GNR-638: Deep learning for image analysis

Prof. Biplab Banerjee
Associate Professor,
CSRE & C-MInDS,
Web: biplab-banerjee.github.io

Many of the slides credit to Internet

Today’s agenda

- ✓What are we going to learn in computer vision?
- ✓Vision datasets
- ✓Vision tasks
- ✓Challenges in visual recognition
- ✓History of the recognition tasks
- ✓Course administrative details

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

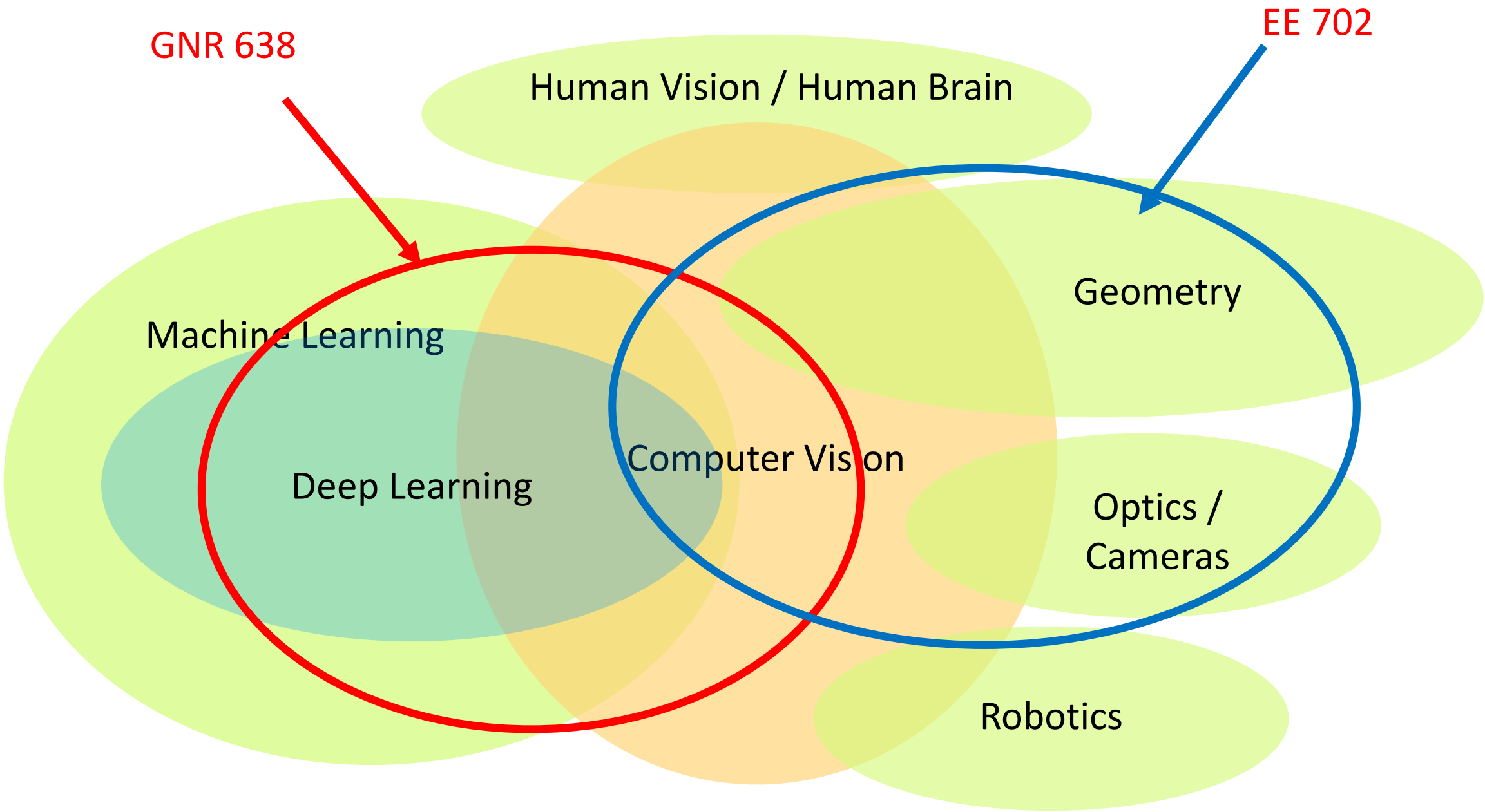
July 7, 1966

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

https://dspace.mit.edu/handle/1721.1/6125


GNR 638
EE 702
Human Vision / Human Brain
Machine Learning
Deep Learning
Computer Vision
Geometry
Optics / Cameras
Robotics

Computer Vision

Make computers understand images and video.

