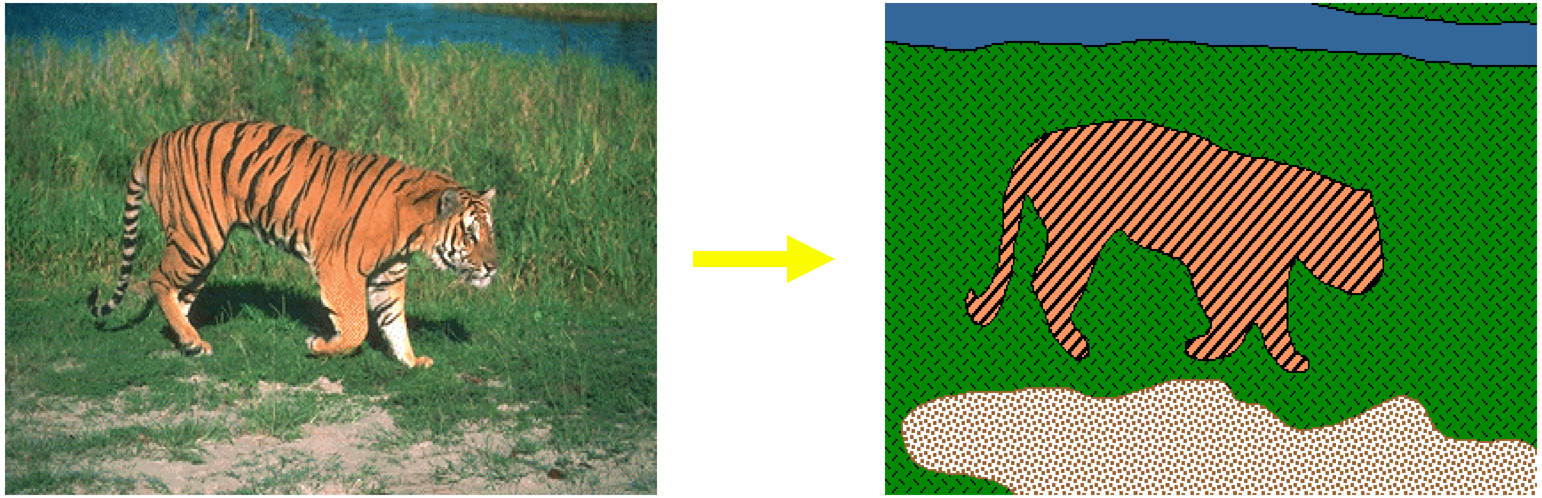
From Pixels to Objects: Recognition and Segmentation

