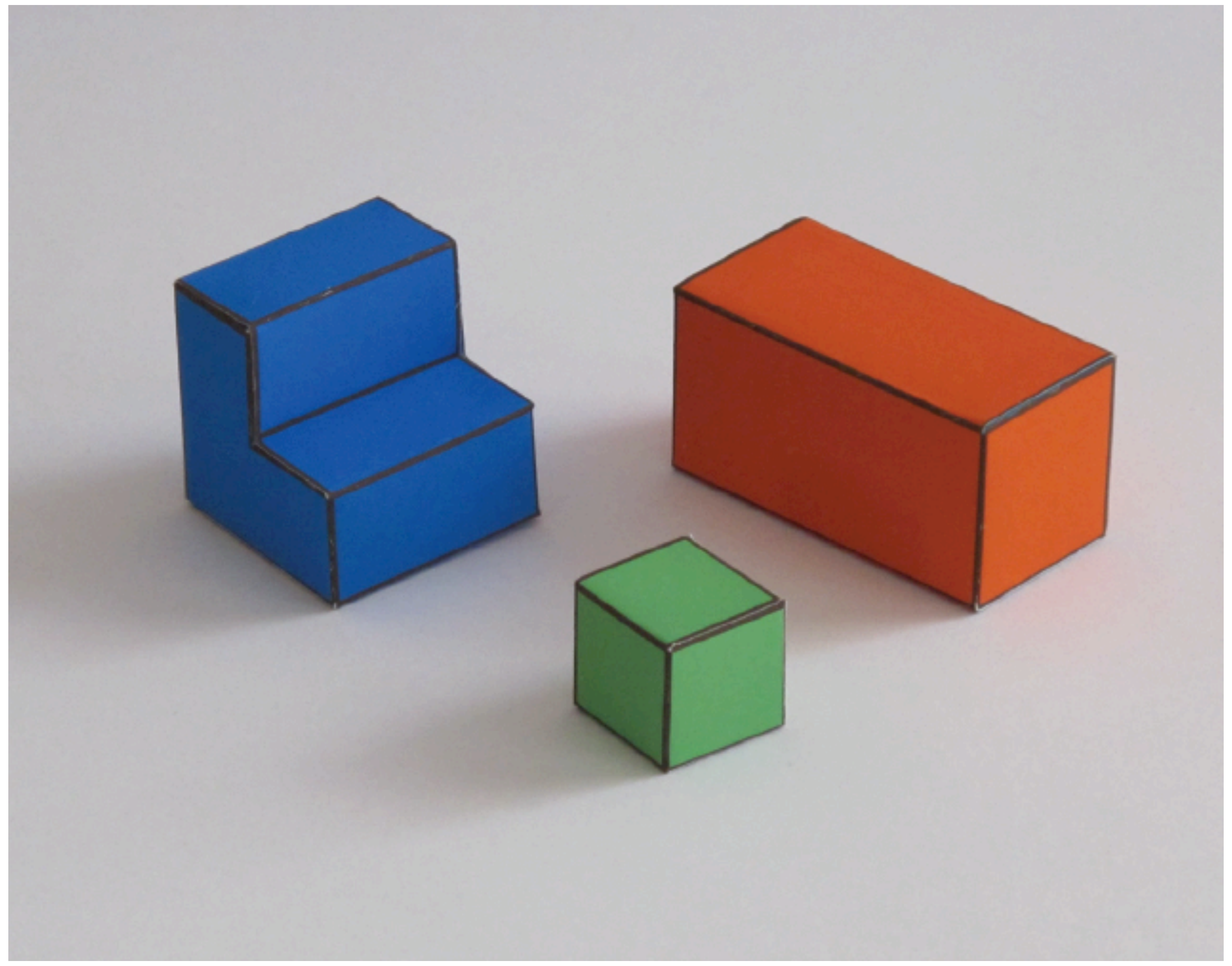


Karsch et al.

Ladicky et al.

…

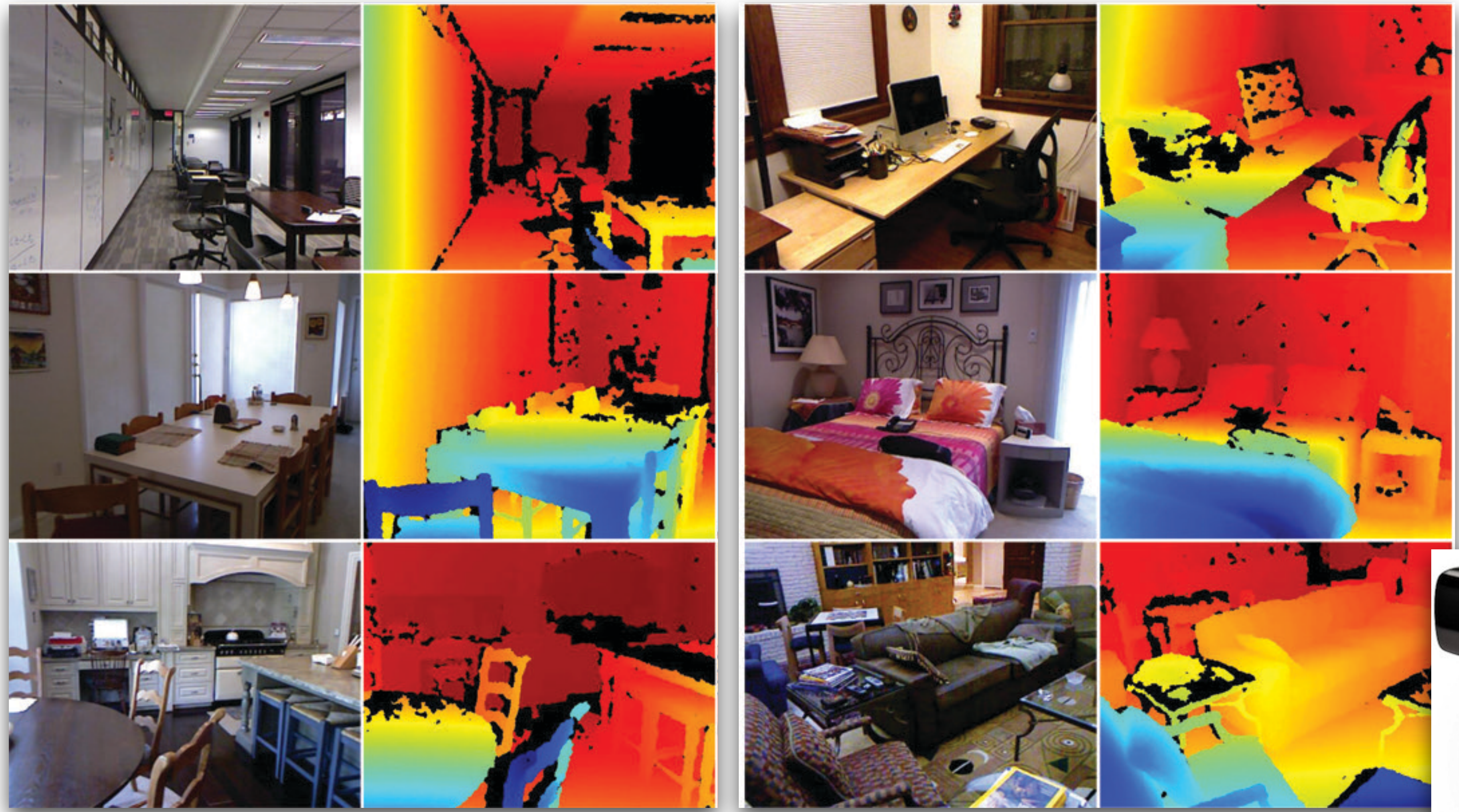
Source: Torralba, Freeman, Isola



3D scene understanding
in the deep net era

Source: Torralba, Freeman, Isola

3D in the deep learning era



Ground truth is collected by using traditional methods

Source: Torralba, Freeman, Isola

Datasets

KITTI

Cityscapes

"Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite", *Geiger et al.*, CVPR'12
"The Cityscapes Dataset for Semantic Urban Scene Understanding", *Cordts et al.*, CVPR'16

Source: Torralba, Freeman, Isola

Datasets

Cityscapes

KITTI

Source: Torralba, Freeman, Isola

Depth estimation

Data

$\{x_i, y_i\}_{i=1}^N$

$\rightarrow$

Learner

Objective

scale invariant MSE in log space

Hypothesis space

Deep Neural Network

Optimizer

SGD

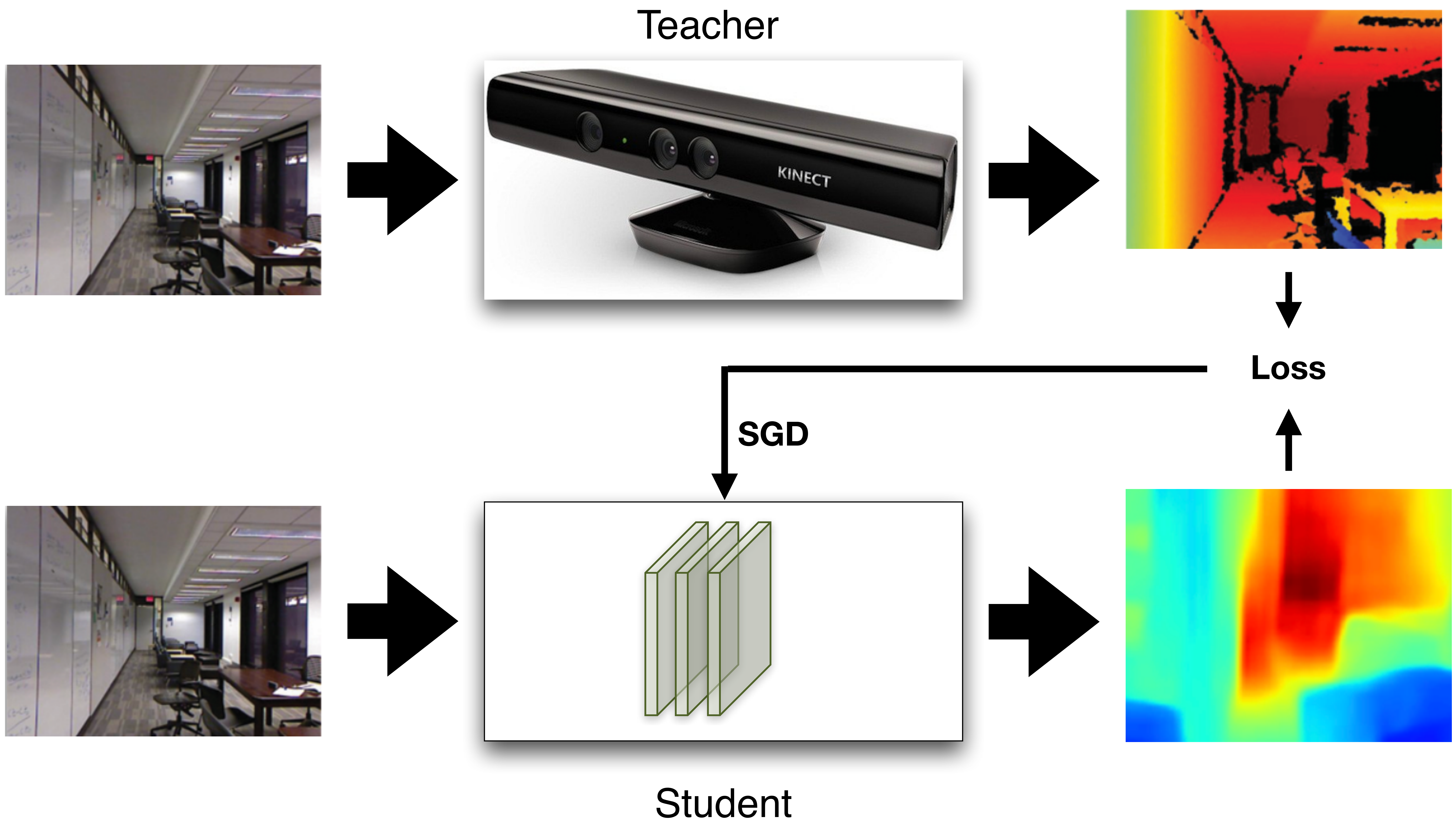
$\rightarrow f$

Regular old supervised learning!