What kind of scene?

Where are the cars?

How far is the building?

…

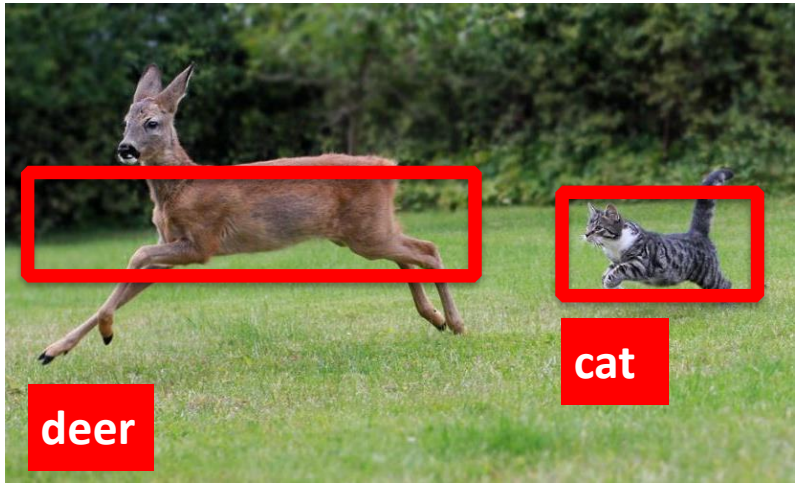
Requires multi-level processing, integrating visual and semantic cues