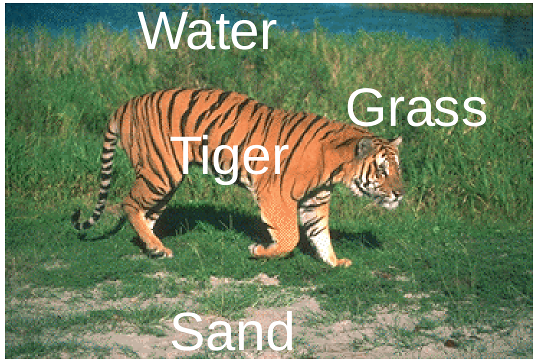
Jitendra Malik

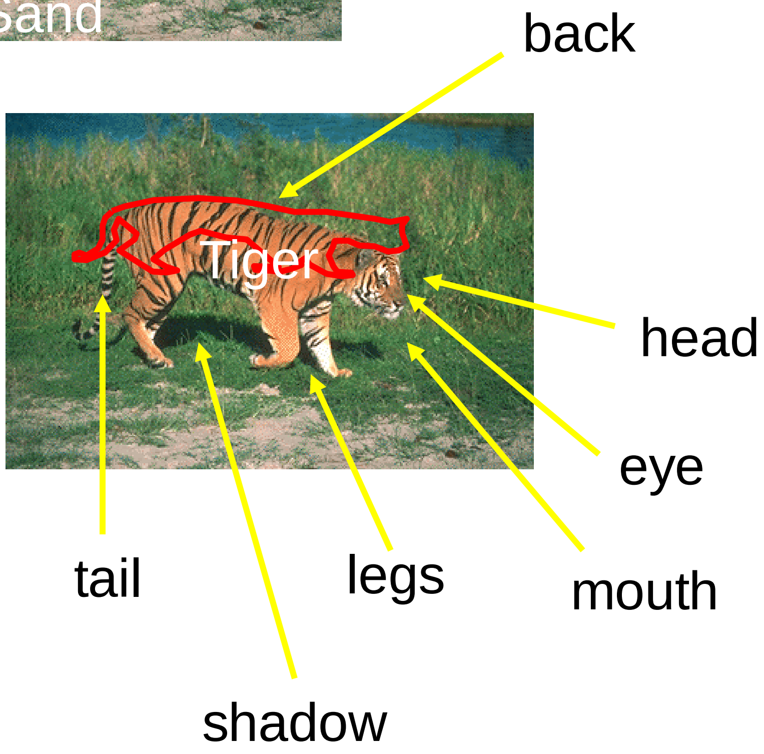
From Images to Objects

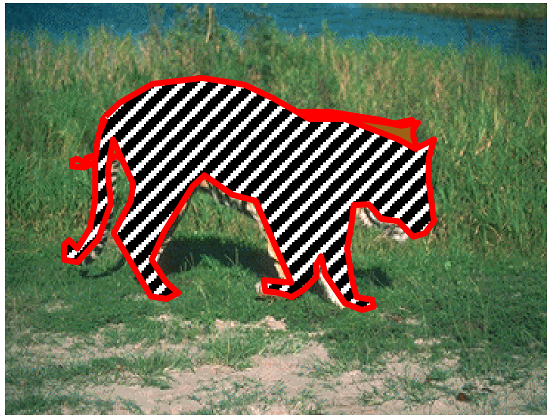


"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees." --Max Wertheimer

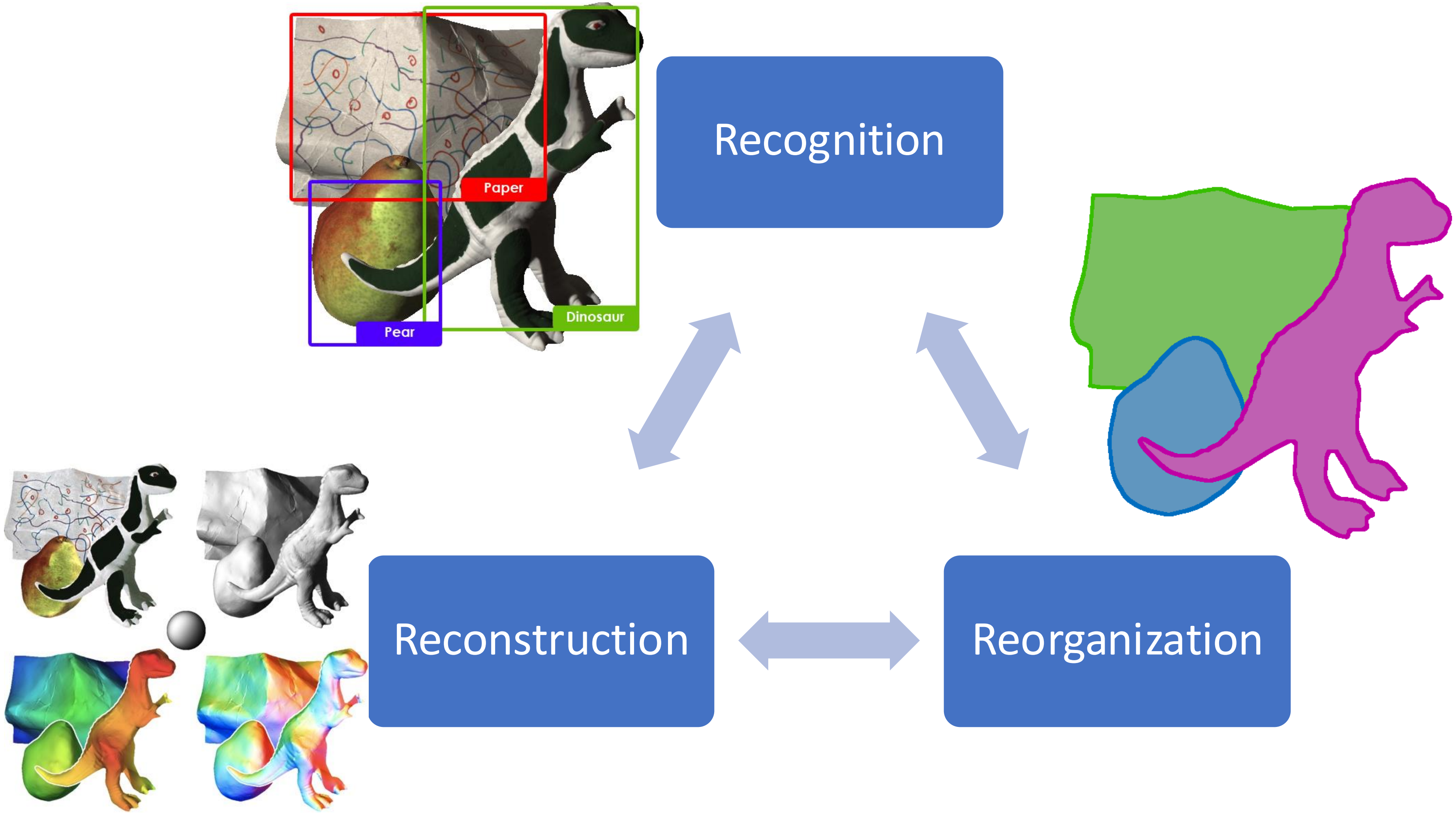
From Pixels to Perception





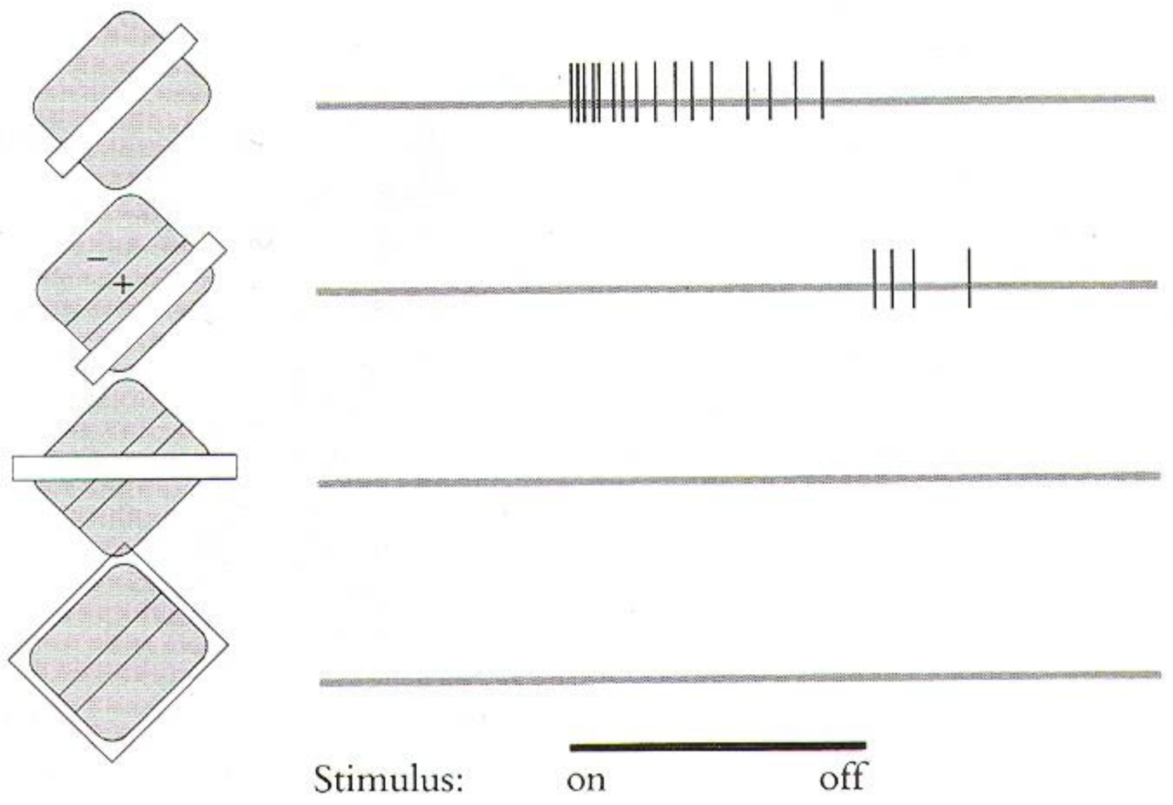


The 3R's of Vision: Recognition, Reconstruction & Reorganization



Talk at POCV Workshop, CVPR 2012

Hubel and Wiesel (1962) discovered orientation sensitive neurons in V1



Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

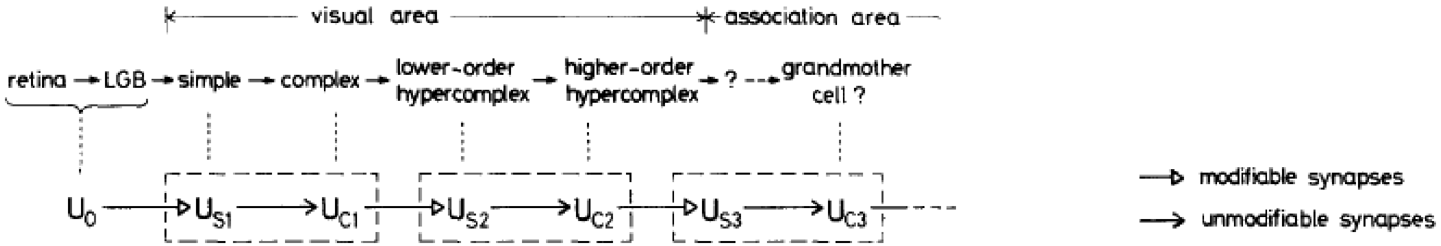


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

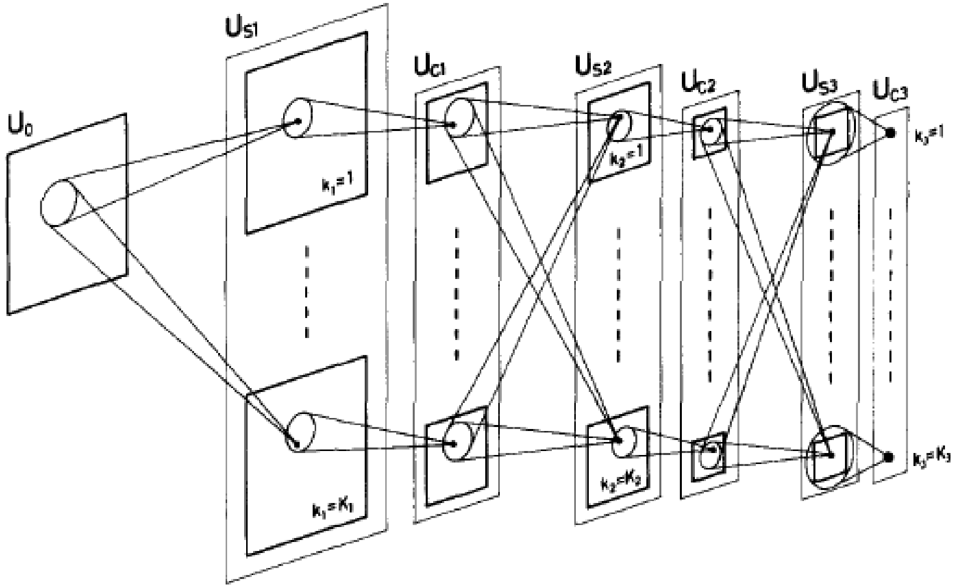


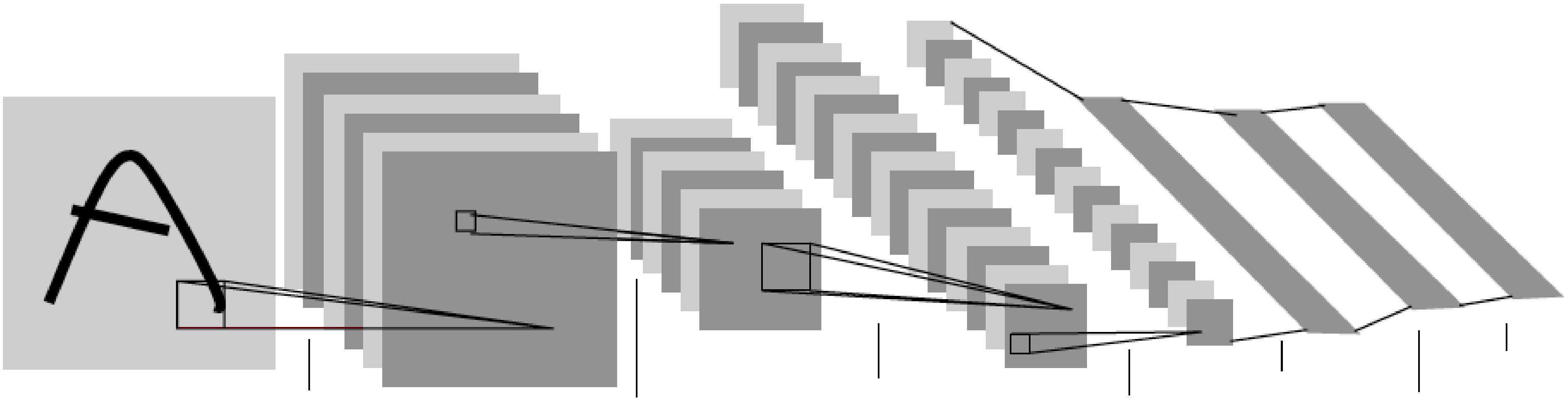
Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Biol. Cybernetics 36, 193–202 (1980)

Convolutional Neural Networks (LeCun et al)

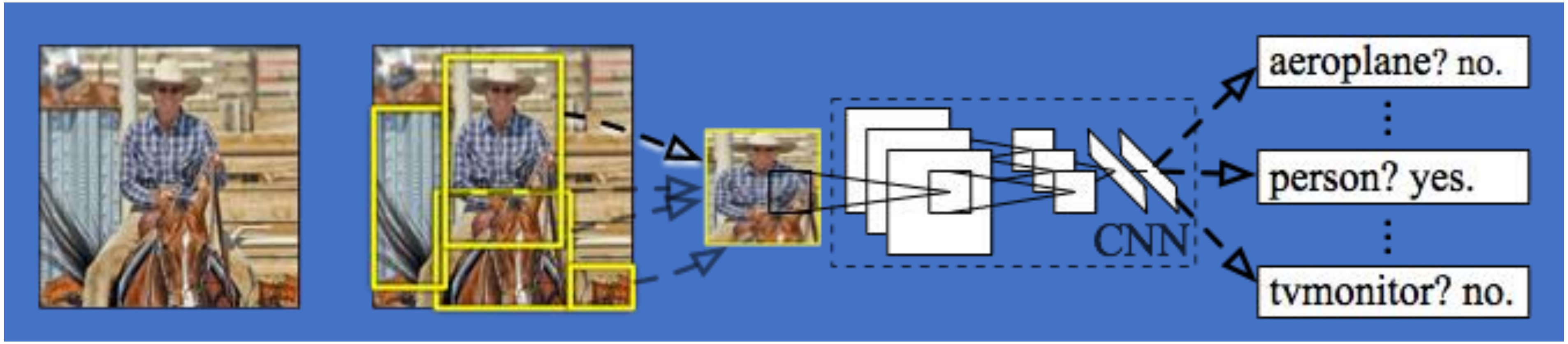
Used backpropagation to train the weights in this architecture

- First demonstrated by LeCun et al for handwritten digit recognition(1989)
- Applied in sliding window paradigm for tasks such as face detection in the 1990s.
- However was not competitive on standard computer vision object detection benchmarks in the 2000s.
- Thanks to availability of faster computing (GPUs) and large amounts of labeled data (Imagenet) we have seen an amazing renaissance led by Krizhevsky, Sutskever & Hinton (2012)



R-CNN: Regions with CNN features

Girshick, Donahue, Darrell & Malik (CVPR 2014)



Input image

Extract region proposals (~2k / image)