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

Source: Torralba, Freeman, Isola

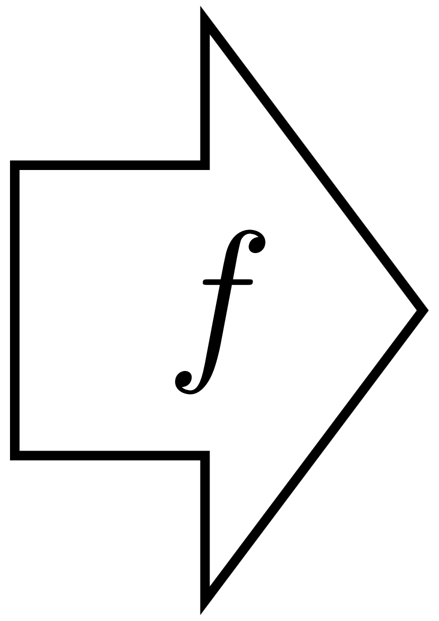
Depth estimation



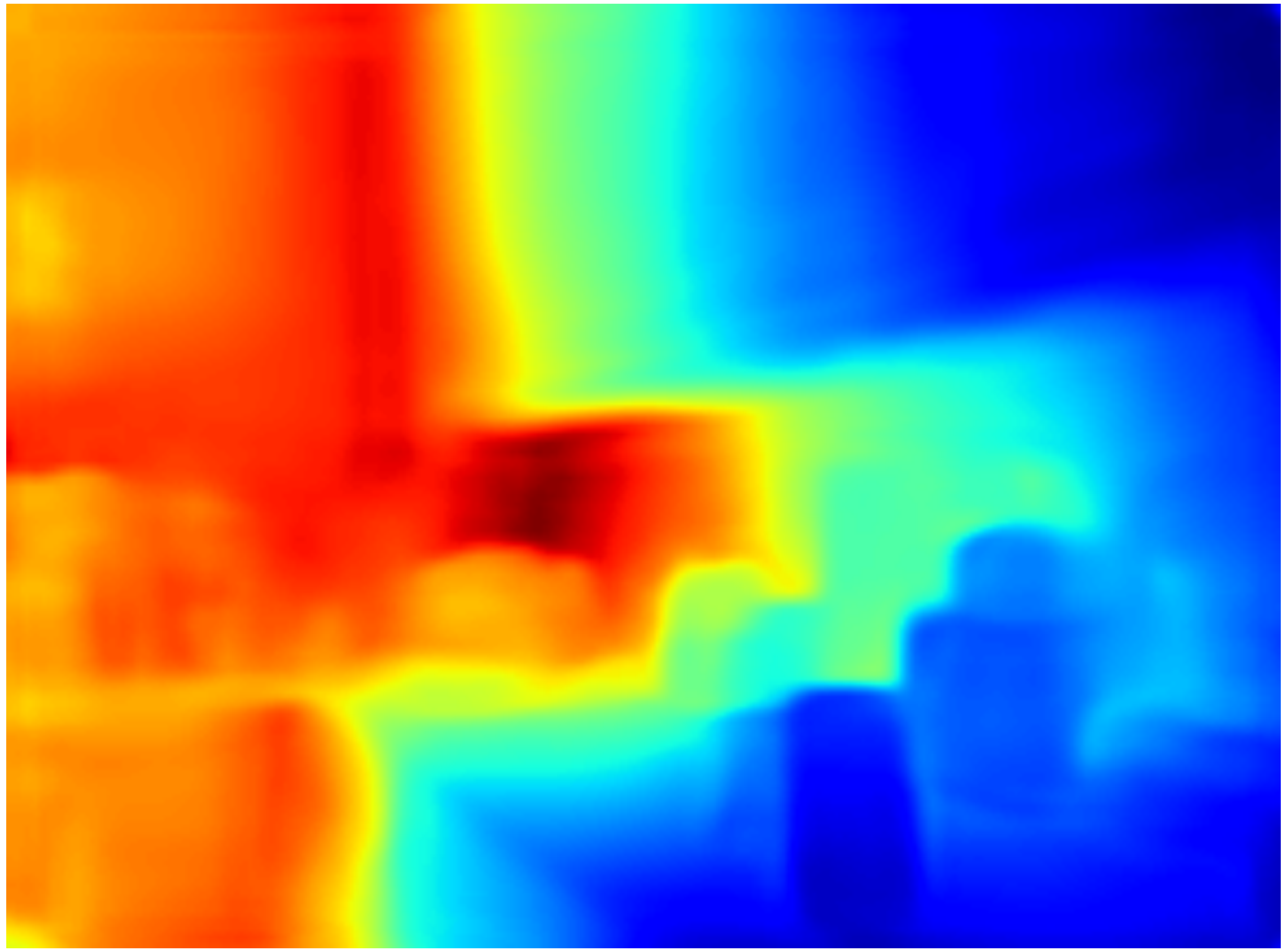
Source: Torralba, Freeman, Isola

Depth estimation

Input image



Predicted depth map



[Result of Eigen et al., NIPS, 2014]

Source: Torralba, Freeman, Isola

Regression problem

Estimate log depth instead of depth (matches human capabilities better).
Defining y_i the ground truth depth on pixel i, and y^*_i its estimated depth:

Standard L2 error:

$$D_{L2}(y, y^*) = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log y_i^*)^2$$

Scale invariant error:

$$D_{SI}(y, y^*) = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log y_i^* + \alpha(y, y^*))^2$$

with $\alpha(y, y^*) = \frac{1}{n}\sum_{j=1}^{n}(\log y_j - \log y_j^*)$

Source: Torralba, Freeman, Isola

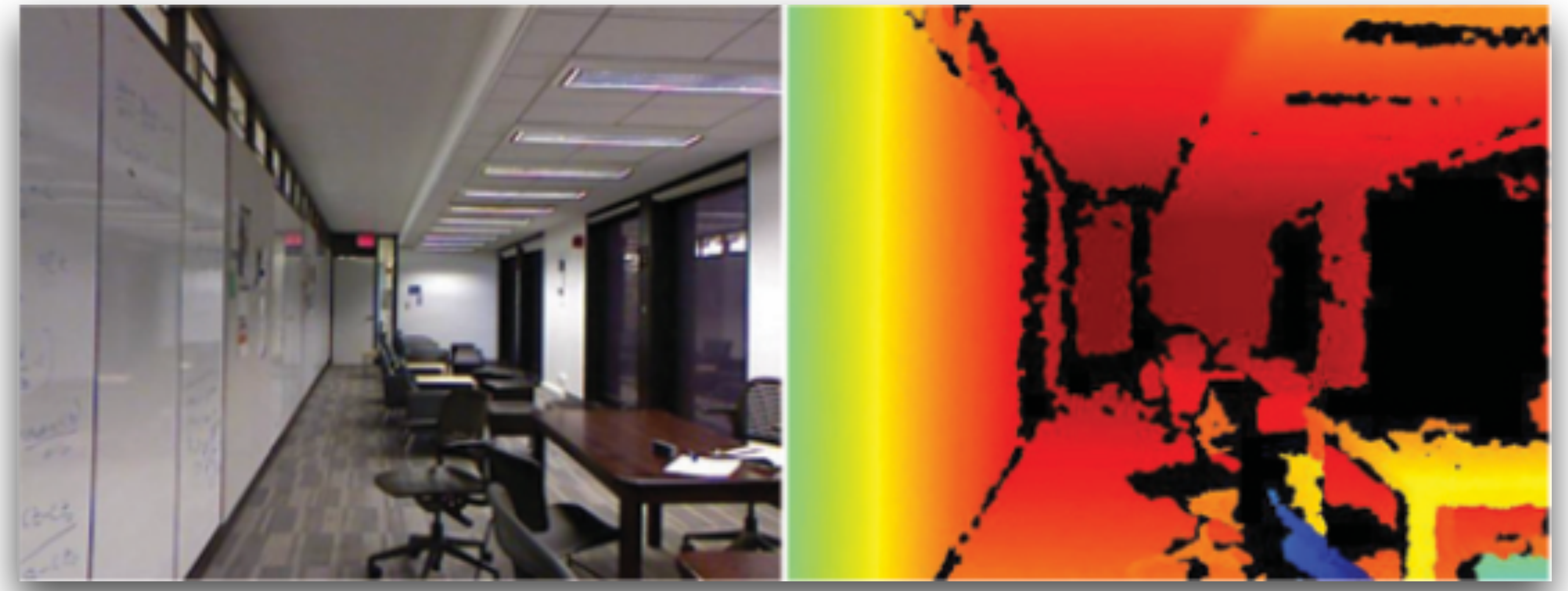
Training:

- **Training loss**: Mixture of both error measures (best \lambda=0.5):