Relationship with Other Fields

- Image Processing: Image → Image

Relationship with Other Fields

- Computer Vision: Image → Knowledge

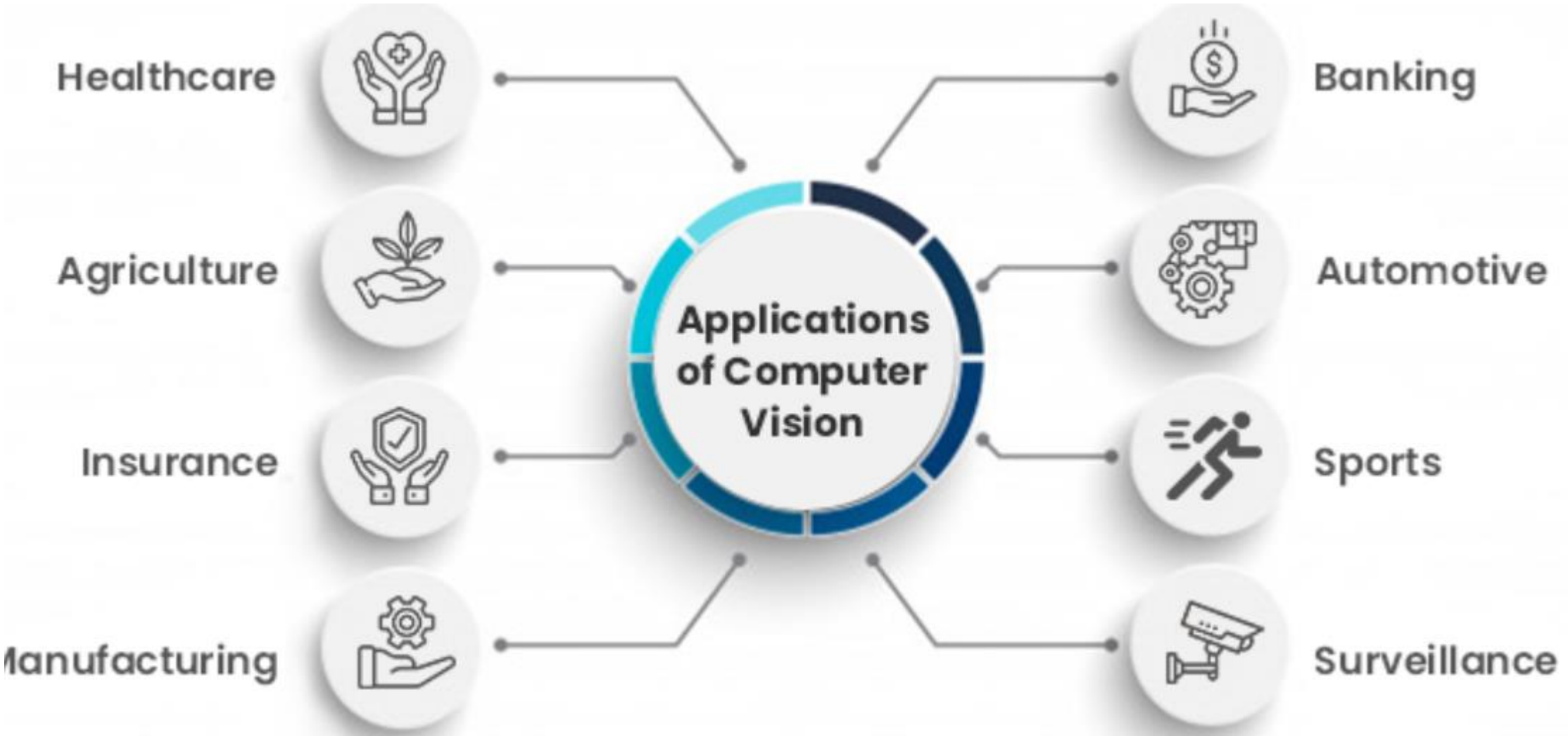


Relationship with Other Fields

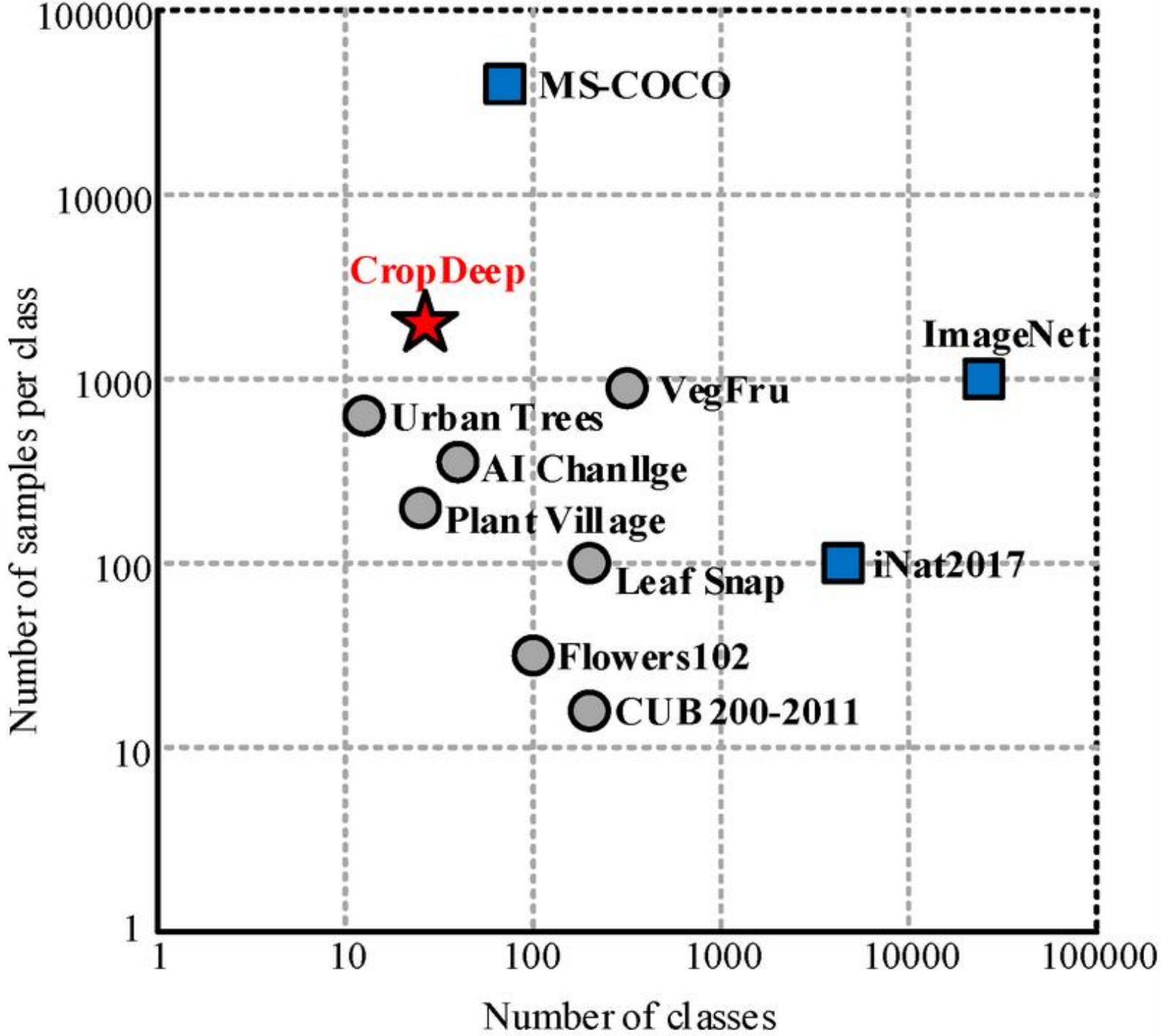
- Computer Graphics: Knowledge → Image

Vertices, Locations, Objects,
Shapes, Colors, Material properties,
Lighting settings, Camera settings, etc.

→


Healthcare
Agriculture
Insurance
Manufacturing
Applications of Computer Vision
Banking
Automotive
Sports
Surveillance


The European Camera of the Year 1983.
Fortunately, fine photography is a universal language.
PENTAX
PENTAX


100000
10000
1000
100
10
1
Number of samples per class
1
10
100
1000
10000
100000
Number of classes
MS-COCO
CropDeep
ImageNet
VegFru
Urban Trees
AI Chanllge
Plant Village
iNat2017
Leaf Snap
Flowers102
CUB 200-2011

ImageNet dataset (wiki)

ImageNet-21K [edit]

The full original dataset is referred to as ImageNet-21K. ImageNet-21k contains 14,197,122 images divided into 21,841 classes. Some papers round this up and name it ImageNet-22k.[29]

The full ImageNet-21k was released in Fall of 2011, as `fall11_whole.tar`. There is no official train-validation-test split for ImageNet-21k. Some classes contain only 1-10 samples, while others contain thousands.[29]

ImageNet-1K [edit]

There are various subsets of the ImageNet dataset used in various context, sometimes referred to as "versions".[16]