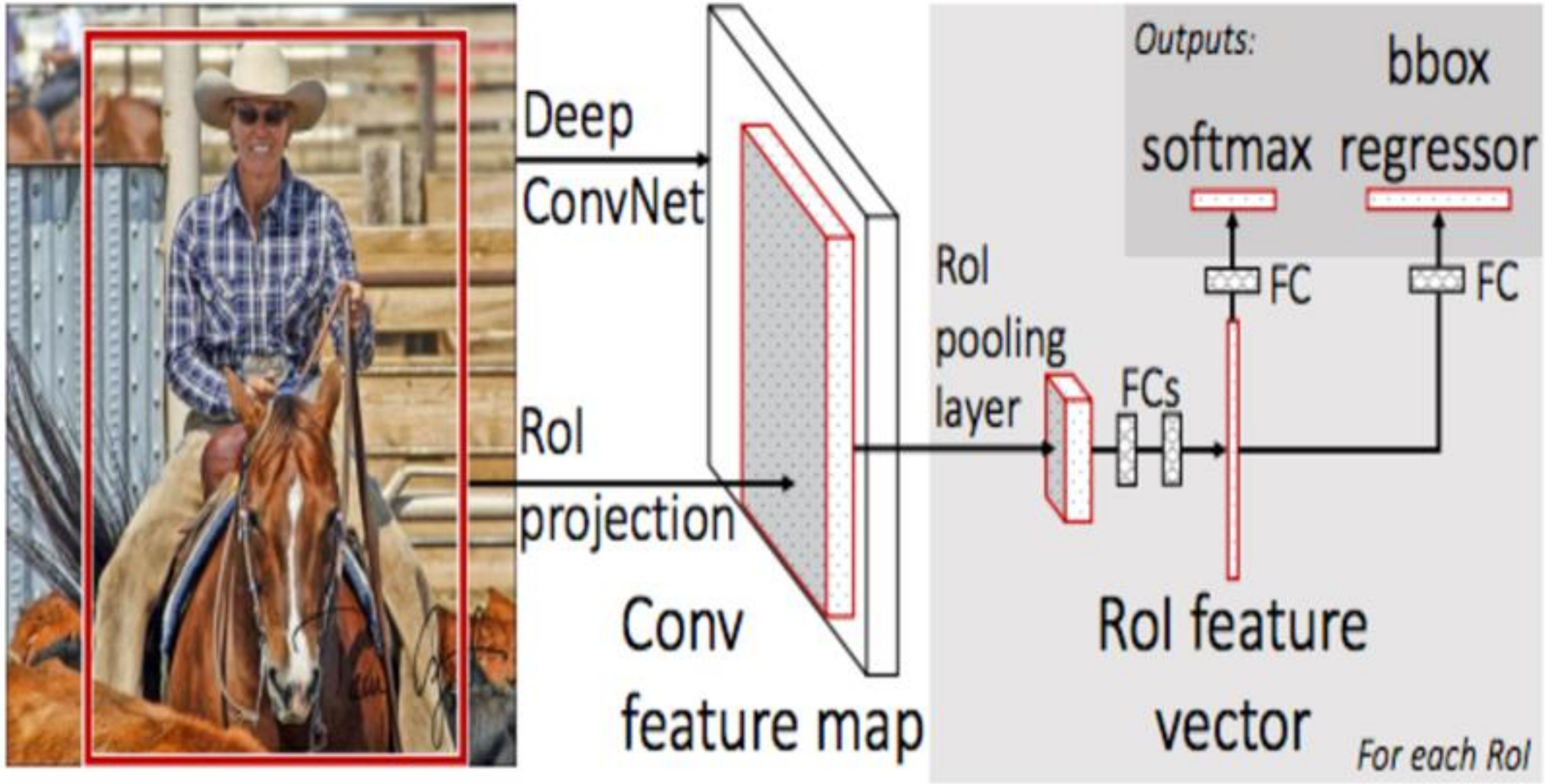
Compute CNN features

Classify regions (linear SVM)

This and the Multibox work from Google showed how to apply these architectures for object detection

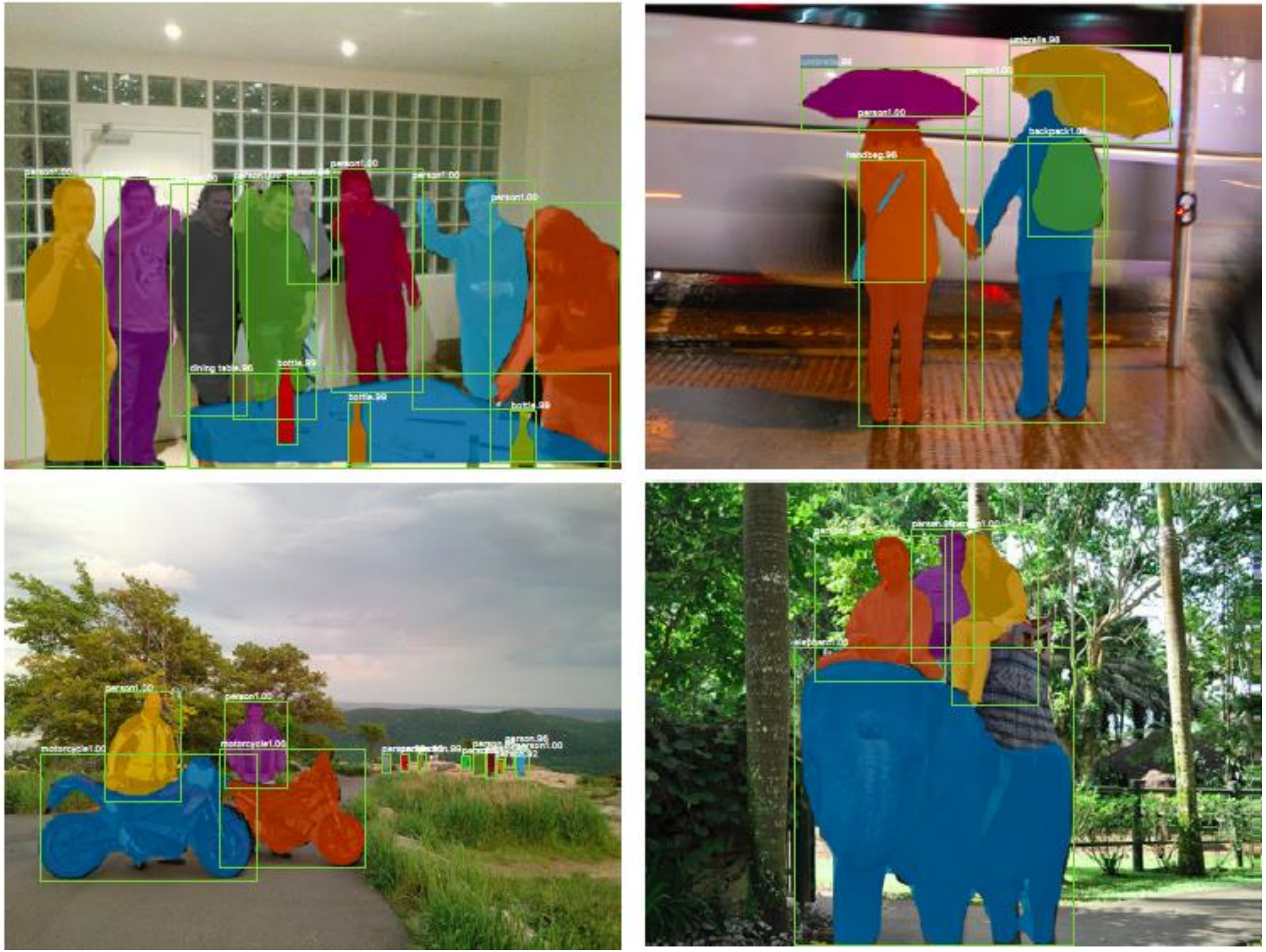
Fast R-CNN (Girshick, 2015)

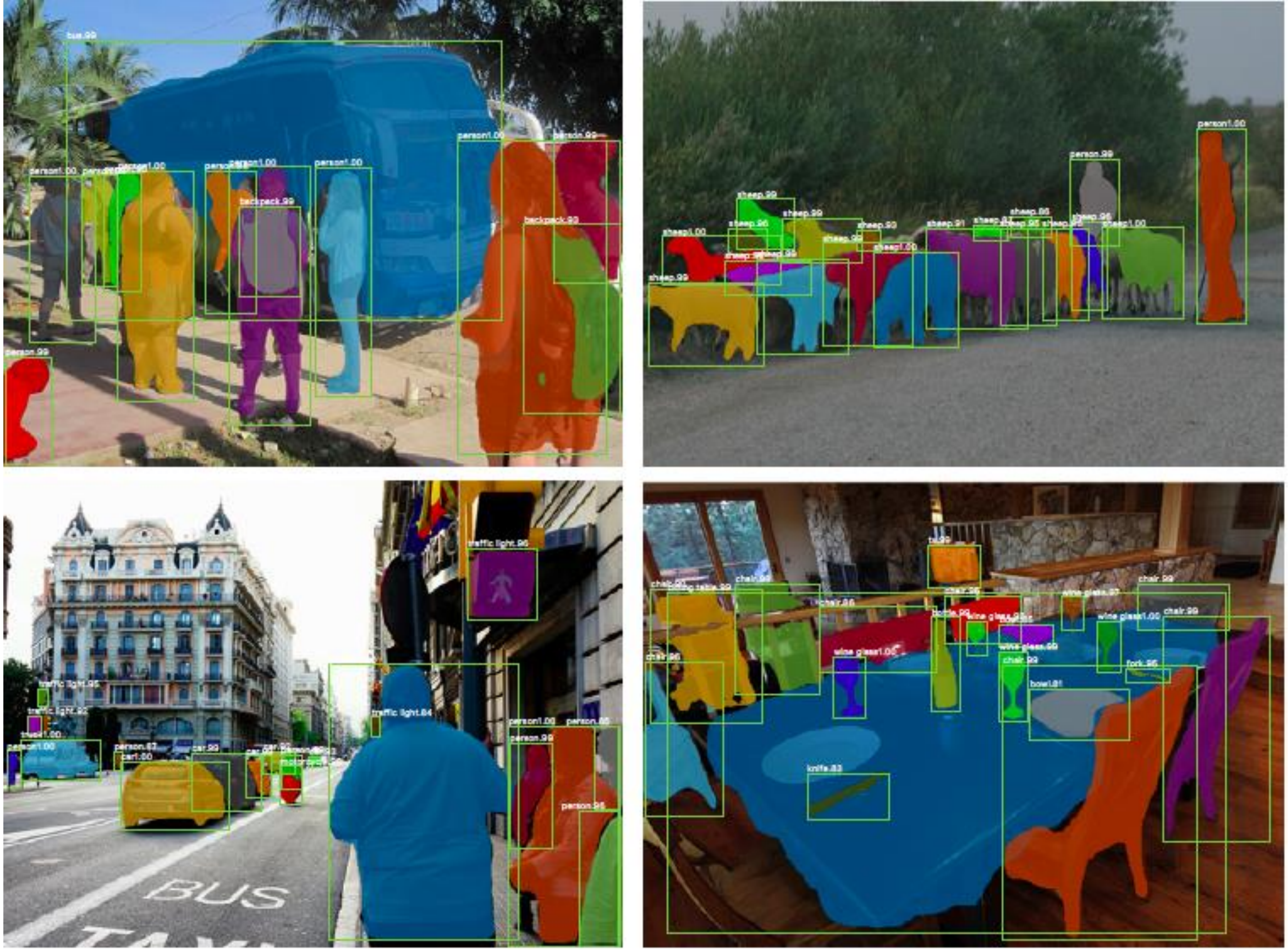
R-CNN with SPP features, no need to warp individual windows



Faster R-CNN doesn’t require external proposals & Mask R-CNN also produces instance segmentations

Mask R-CNN : He, Gkioxari, Dollar & Girshick (2017)




bus.99
person1.00
backpack.99
person.99
sheep1.00
sheep.99
traffic light.96
traffic light.84
car1.00
BUS
chair.99
dining table.99
wine glass1.00
bottle.99
bowl.81
fork.96
knife.83

We have some very good models for image analysis

- For recognition tasks, there has been continual improvement with Mask R-CNN, Detectron (Facebook), CLIP and GPT-4v (Open AI) showing amazing capabilities
- For reorganization (segmentation) tasks, Meta's SAM model dominates (it is freely available)
- For 3D reconstruction, we have progress, but still much to do.