Standard L2 error: **Scale invariant error:**

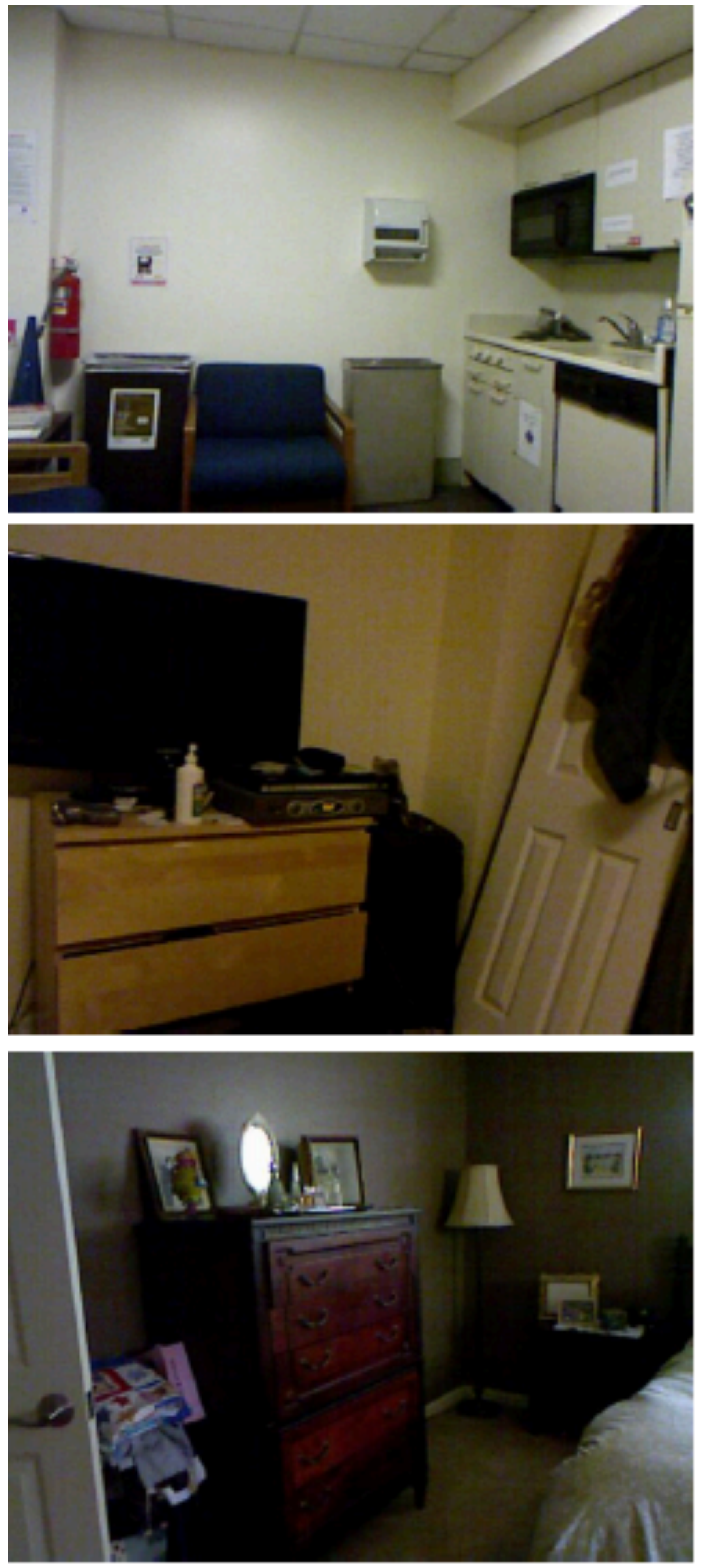
$$J = \lambda D_{L2}(y, y^*) + (1 - \lambda) D_{SI}(y, y^*)$$

Depth contains missing values. Only evaluate on valid pixels.



Source: Torralba, Freeman, Isola

Results (best)



Input

Prediction   Ground truth

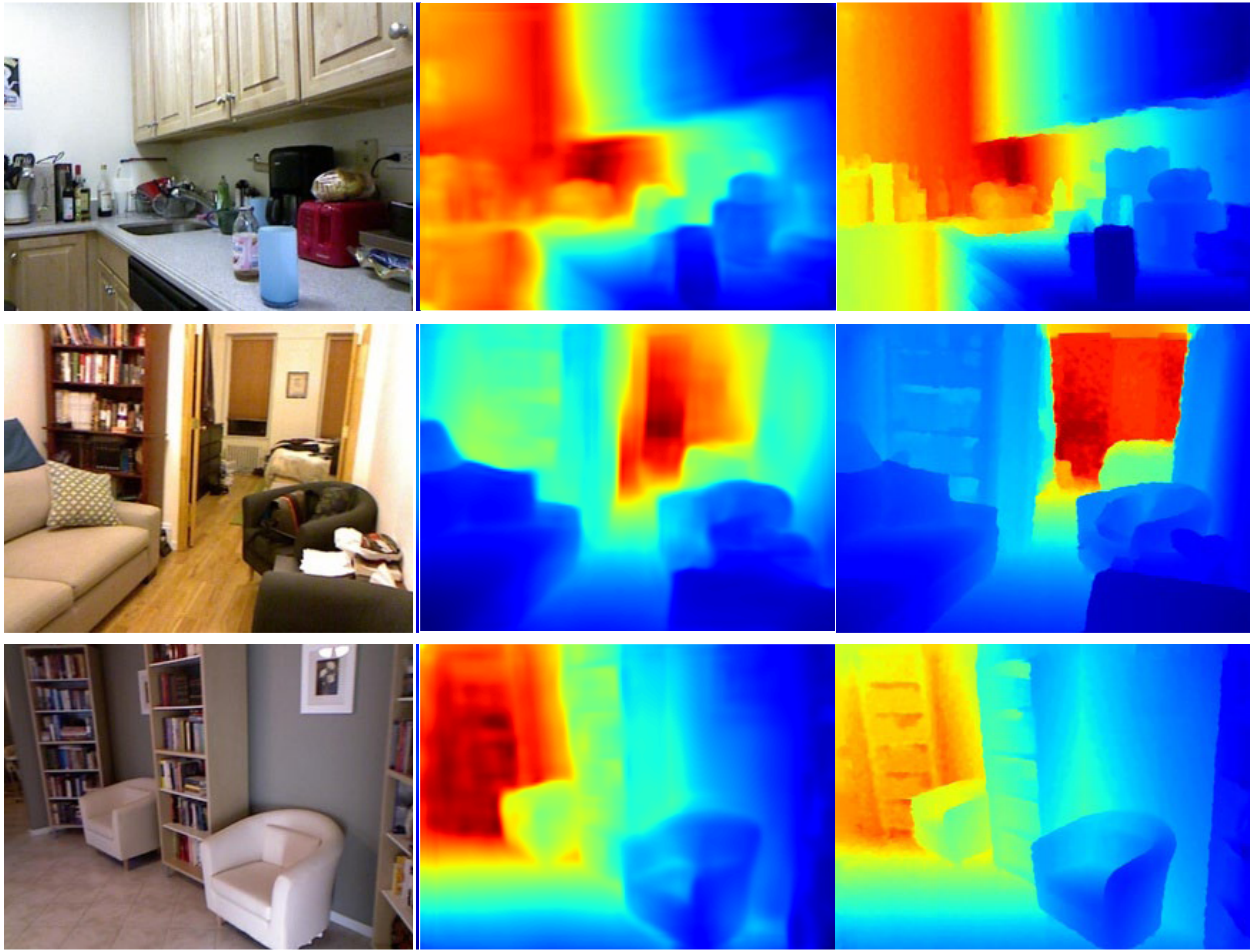
Source: Torralba, Freeman, Isola

Results (worst)