One of the most highly used subset of ImageNet is the "ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012–2017 image classification and localization dataset". This is also referred to in the research literature as ImageNet-1K or ILSVRC2017, reflecting the original ILSVRC challenge that involved 1,000 classes. ImageNet-1K contains 1,281,167 training images, 50,000 validation images and 100,000 test images.[30]

Each category in ImageNet-1K is a leaf category, meaning that there are no child nodes below it, unlike ImageNet-21K. For example, in ImageNet-21K, there are some images categorized as simply "mammal", whereas in ImageNet-1K, there are only images categorized as things like "German shepherd", since there are no child-words below "German shepherd".[22]

LAION-400-MILLION OPEN DATASET

by: Christoph Schuhmann, 20 Aug, 2021

We present LAION-400M: 400M English (image, text) pairs - see also our Data Centric AI NeurIPS Workshop 2021 paper

Concept and Content

The LAION-400M dataset is entirely openly, freely accessible.

WARNING: be aware that this large-scale dataset is non-curated. It was built for research purposes to enable testing model training on larger scale for broad researcher and other interested communities, and is **not** meant for any real-world production or application.

We have filtered all images and texts in the LAION-400M dataset with OpenAI's CLIP by calculating the cosine similarity between the text and image embeddings and dropping those with a similarity below 0.3. The threshold of 0.3 had been determined through human evaluations and seemed to be a good heuristic for estimating semantic image-text-content matching.