PIETRO

Please name tge objects you see.

CHATGPT

I see the following objects in the image:

1. Wooden table
2. White coffee cup with a saucer
3. Clear glass water bottle
4. Clear glass filled with water
5. Metal milk pitcher
6. Wicker-wrapped sugar packet or biscuit
7. Wooden chairs
8. Metal and rope barrier posts
9. Trees
10. Fallen leaves on the ground
11. Outdoor café seating with people
12. Yellow meta' ↓ st (possibly a part of an umbrel or awning)

PIETRO

Do you see a spoon?

CHATGPT

Yes, there appears to be a spoon on the saucer next to the white coffee cup. My apologies for missing that in the initial list.

PIETRO

Is there more water in the cup, in the glass or in the bottle?

CHATGPT

Based on the image, it appears that there's more water in the bottle, followed by the glass, and the least amount in the cup (which seems to contain coffee or another beverage rather than water).

PIETRO

Describe this picture

CHATGPT

The picture depicts a classical-style painting with three female figures set against a scenic background.

- The female figure on the left is reclining on a rocky terrain, her body positioned sideways while she rests her head on one arm. Her long hair flows down, and she wears a draped white garment that covers most of her body.
- The central figure stands with her back to the viewer. She is holding up a long, flowing piece of cloth or hair. She is draped i similar white

Message

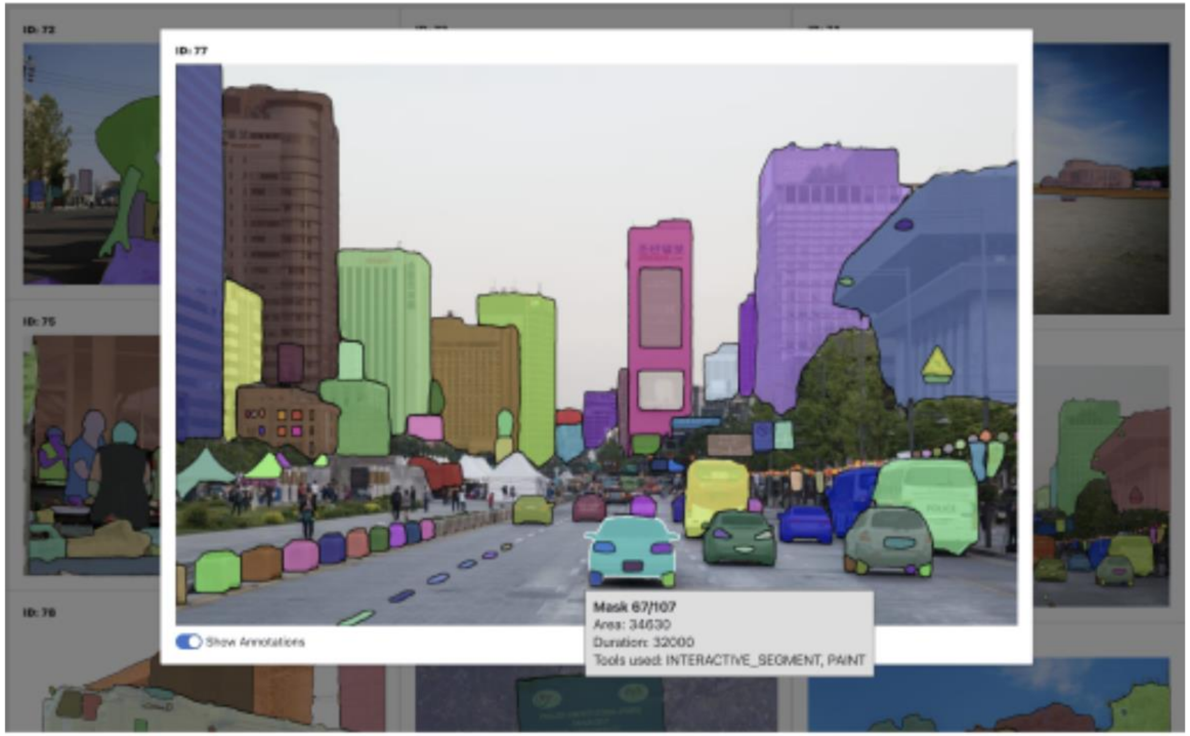
Courtesy of Pietro Perona


AI Computer Vision Research
Segment Anything Model (SAM): a new AI model from Meta AI that can "cut out" any object, in any image, with a single click
SAM is a promptable segmentation system with zero-shot generalization to unfamiliar objects and images, without the need for additional training.
Try the demo

11M images, 1B+ masks

After annotating enough masks with SAM's help, we were able to leverage SAM's sophisticated ambiguity-aware design to annotate new images fully automatically. To do this, we present SAM with a grid of points on an image and ask SAM to segment everything at each point. Our final dataset includes more than 1.1 billion segmentation masks collected on ~11 million licensed and privacy preserving images.

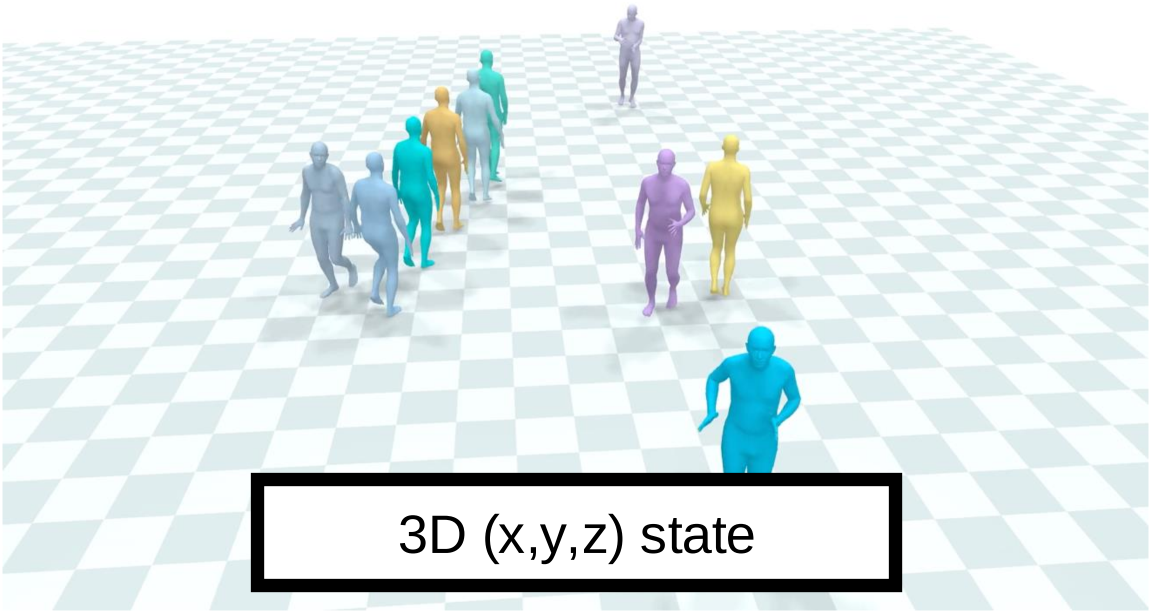
→ **Explore the dataset** → **Download full dataset**



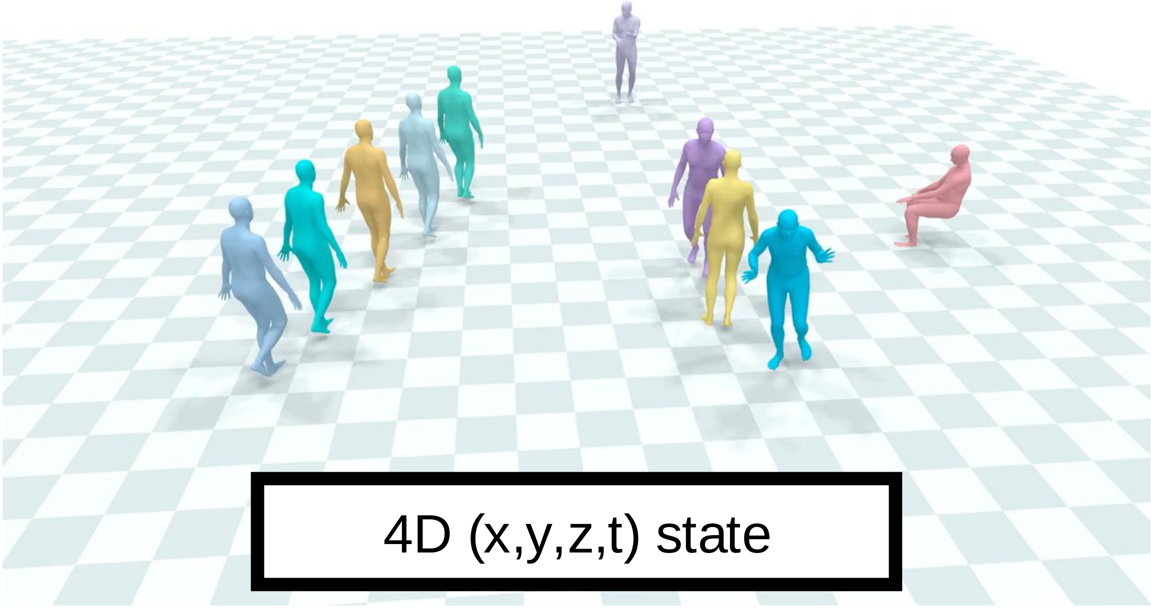
3D reconstruction

- 3D reconstruction of objects is a topic of vibrant research. NeRF models are very popular when multiple views are available of an object. However the problem of 3D reconstruction from a single view is still very much open.
- For the very important case of humans we now have quite powerful models