Input **Prediction** **Ground truth**

Source: Torralba, Freeman, Isola

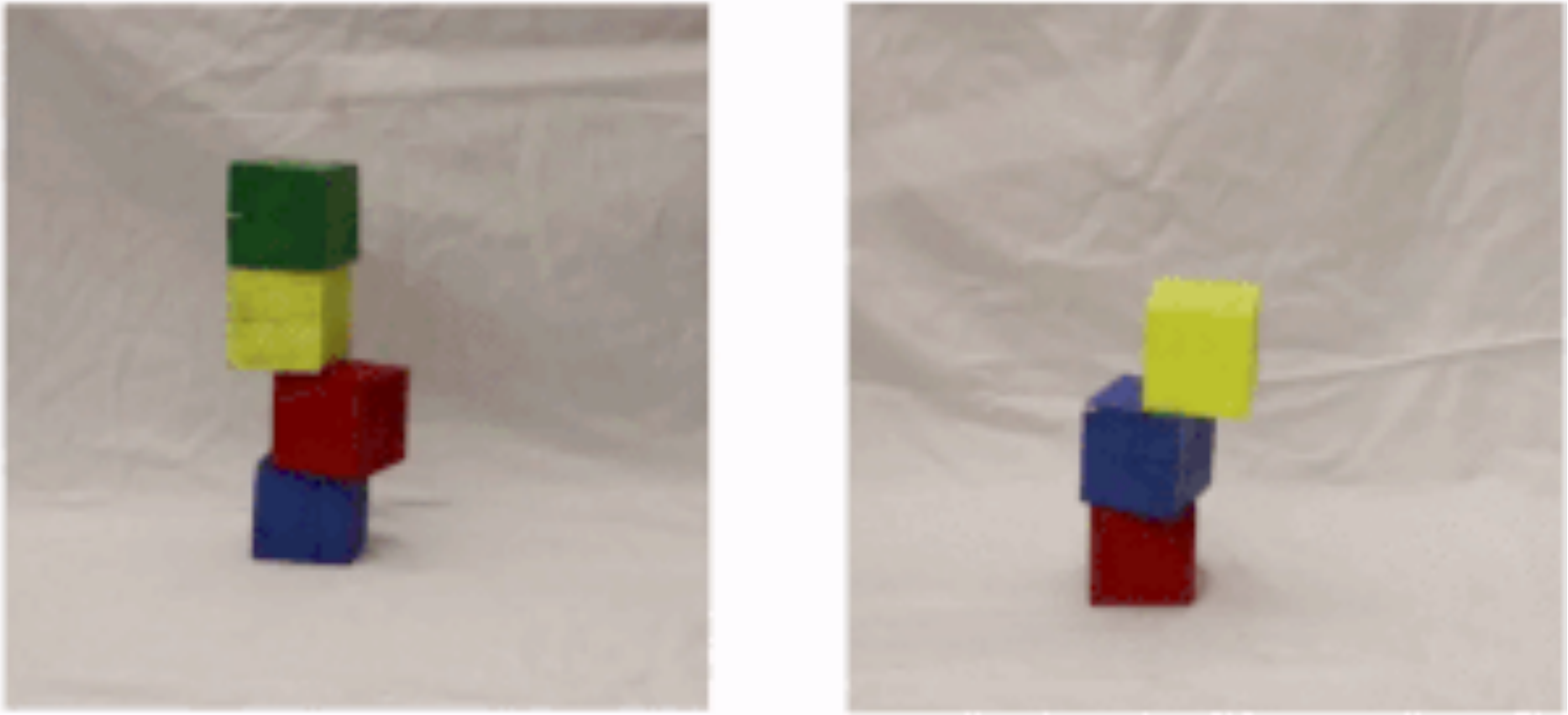
Results



Input Prediction Ground-truth

Source: Torralba, Freeman, Isola

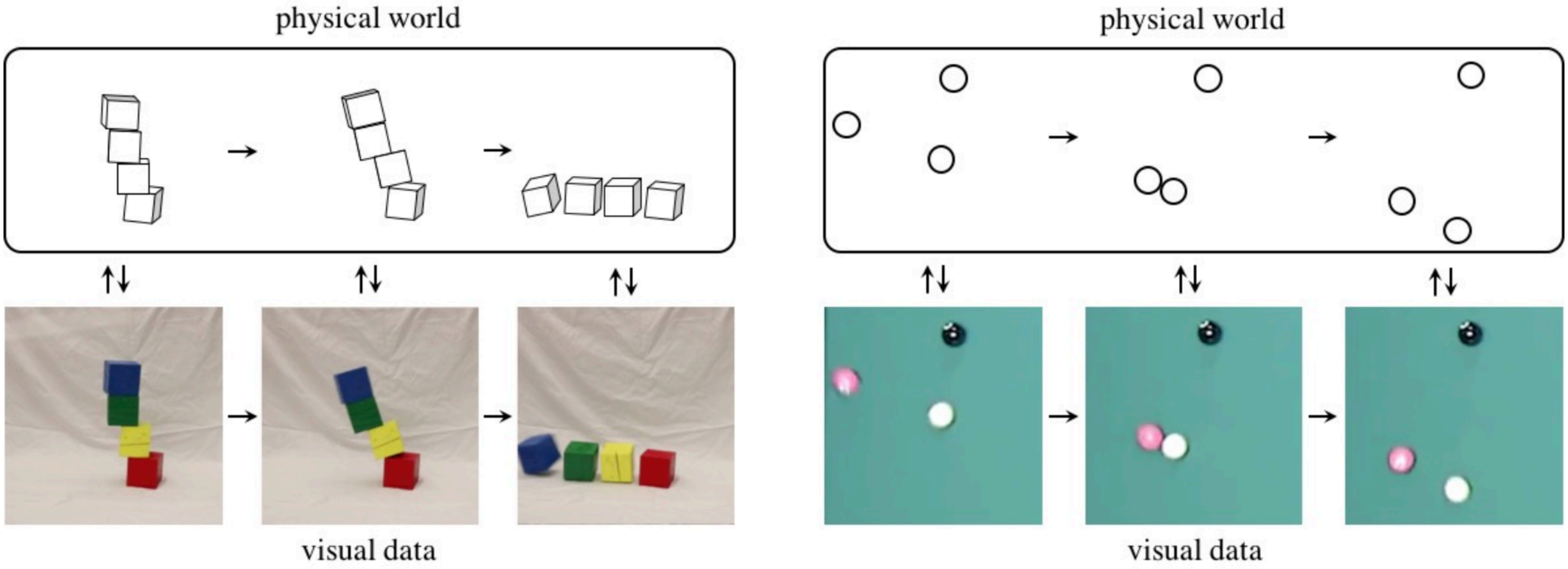
Intuitive physics



["Learning to See Physics via Visual De-animation", Wu et al., NIPS 2017]

Source: Torralba, Freeman, Isola

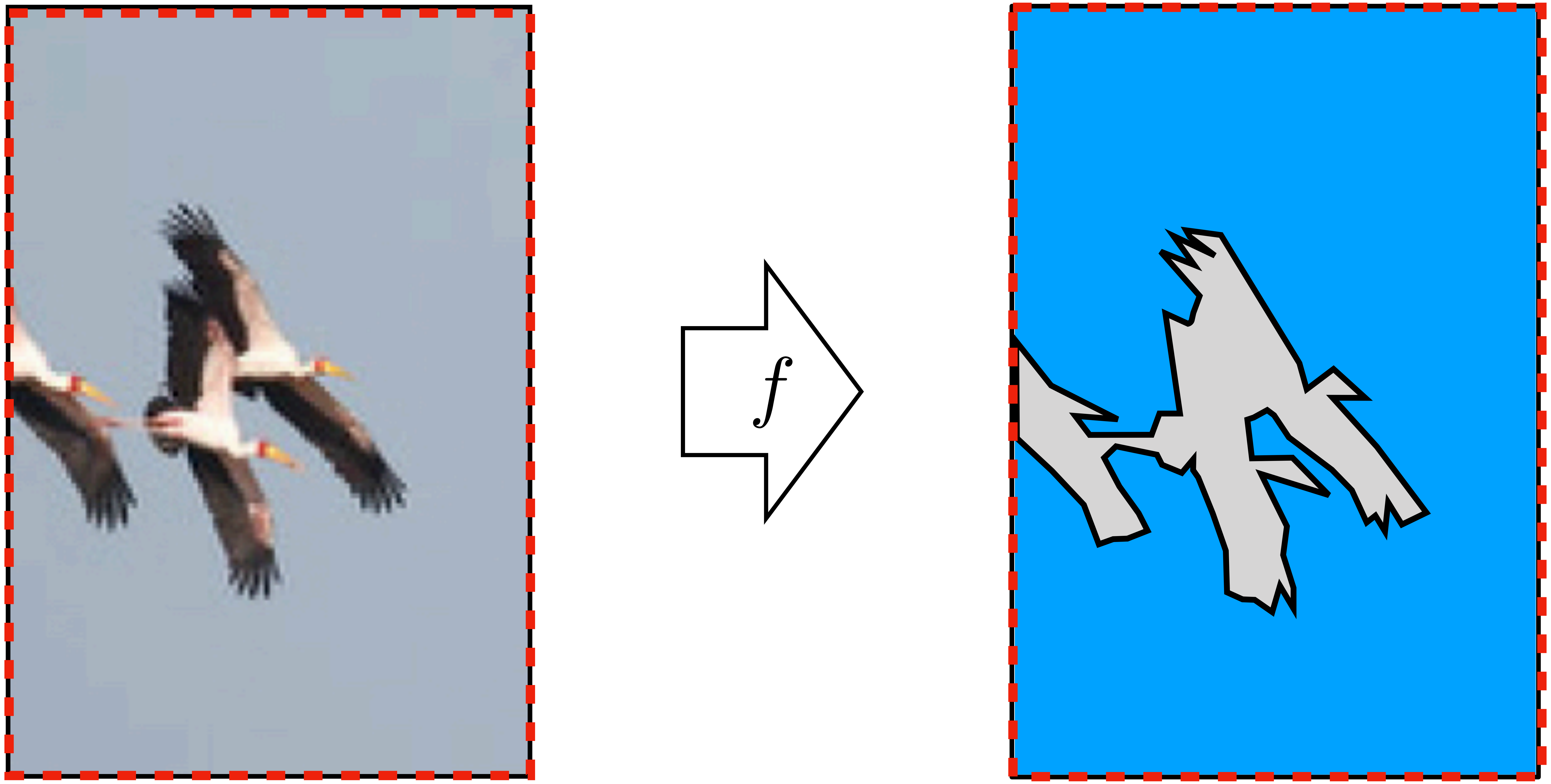
Intuitive physics



["Learning to See Physics via Visual De-animation", Wu et al., NIPS 2017]

Source: Torralba, Freeman, Isola

Semantic segmentation

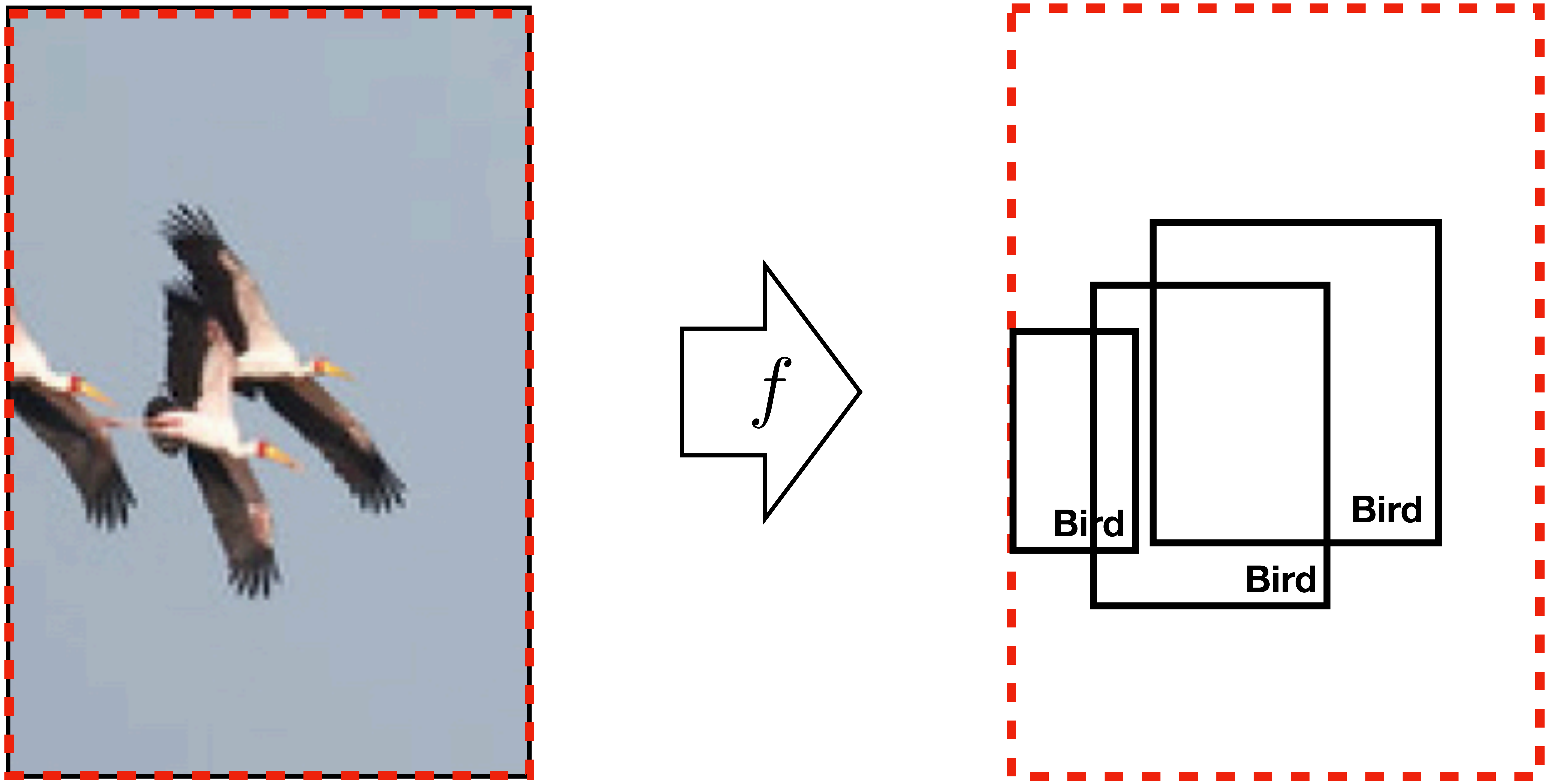


“A bunch of bird stuff”