The image-text-pairs have been extracted from the Common Crawl web data dump and are from random web pages crawled between 2014 and 2021.

blue cat

Russian Blue Postcards (Package of 8)


CAT, LED Lamps, slingly, FamilyTrophy.com - Family...


Halloween, Cat, Paper Plate, Toddler, Preschool


Blue Wooden Cat


Ink sketch of a cat


Blue Eyes Greeting Card


Alicia Vannoy Call Framed Prints - Cat - Kitten Bl...

Blues for a Black Cat and Other Stories


Sitting cat low poly in Gloss Blue Porcelain

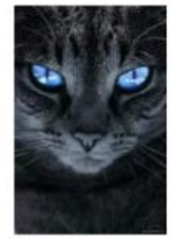

Tonkinese cat drawing BZTAT


Blue Cat Full Moon Postcard


Kitten Digital Art - Soft Kitty by William Paul M...


Blue and Black Panther Football Colors Acrylic Pai...


Tonkinese drawing - picture #10


Cat with blue eyes


Pretty Blue Cat cards

cat with blue eyes


cats with different colored through golden this ca...


cat, white, and eyes image


Siamese cat lying in the bast basket


Catito, Blue Eyes by RBenedetti


, Pam Pam, le Chat Blanc aux Yeux de 2 Couleurs qu...


heterochromia-cat-cross-eyed-alos-5


iPhone Wallpaper Two white kittens, blue eyes, pla...

5D Diamond Painting White Cat with Blue Eyes Kit


Coby the Cat with Piercing Blue Eyes, over 1 Milli...


heterochromia-cat-cross-eyed-alos-17

5D Diamond Painting White Cat with Blue Eyes Kit


ragdoll stock photo © nailiaschwarz


This Cat Has The Most Stunningly Beautiful Blue Ey...