2D (x,y) observation


3D (x,y,z) state


3D (x,y,t) observation


4D (x,y,z,t) state


87% : ride
82% : sit
1% : stand

AVA (Atomic Visual Actions) labels

run/jog
walk
jump
stand
sit

lie/sleep
bend/bow
crawl
swim
dance

get up
fall down
crouch/kneel
martial art

Pose (14)

talk to
watch
listen to
sing to
kiss
hug
grab
lift
kick

give/serve … to …
take … from …
play with kids
hand shake
hand clap
hand wave
fight/hit
push

Person-person (17)

lift/pick up
put down
carry
hold
throw
catch
eat
drink
cut
hit
stir
press
extract
read
write

smoke
sail boat
row boat
fishing
touch
cook
kick
paint
dig
shovel
chop
shoot
take a photo
brush teeth
clink glass

work on a computer
answer phone
climb (e.g., mountain)
play board game
play with pets
drive (e.g., a car)
push (an object)
pull (an object)
point to (an object)
play musical instrument
text on/look at a cellphone
turn (e.g., screwdriver)
dress / put on clothing
ride (e.g., bike, car, horse)
watch (e.g., TV)

open
close
enter
exit

Person-object (49)

The Visual World

Ecological Optics (Gibson), Ethology (Tinbergen, Lorenz)

- The environment consists of the surrounds of animals. The environment is not the same as the physical world, if one means by that the world described by physics. Rather it is the relevant structure, at spatial and temporal scales corresponding to the animal.
- The terrestrial environment contains
    - Surfaces that have a certain layout
    - The Ground is the surface of the earth
    - Objects are persisting substances with closed or nearly closed surfaces. Could be detached or attached.
    - Enclosures are layouts of surfaces that surround the medium to some degree
    - Places that are locations in the environment. The habitat of an animal is made up of places.
    - ….

Taxonomy and Partonomy

- Taxonomy: E.g. Cats are in the order Felidae which in turn is in the class Mammalia
  - Recognition can be at multiple levels of categorization, or be identification at the level of specific individuals , as in faces.
- Partonomy: Objects have parts, they have subparts and so on. The human body contains the head, which in turn contains the eyes.
- These notions apply equally well to scenes and to activities.
- Psychologists have argued that there is a “basic-level” at which categorization is fastest (Eleanor Rosch et al).
- In a partonomy each level contributes useful information for recognition.

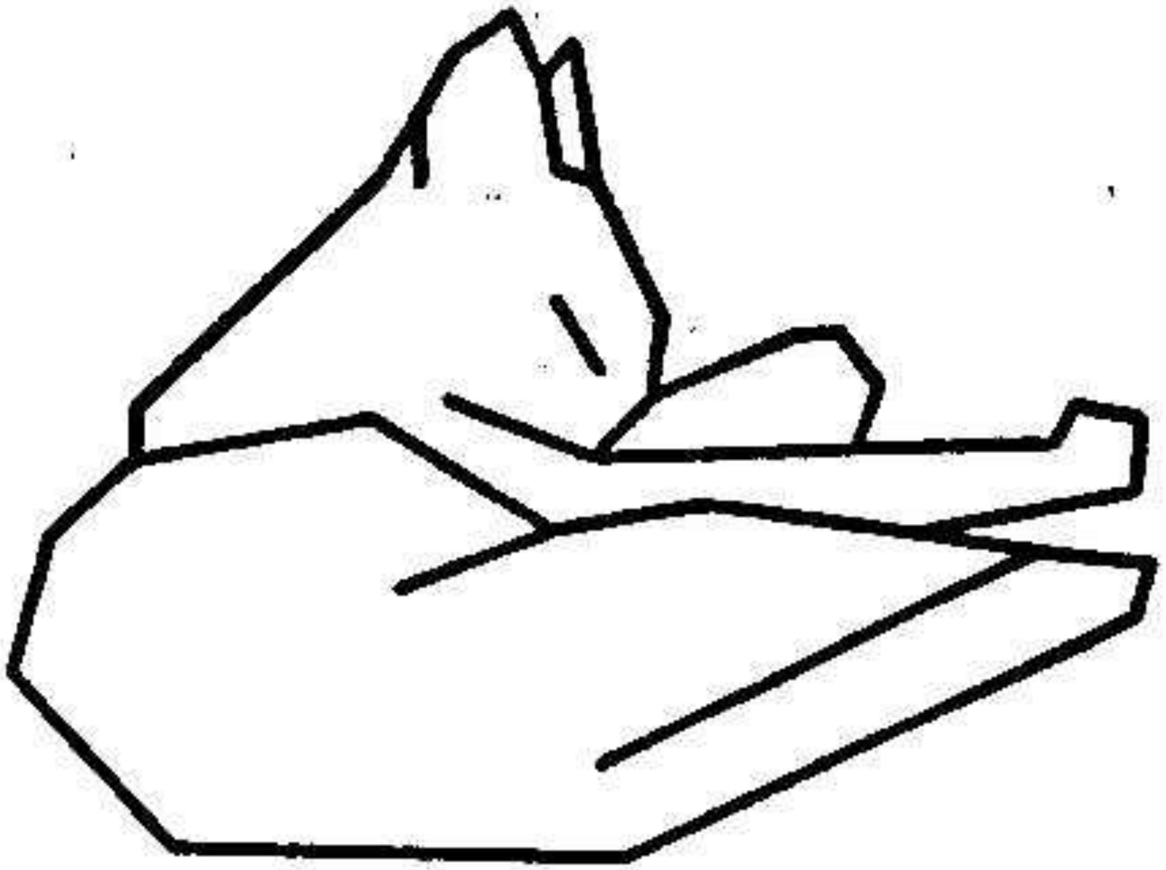
Functions of Visual Recognition

- Identifying conspecifics and their behavior
  - Friends, foes, mates, offspring, dominance hierarchies…
  - Emotions, gestures, actions…
- Recognizing environmental features for navigation
  - Generic terrain features e.g. obstacles, water margins, brinks of cliffs..
  - Specifically remembered landmarks for way-finding
- Recognizing predators, prey, edible or non-edible, tools and other cultural artifacts…

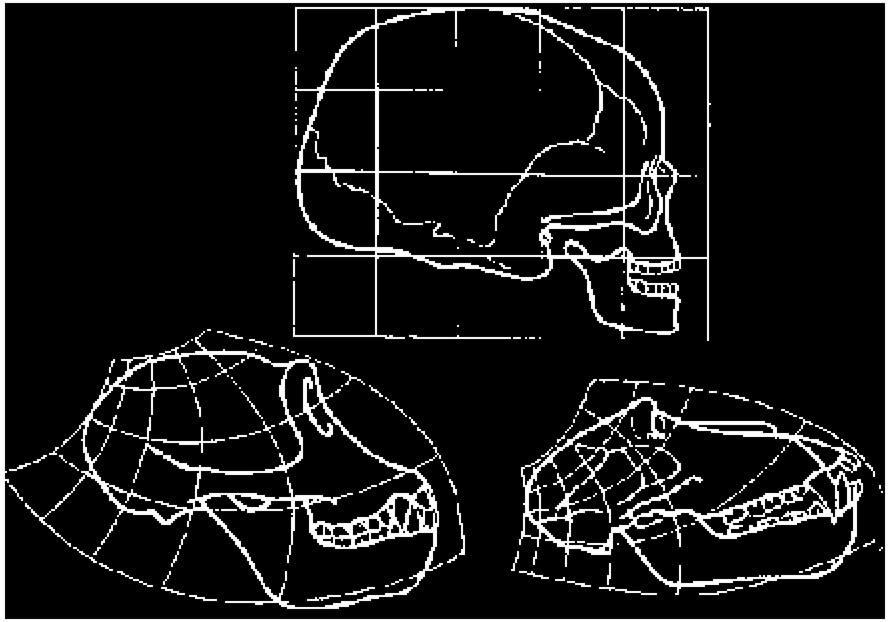
Two aspects of object recognition

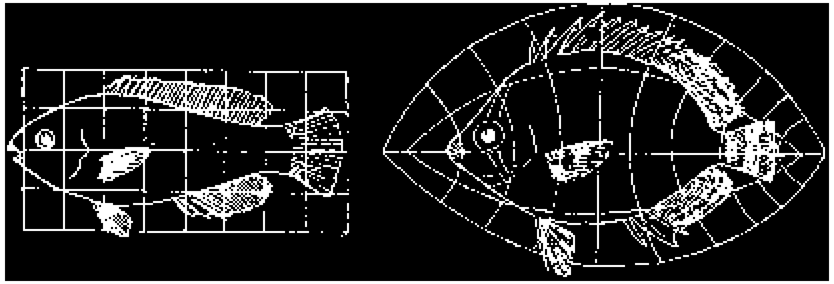
- Shape
- Texture

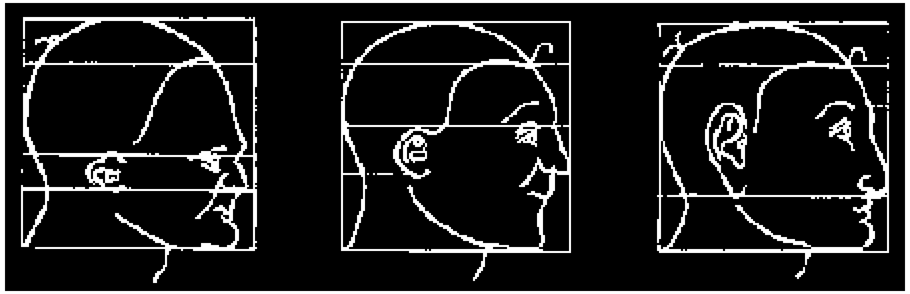
Attneave's Cat (1954)
Line drawings convey most of the information



Modeling shape variation in a category

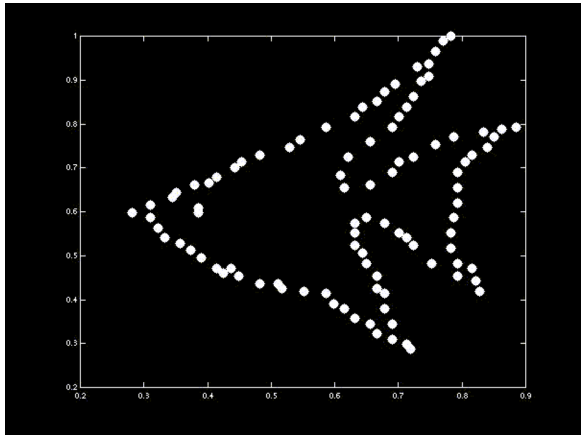






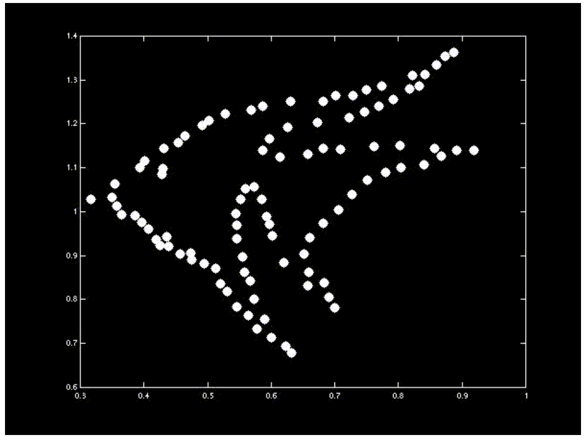
- D’Arcy Thompson: *On Growth and Form*, 1917
  - studied transformations between shapes of organisms