Source: Torralba, Freeman, Isola

Object detection

Classification and localization

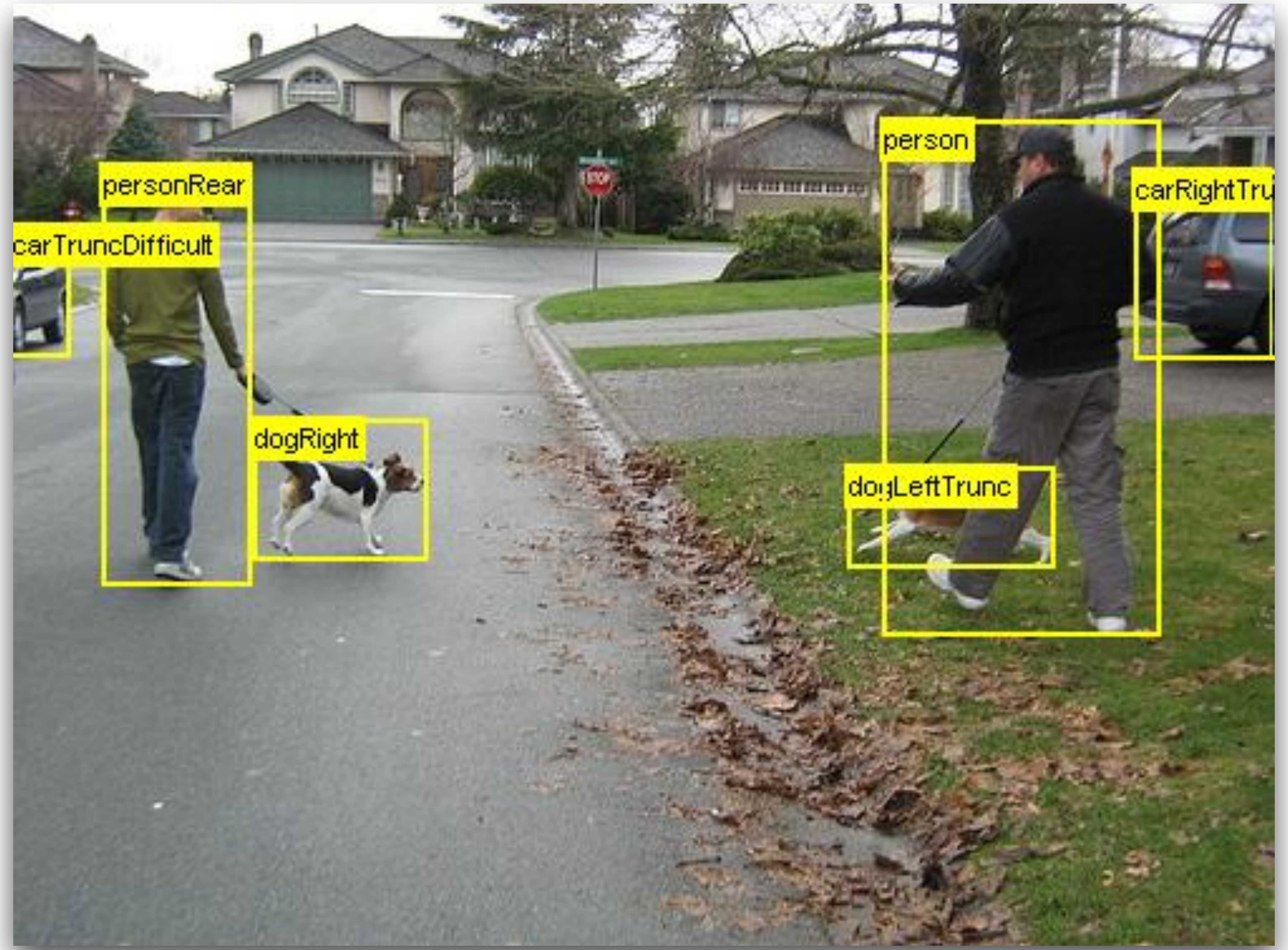


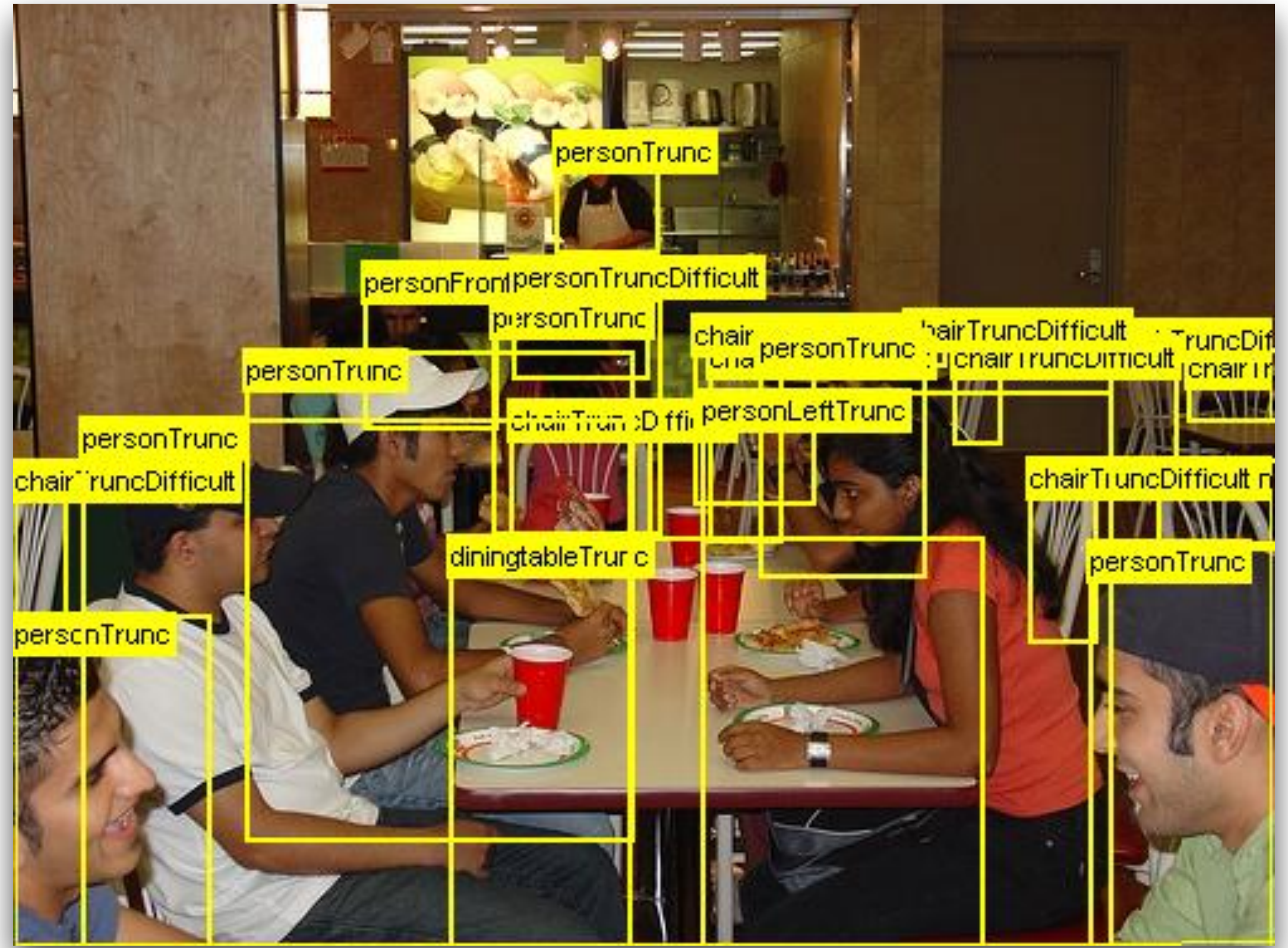
Each bounding box is:
[x,y,w,h]

Challenge: unbounded number of detections, possibly multiple detections per pixel