A cat conquers the Internet with blue eyes


cat, white, and eyes image

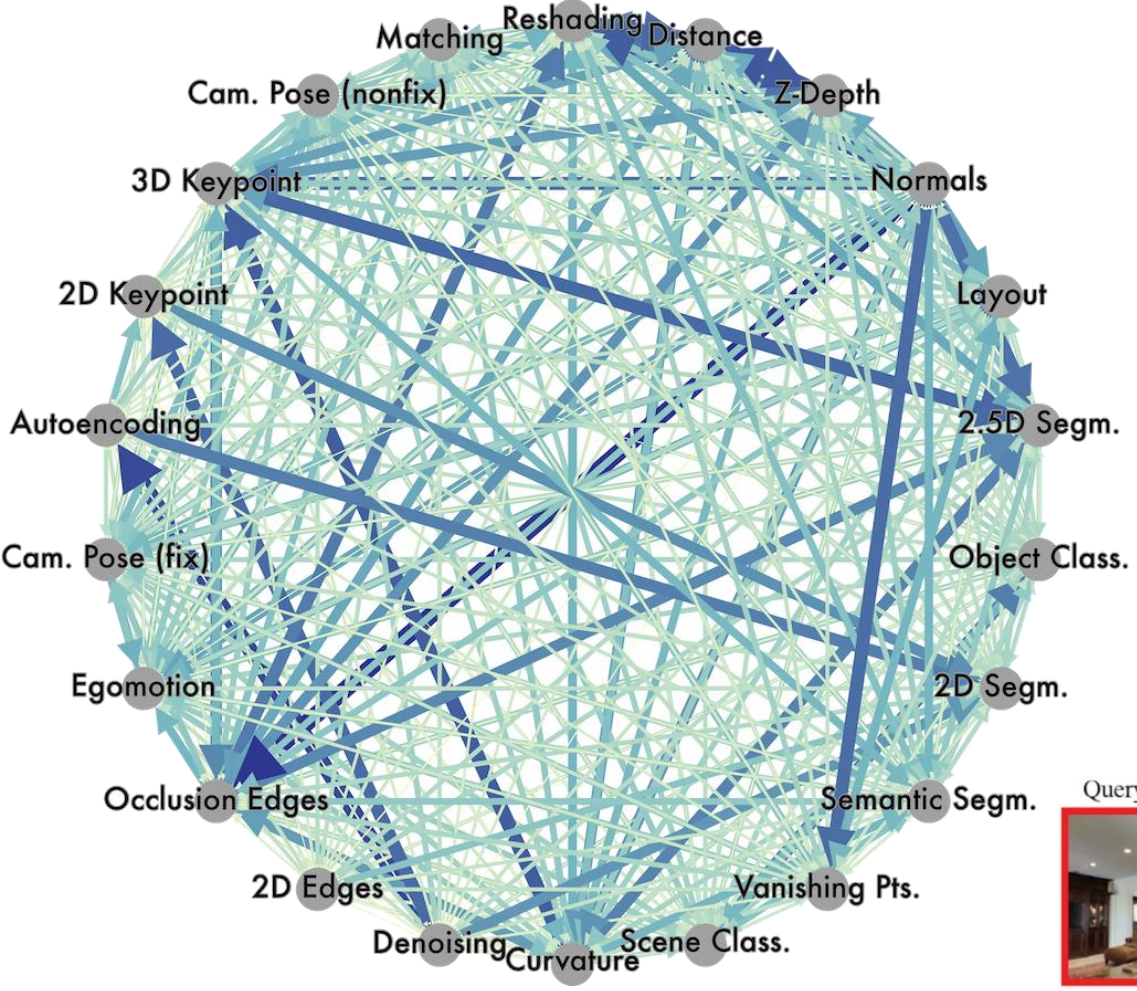

blue eyes and cute kitty image

Video recognition - ActivityNet

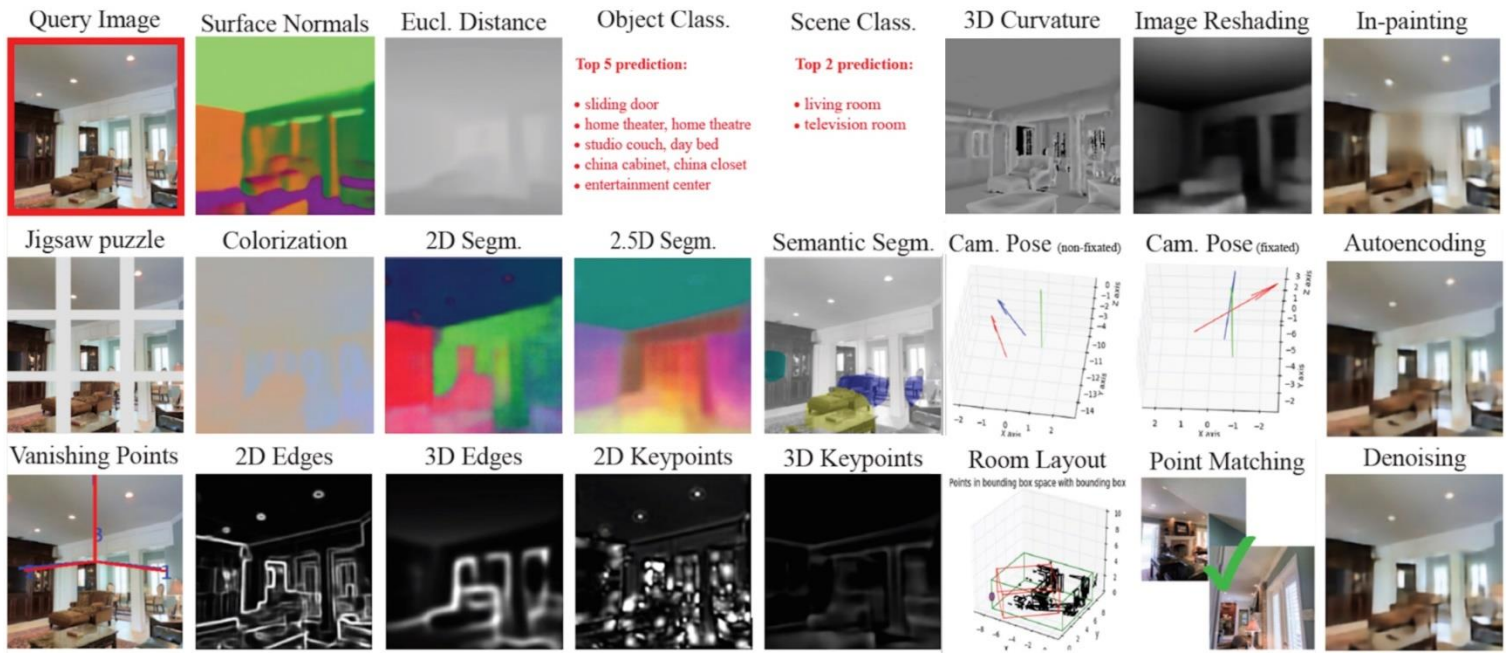
Figure 1. *ActivityNet* organizes a large number of diverse videos that contain human activities into a semantic taxonomy. **Top-row** shows the root-leaf path for the activity *Cleaning windows*. **Bottom-row** shows the root-leaf path for the activity *Brushing teeth*. Each box illustrates example videos that lie within the corresponding taxonomy node. Green intervals indicate the temporal extent of the activity. All figures are best viewed in color.