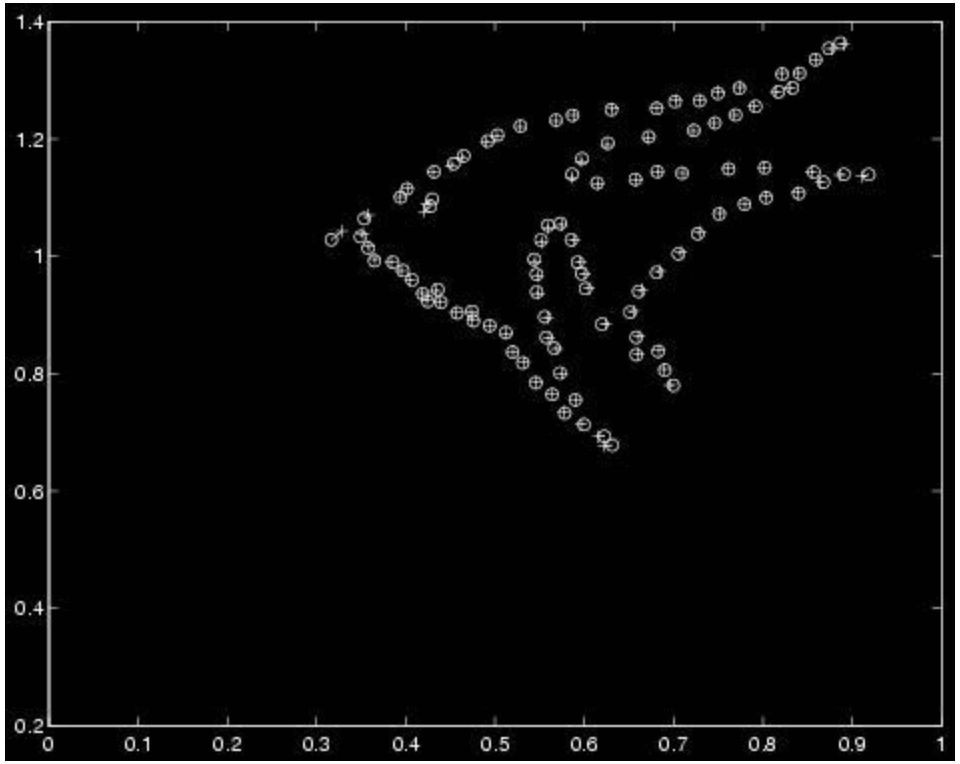
Matching Example

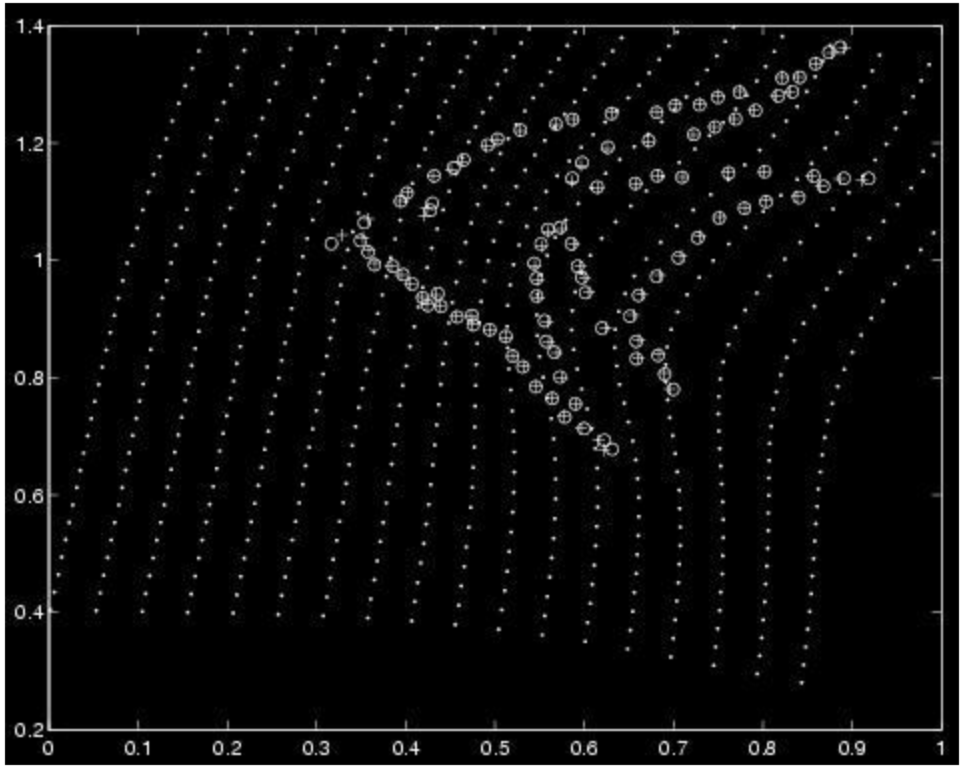


model

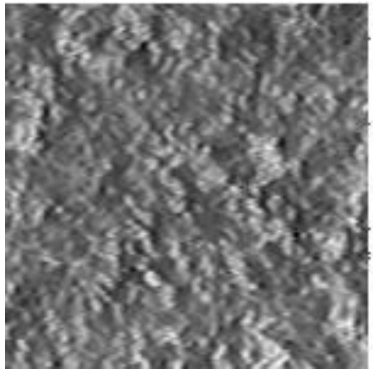
target



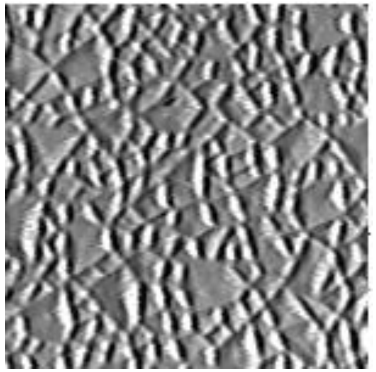




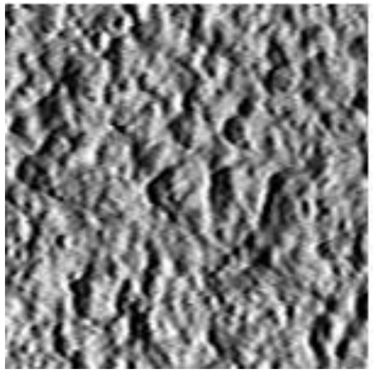
Example Natural Materials



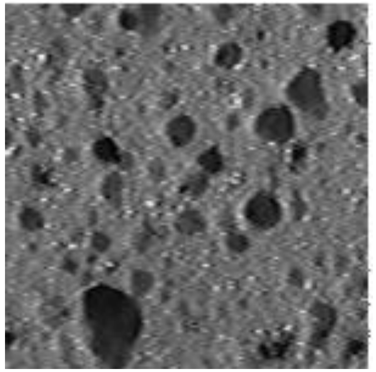
Terrycloth



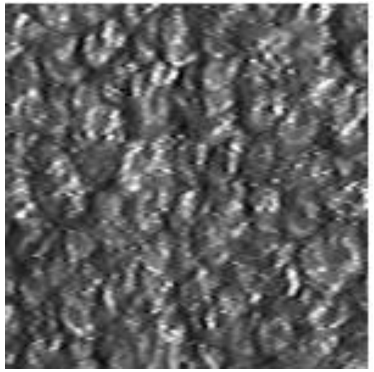
Rough Plastic



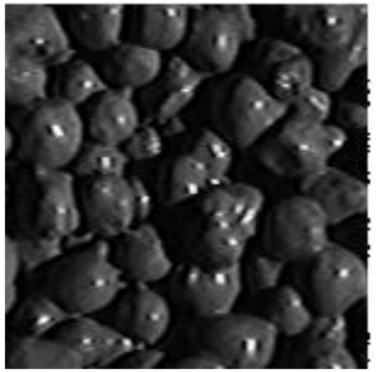
Plaster-b



Sponge



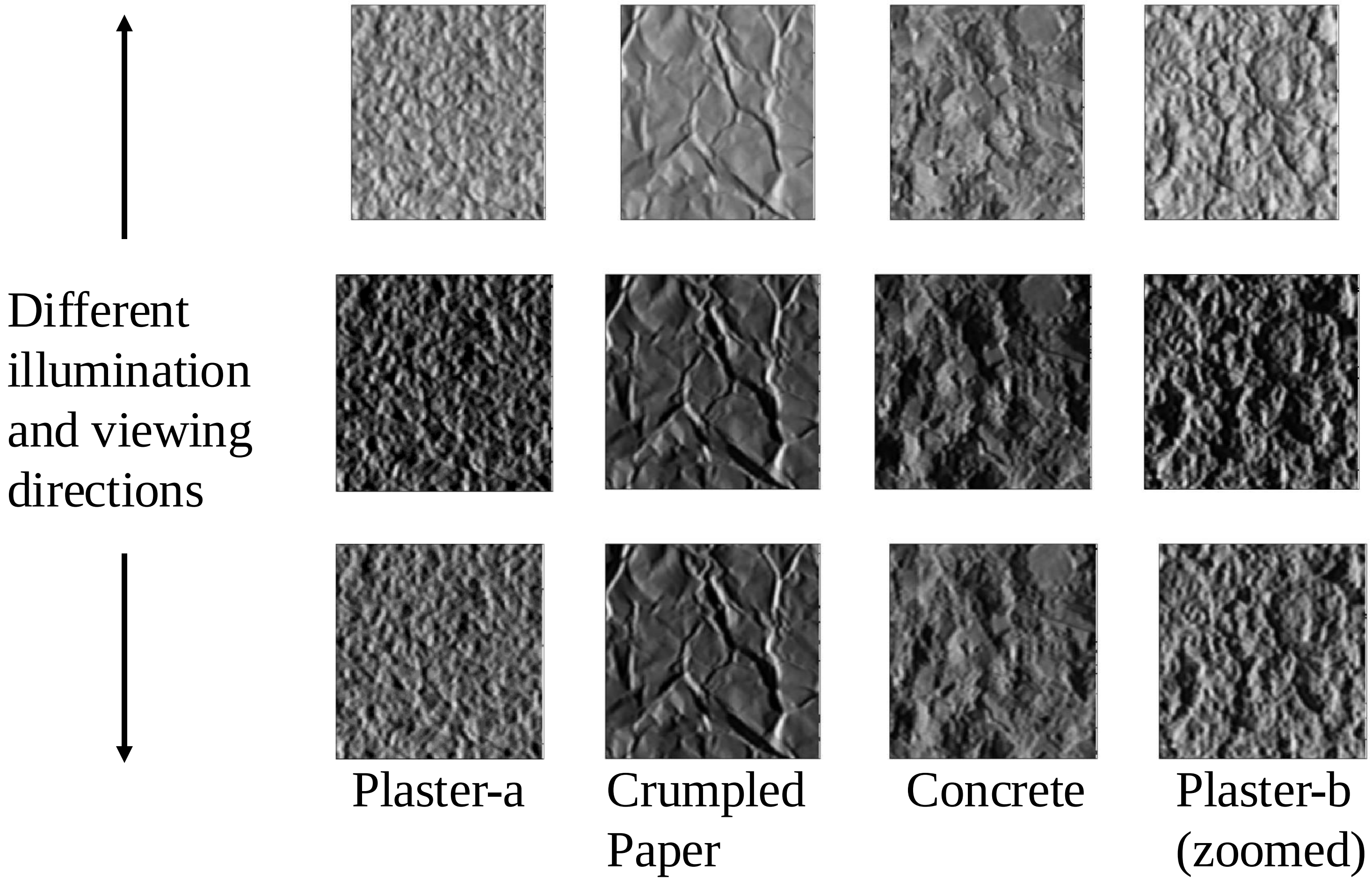
Rug-a



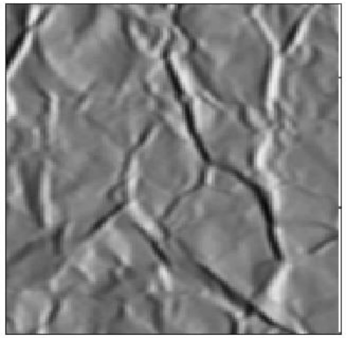
Painted Spheres

Columbia-Utrecht Database (http://www.cs.columbia.edu/CAVE)

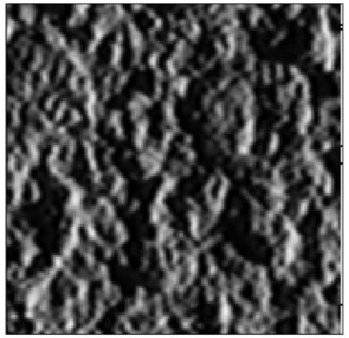
Materials under different illumination and viewing directions



Texture Recognition



?



?

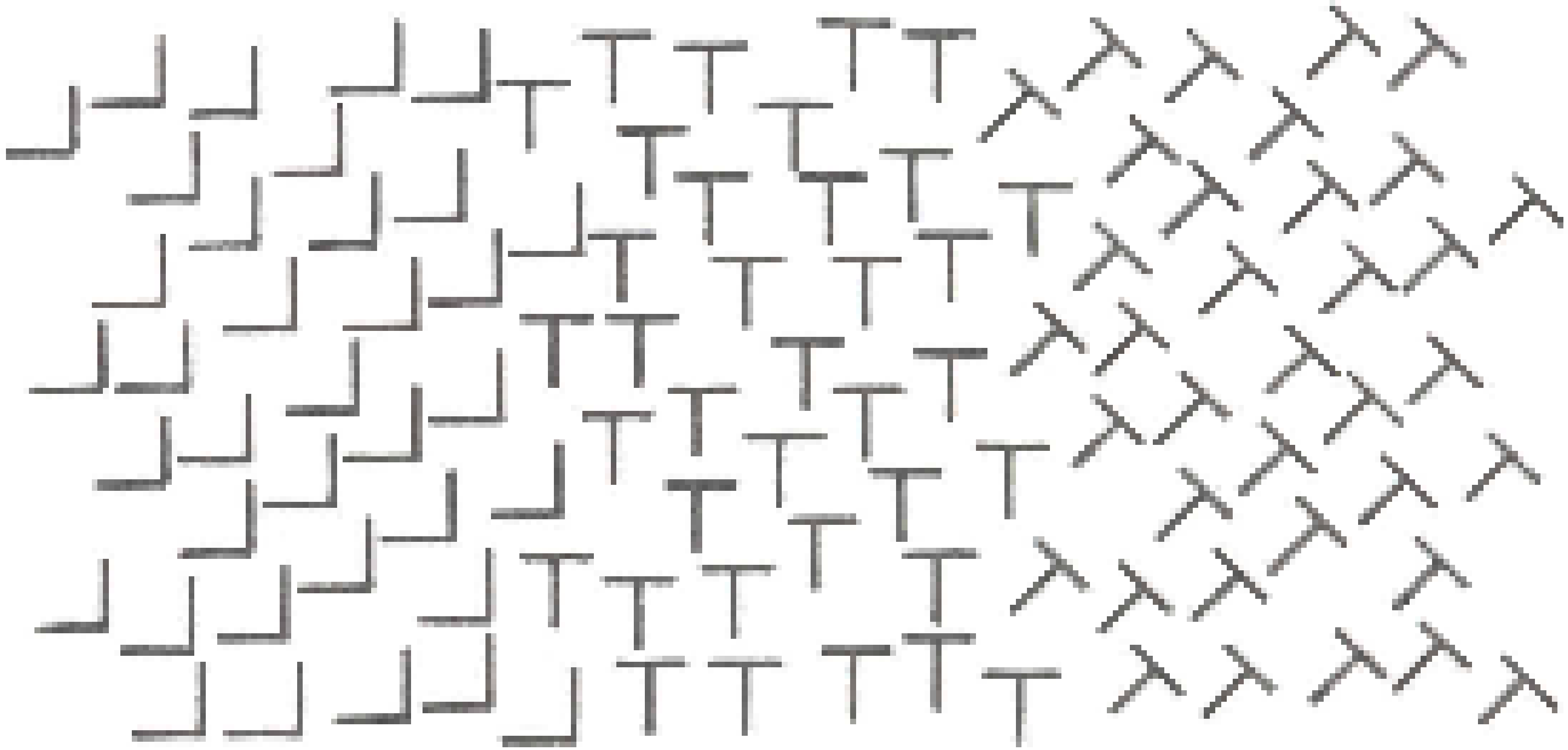
Felt?