Source: Torralba, Freeman, Isola

PASCAL Visual Object Challenge





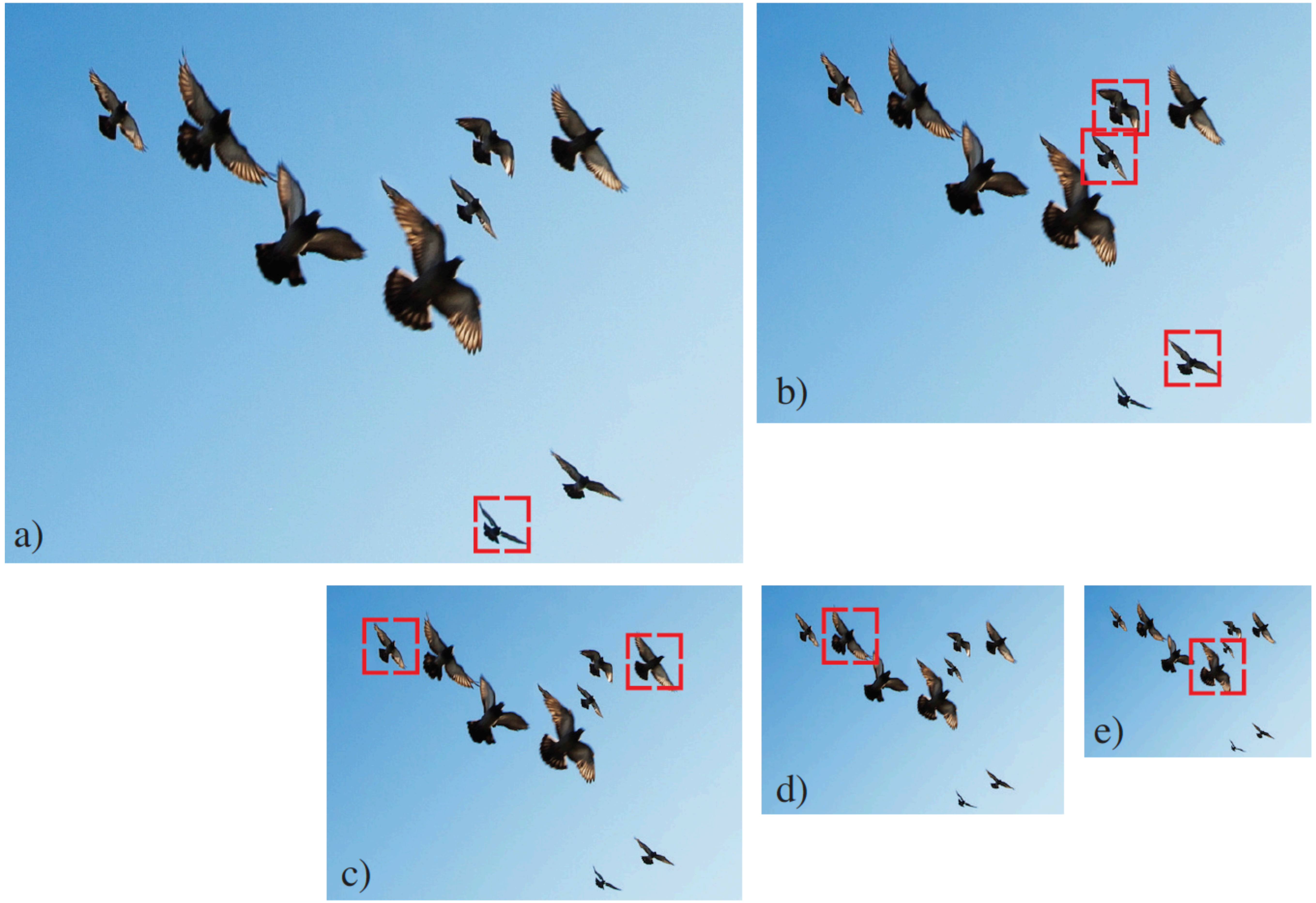
Source: Torralba, Freeman, Isola

Searching for objects

Scanning window approach
& Image pyramids

Source: Torralba, Freeman, Isola

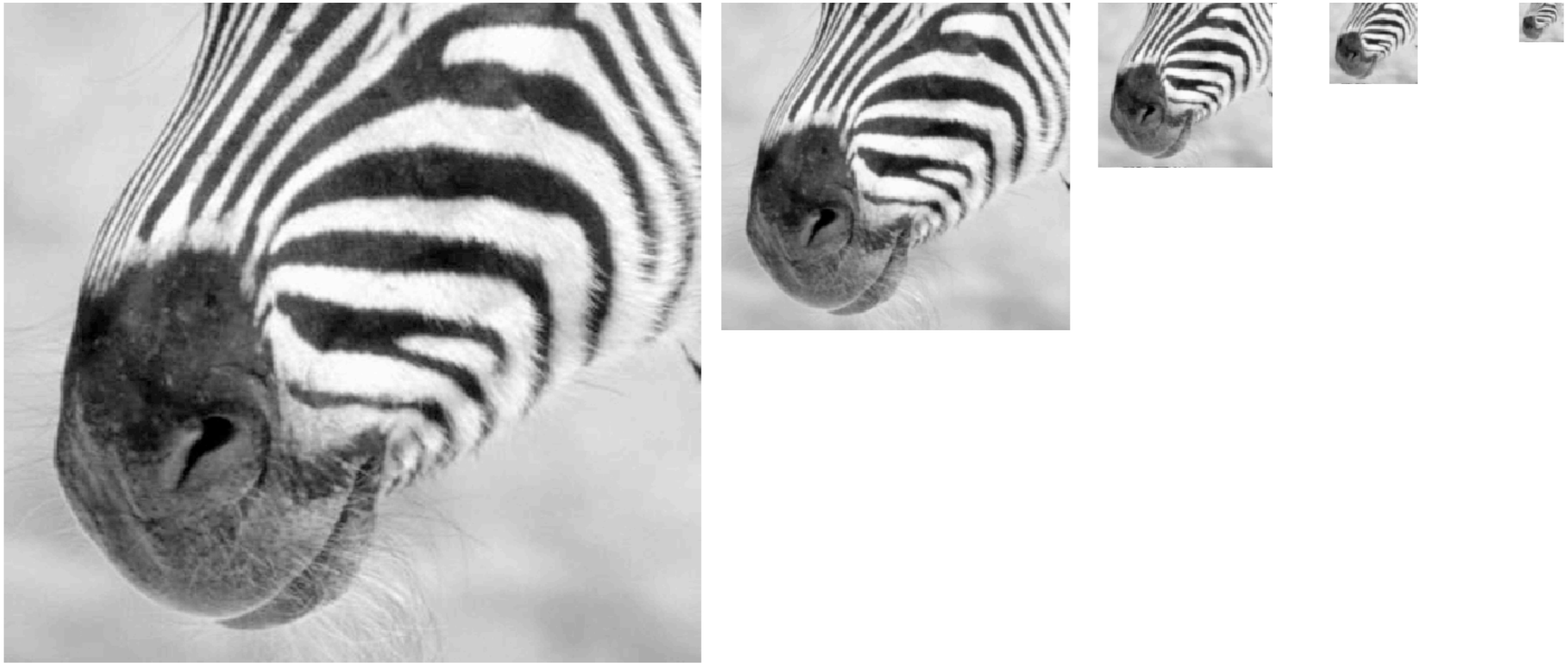
Image pyramids



Source: Torralba, Freeman, Isola

The Gaussian pyramid

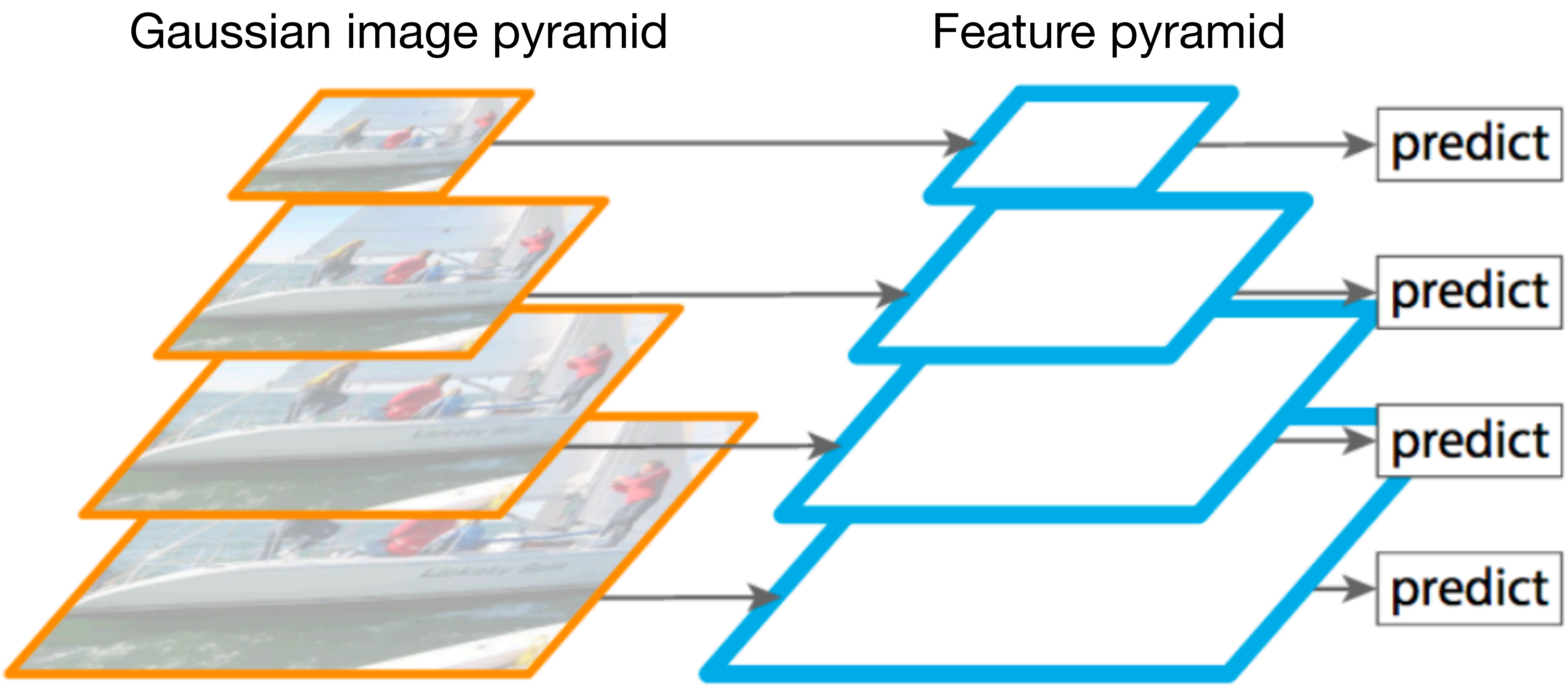
512×512 256×256 128×128 64×64 32×32



(original image)

Source: Torralba, Freeman, Isola

Could detect on each level of Gaussian pyramid…



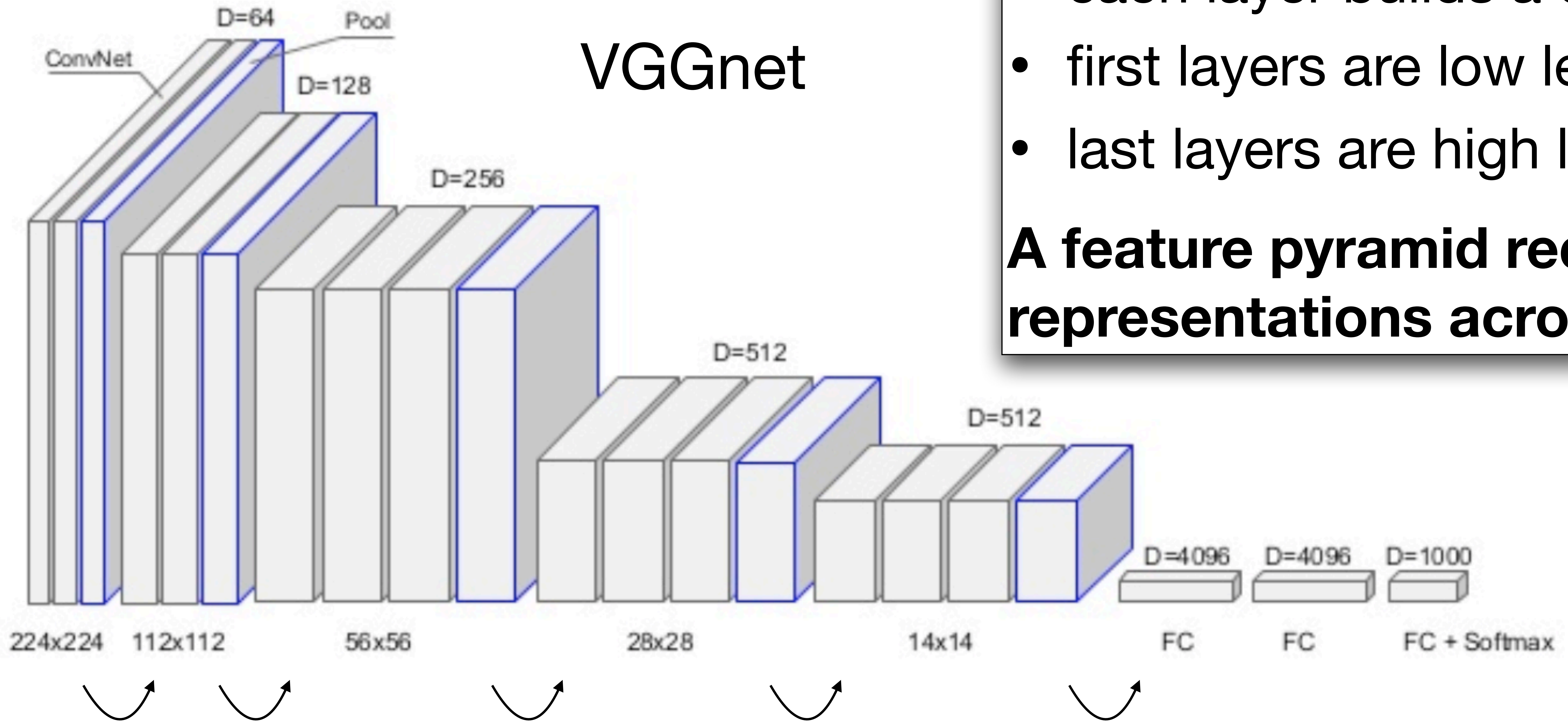
[Lin et al., “Feature Pyramid Networks for Object Detection”, 2017]

Source: Torralba, Freeman, Isola

Image and features pyramids

Each pooling reduces the resolution by
a factor of 2

VGGnet



ConvNet architectures build:

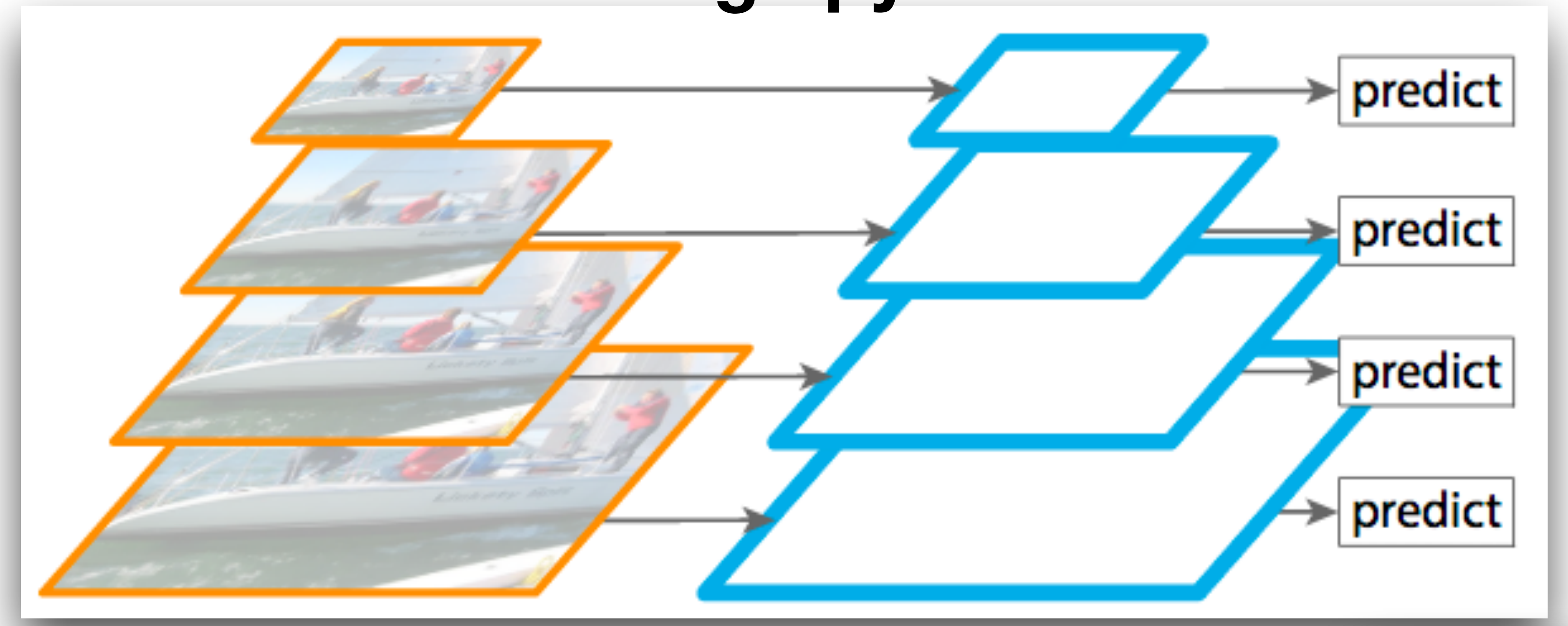
- Multiscale feature hierarchies, but
- each layer builds a different representation
- first layers are low level, while
- last layers are high level.

A feature pyramid requires a uniform representations across scales.

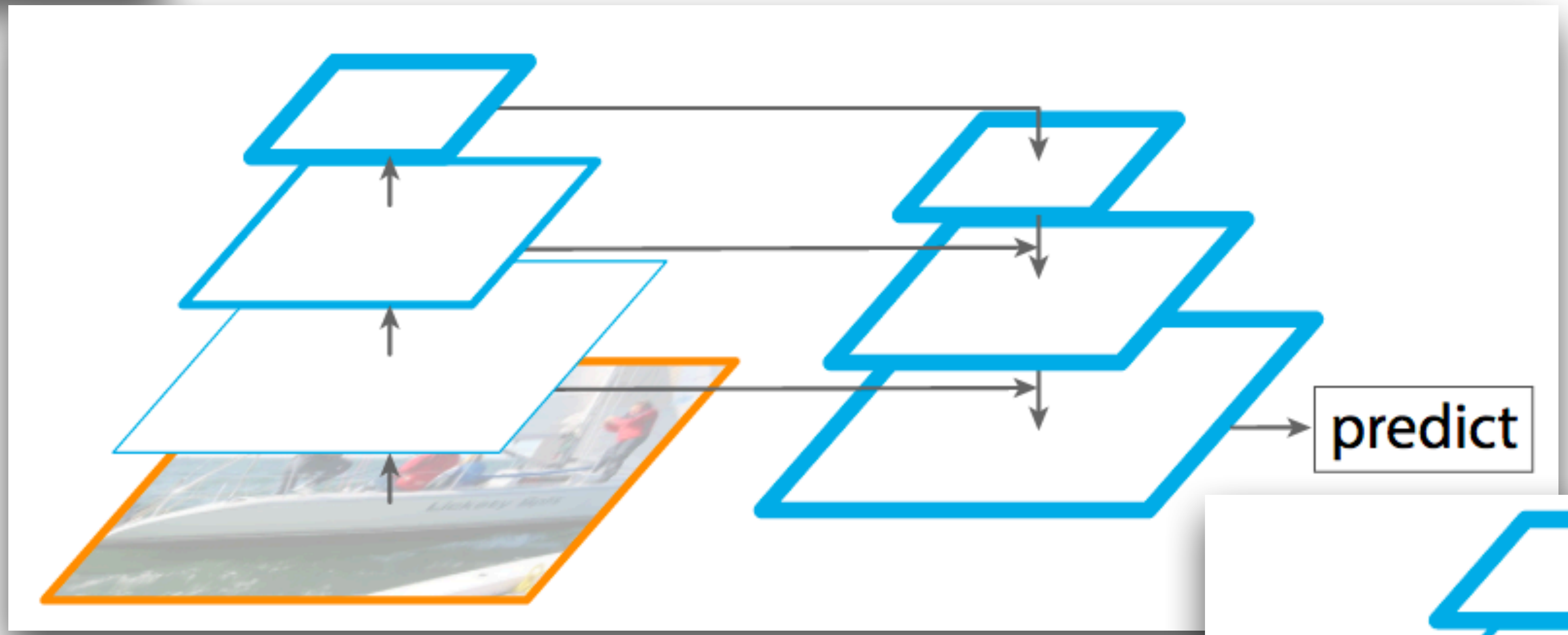
Source: Torralba, Freeman, Isola

Image and features pyramids

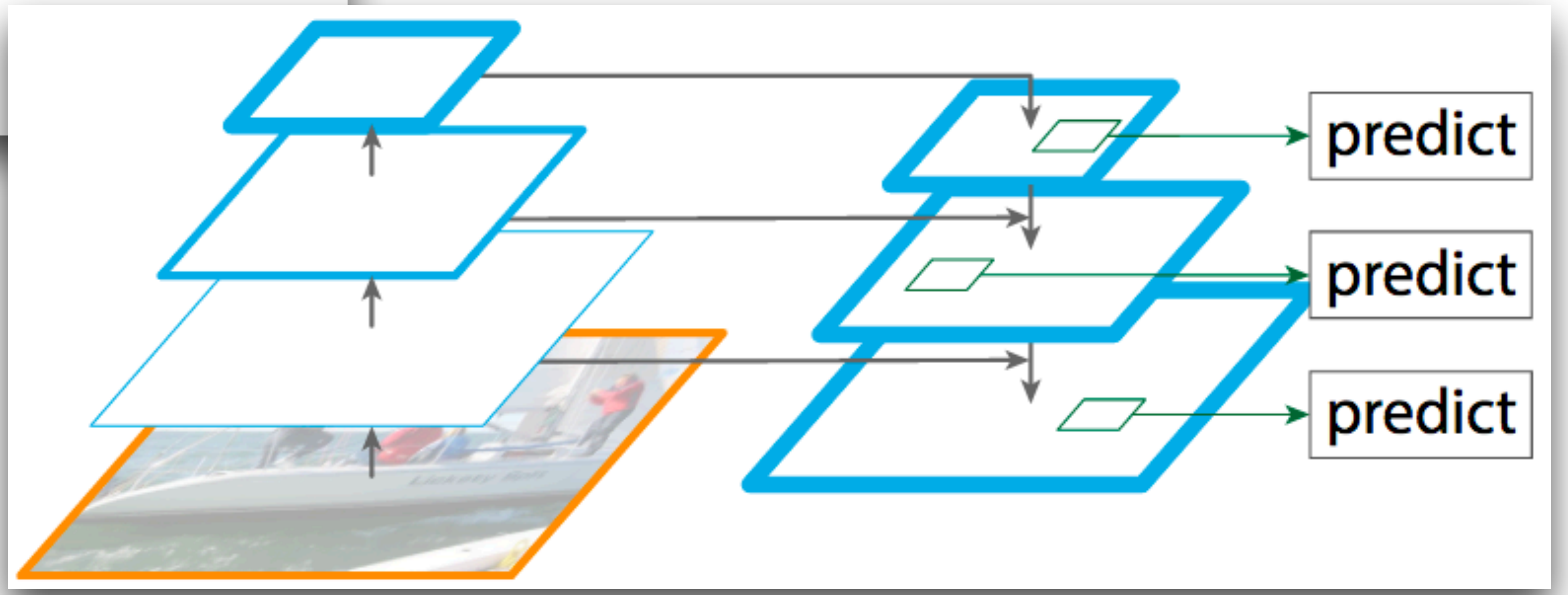
Image pyramid



Encoder-decoder architecture (U-Net)