Polyester?
Terrycloth?
Rough Plaster?
Leather?
Plaster?
Concrete?
Crumpled Paper?
Sponge?
Limestone?
Brick?

Julesz's texton theory

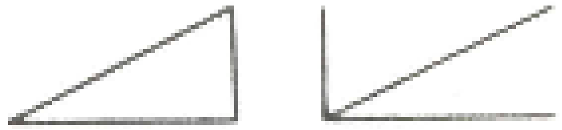
1. Human Vision operates in two distinct modes
   1. Pre-attentive vision - parallel, instantaneous, without scrutiny, independent of the number of patterns
   2. Attentive vision - serial search by focal attention in 50 ms steps limited to a small aperture as in form recognition

2. Textons are
   1. Elongated blobs - e.g. rectangles, ellipses, line segments with specific orientations, widths and lengths
   2. Terminators - ends of line segments
   3. Crossings of line segments

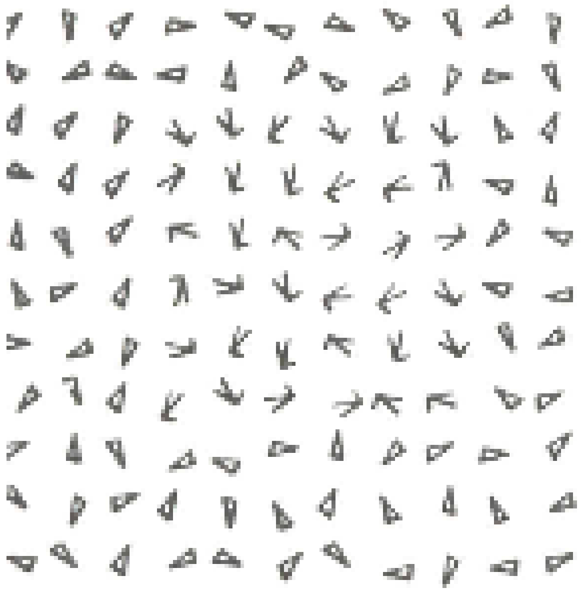
Orientation is a texton



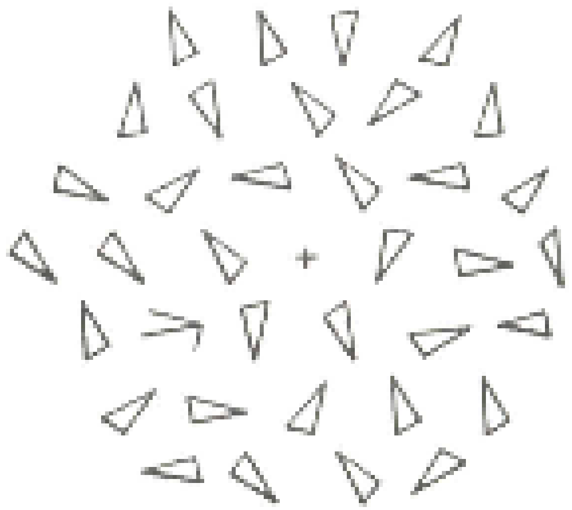
Terminators are textons



(a)



(b)