Feature pyramid



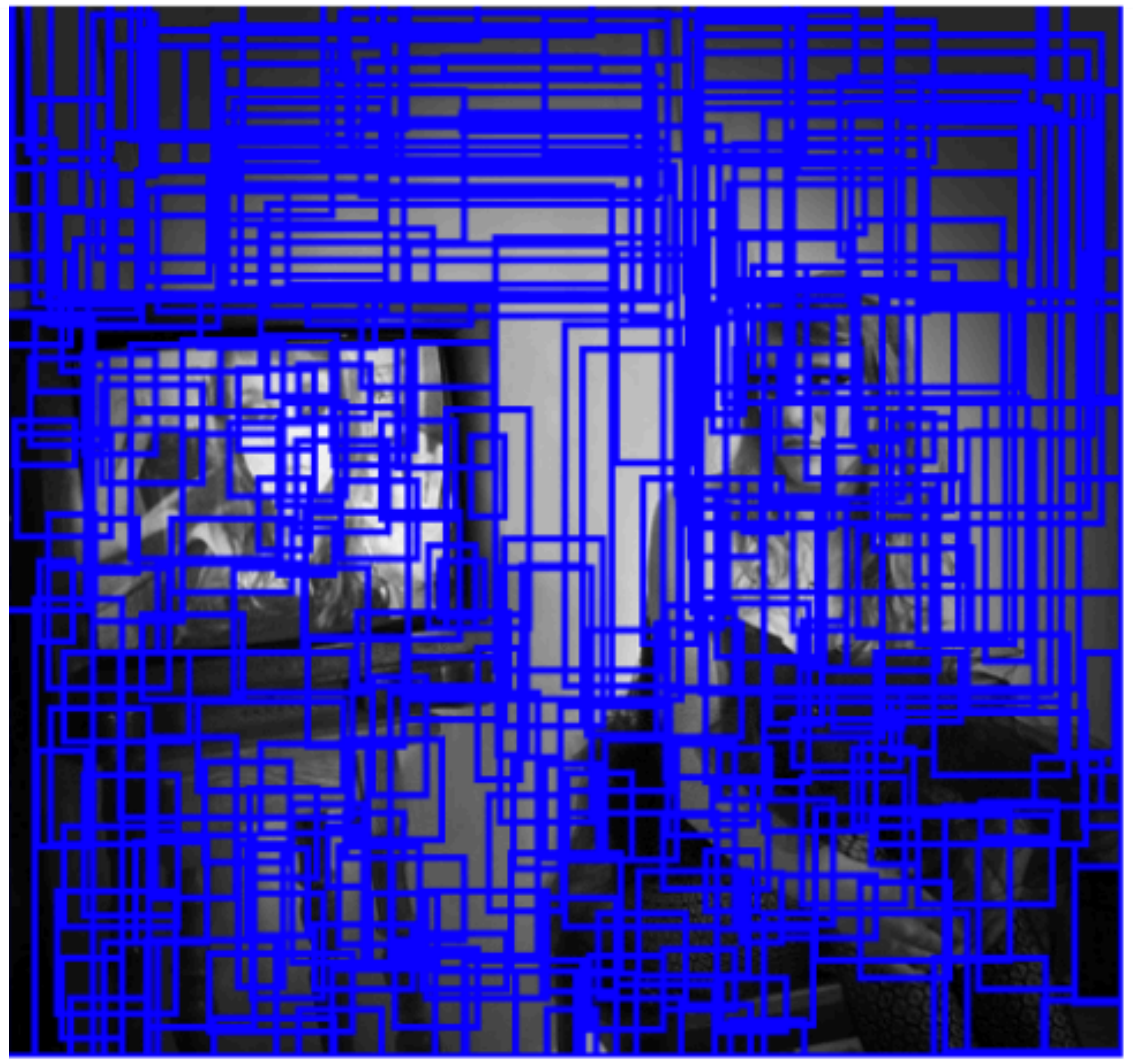
Source: Torralba, Freeman, Isola

Searching for objects

Scanning window approach
& Image pyramids

Selective search

Input image

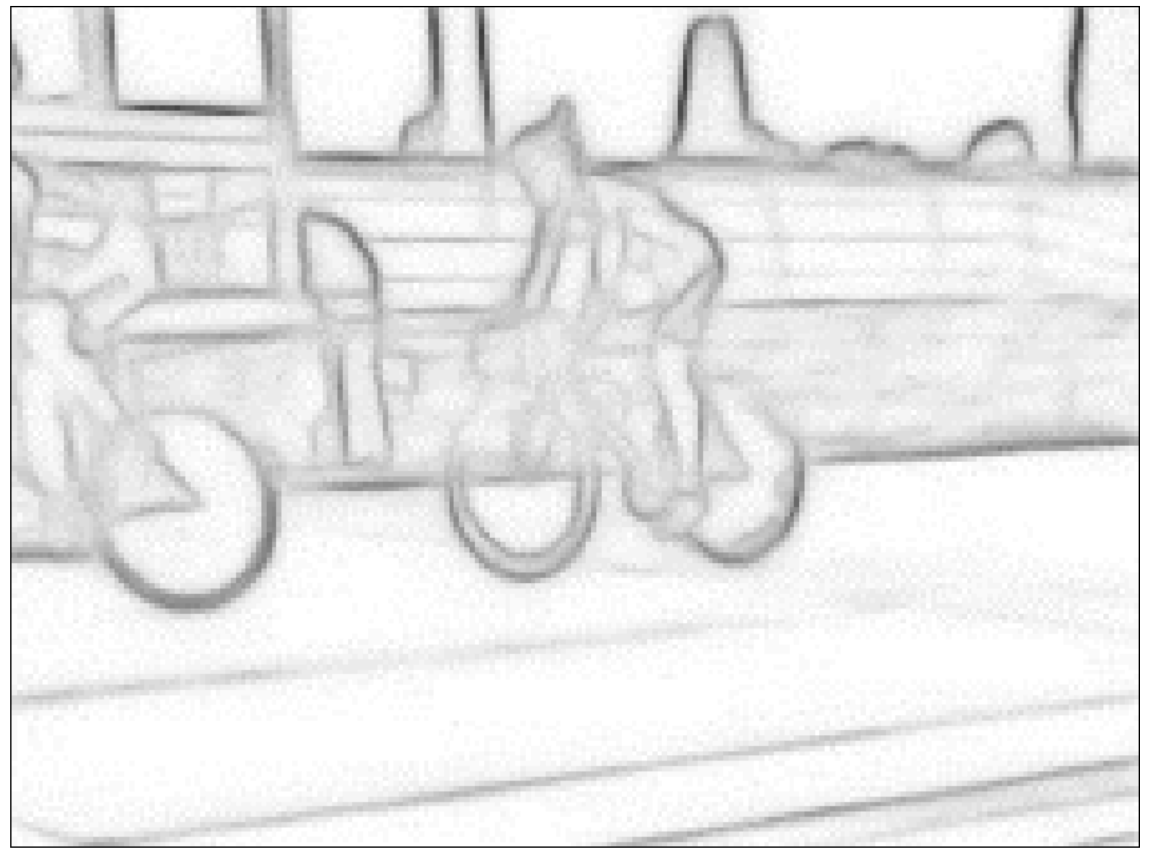


Candidate bounding boxes

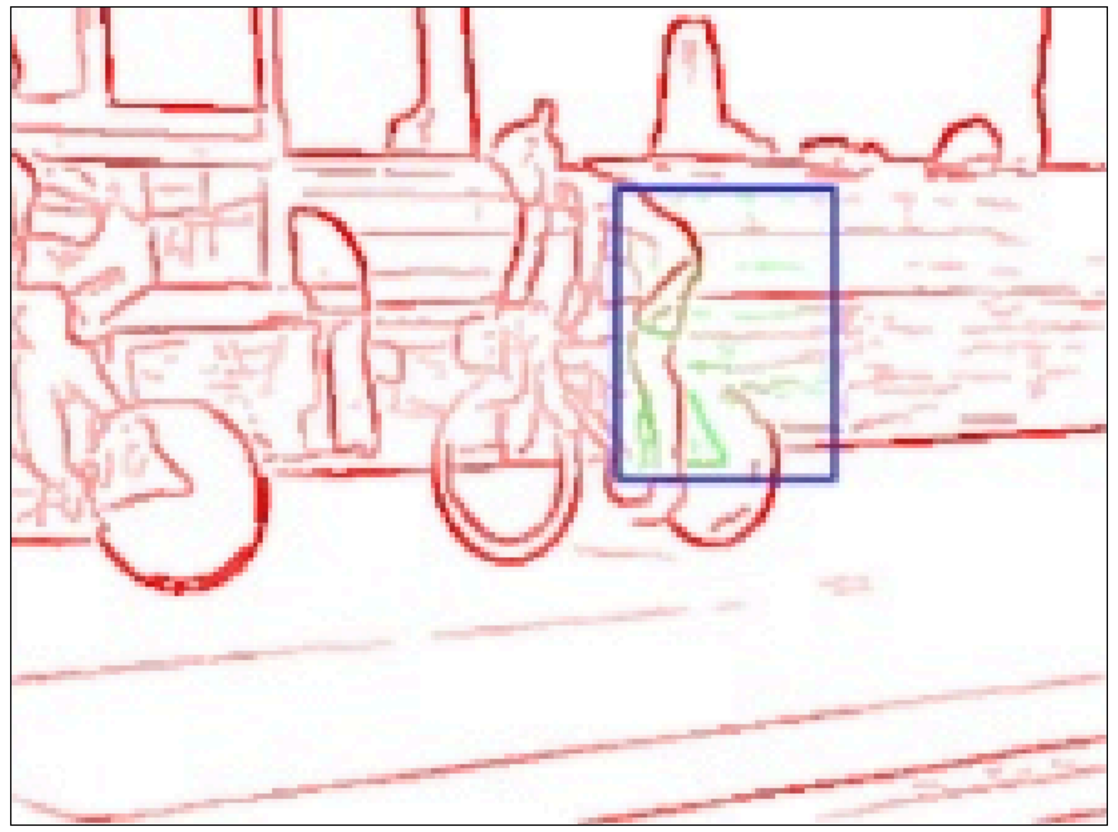
Selective search

Stage 1: generate candidate bounding boxes

Input image



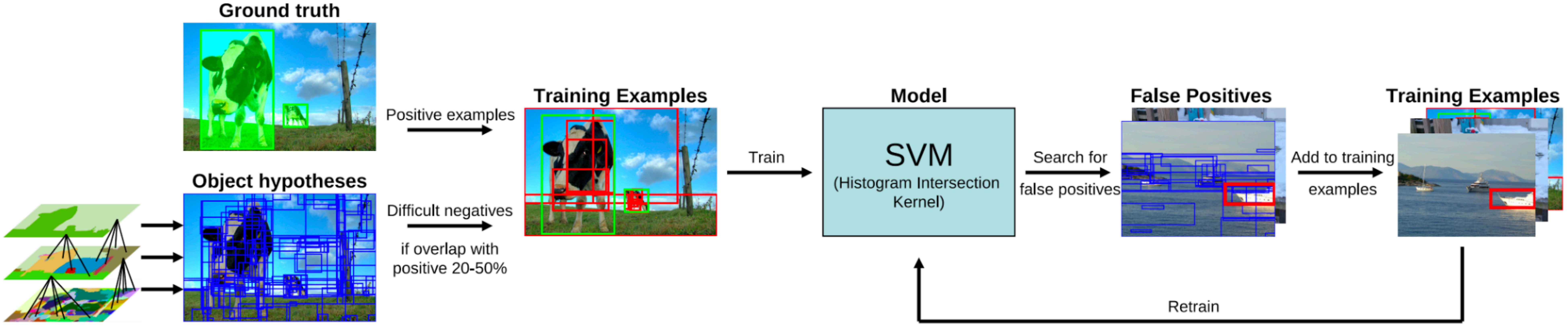
Edge detection



Bounding box proposal

[Zitnick and Dollar, "Edge Boxes…", 2014]

Stage 2: apply classifier to each candidate bounding box



[Uijlings et al., "Selective Search for Object Recognition", 2013]

Next time: More object detection