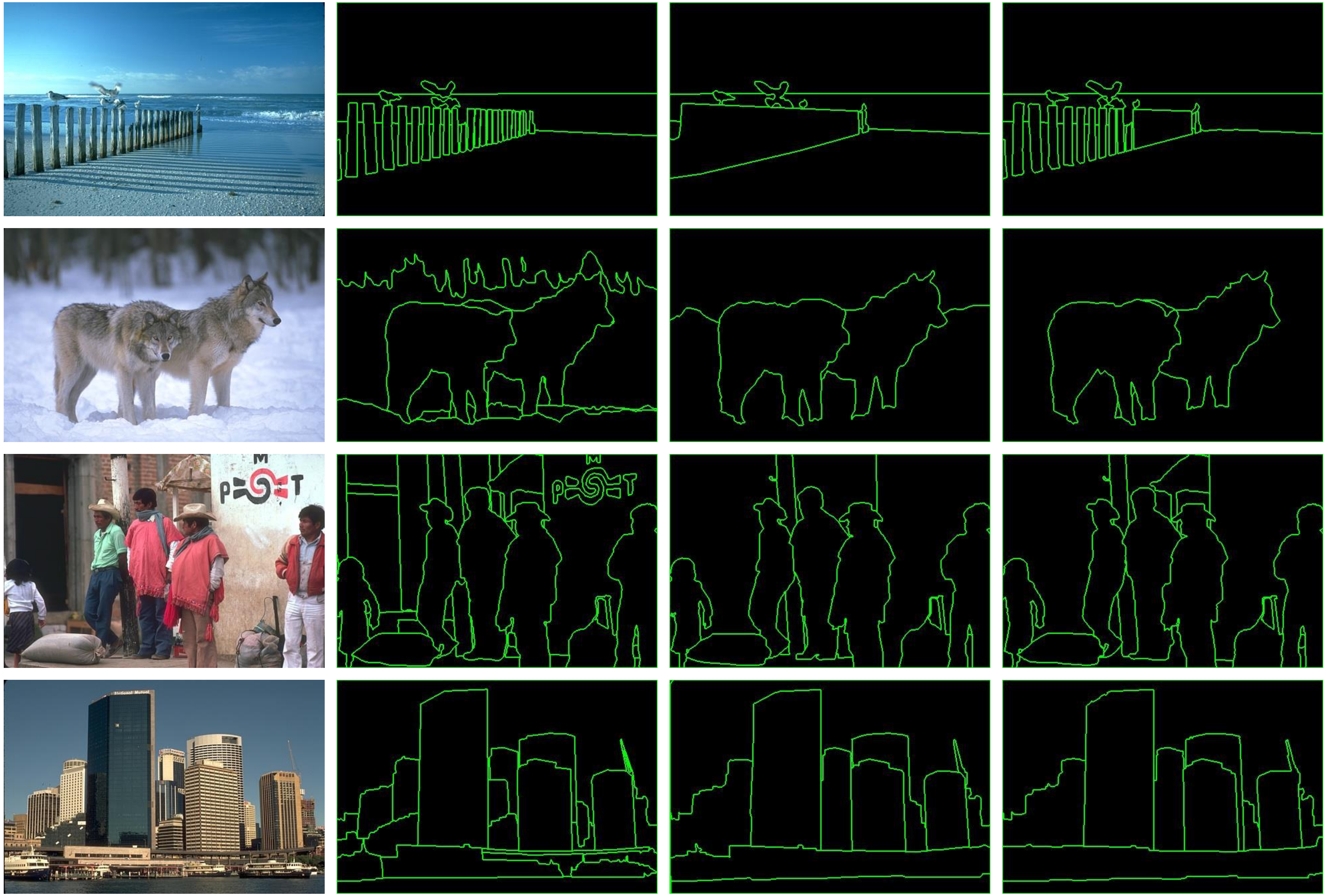


(c)

Boundaries of image regions defined by a number of cues

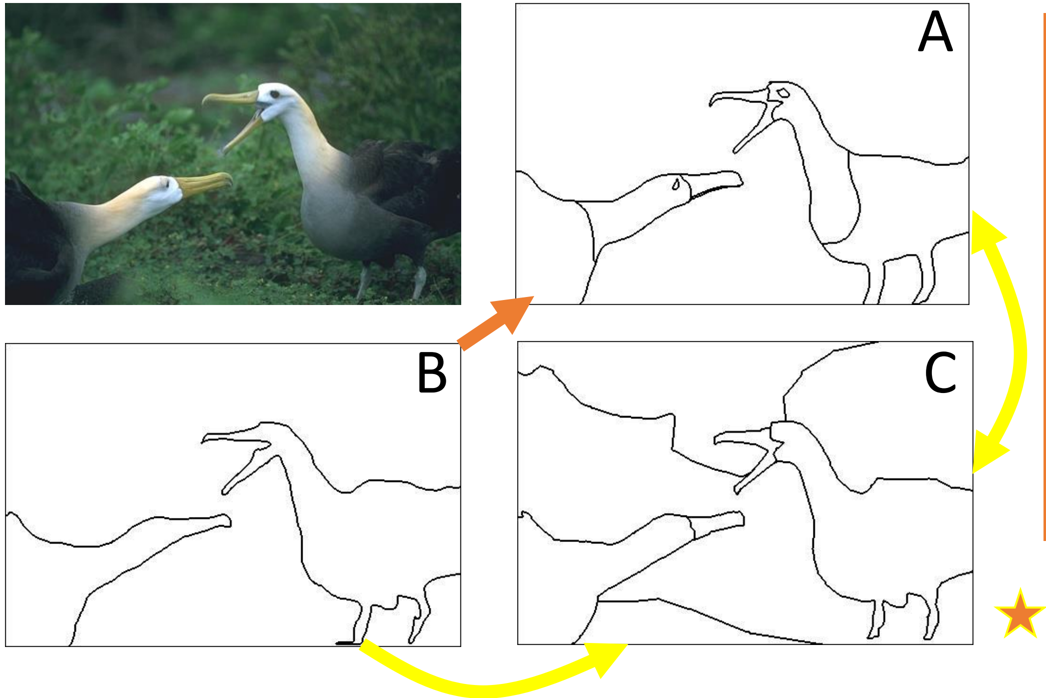
- Brightness
- Color
- Texture
- Motion (in video)
- Binocular Diparity (if available)
- Familiar objects

Berkeley Segmentation DataSet [BSDS]

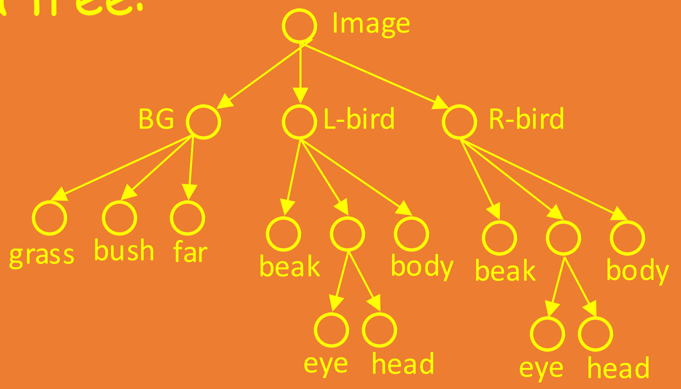


D. Martin, C. Fowlkes, D. Tal, J. Malik. "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics", ICCV, 2001

Consistency

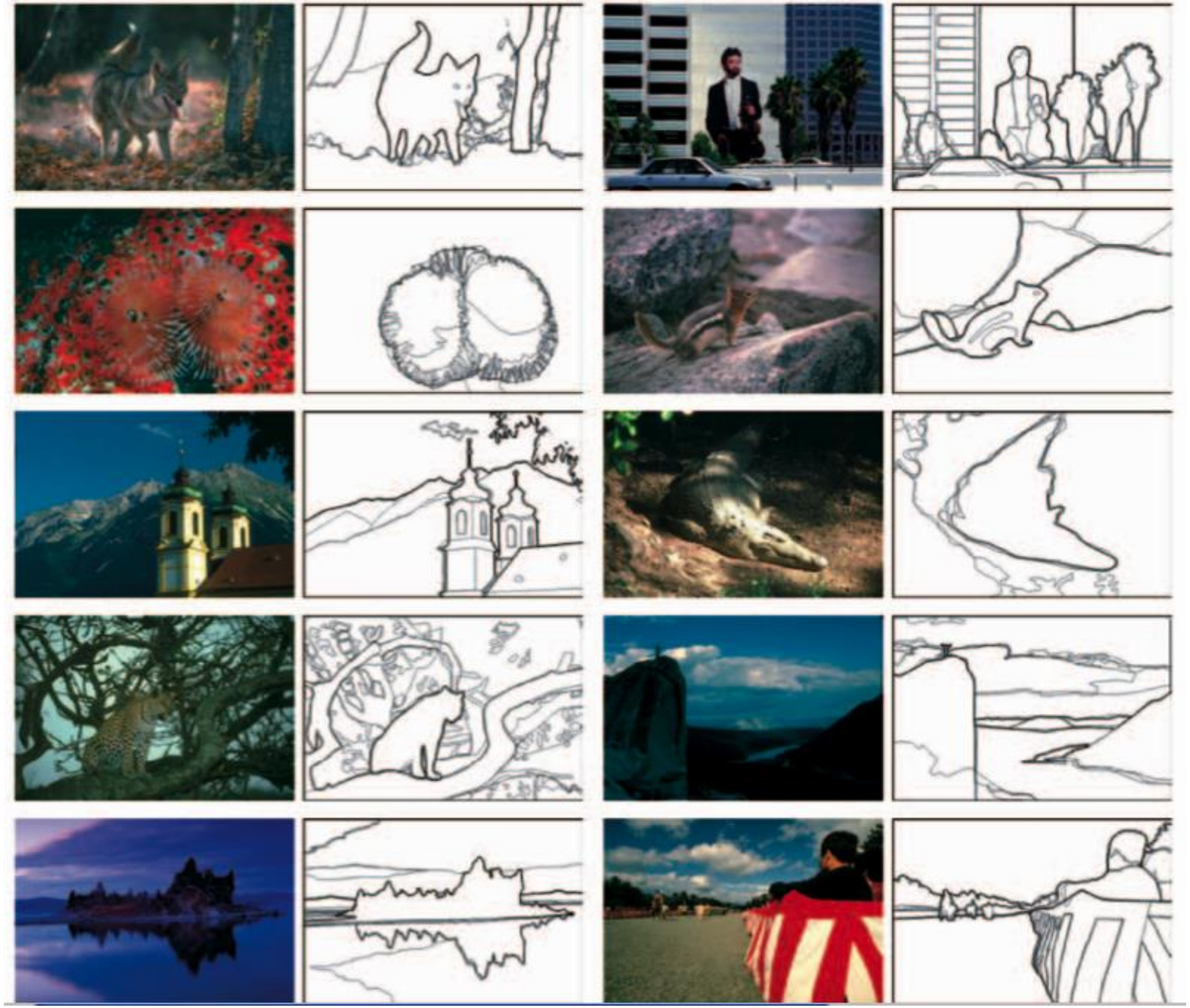


Perceptual organization forms a tree:



★ Two segmentations are consistent when they can be explained by the same segmentation tree (i.e. they could be derived from a single perceptual organization).

- A,C are refinements of B
- A,C are mutual refinements
- A,B,C represent the same percept
    - <u>Attention</u> accounts for differences



Insights from Phylogeny: Camouflage and Breaking Camouflage

- Evolutionary “arms race” between prey, evolving camouflage, and predators evolving recognition abilities

1. Camouflage strategies (crypsis, disruptive coloration, countershading, shadow elimination) suggest that grouping and figure ground inference play a key role in recognition in the wild; else top down template matching could always work
2. Vision with controlled eye movements as important as snapshot vision