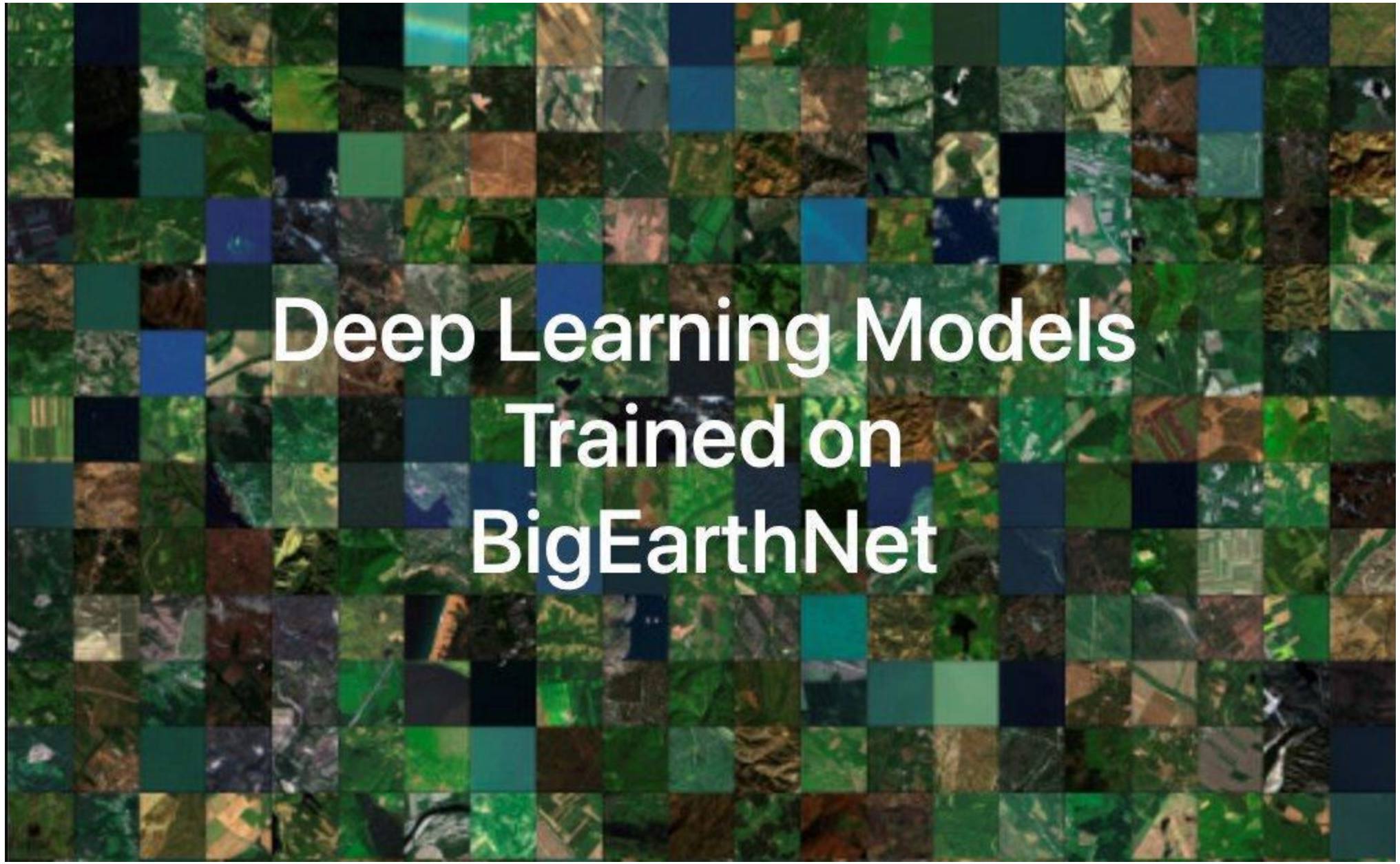
The daily action dataset

Figure 3. Frame and annotation samples of CDAD, including multi-person action (e.g. shake hands, hug), single person action (e.g. raise hand), human-object interaction action (e.g. smoke, sweep). Spatial-temporal annotations are provided for all actions. "Green" indicates temporal annotation of action instances, "blue" represents background (no action), and red rectangles are spatial annotations.

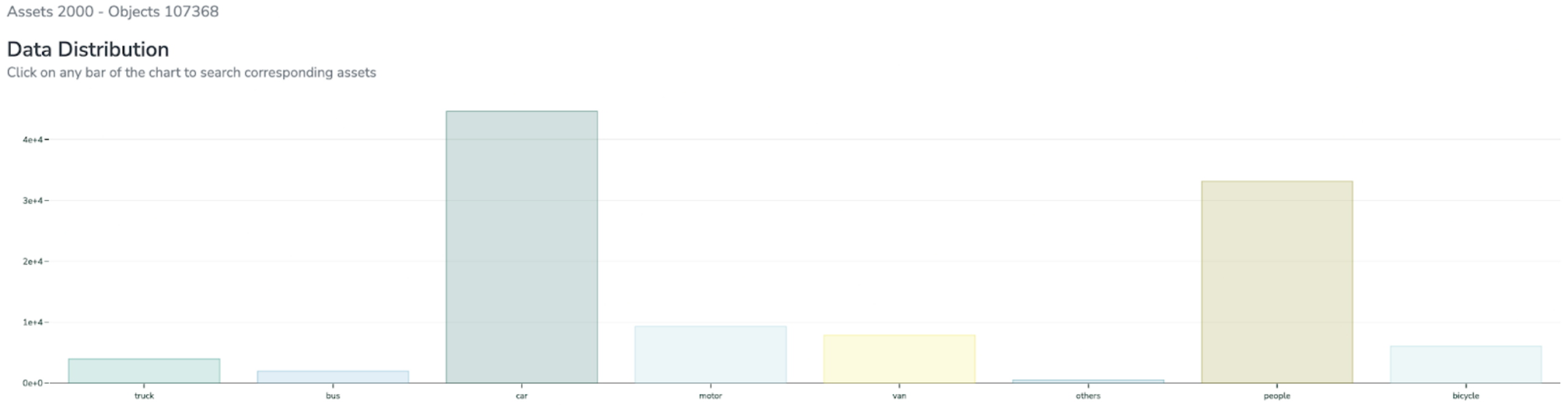

Reshading
Matching
Distance
Cam. Pose (nonfix)
Z-Depth
3D Keypoint
Normals
2D Keypoint
Layout
Autoencoding
2.5D Segm.
Cam. Pose (fix)
Object Class.
Egomotion
2D Segm.
Occlusion Edges
Semantic Segm.
2D Edges
Vanishing Pts.
Denoising
Curvature
Scene Class.

Taskonomy


Query Image
Surface Normals
Eucl. Distance
Object Class.
Top 5 prediction:
• sliding door
• home theater, home theatre
• studio couch, day bed
• china cabinet, china closet
• entertainment center
Scene Class.
Top 2 prediction:
• living room
• television room
3D Curvature
Image Reshading
In-painting
Jigsaw puzzle
Colorization
2D Segm.
2.5D Segm.
Semantic Segm.
Cam. Pose (non-fixated)
Cam. Pose (fixated)
Autoencoding
Vanishing Points
2D Edges
3D Edges
2D Keypoints
3D Keypoints
Room Layout
Point Matching
Denoising

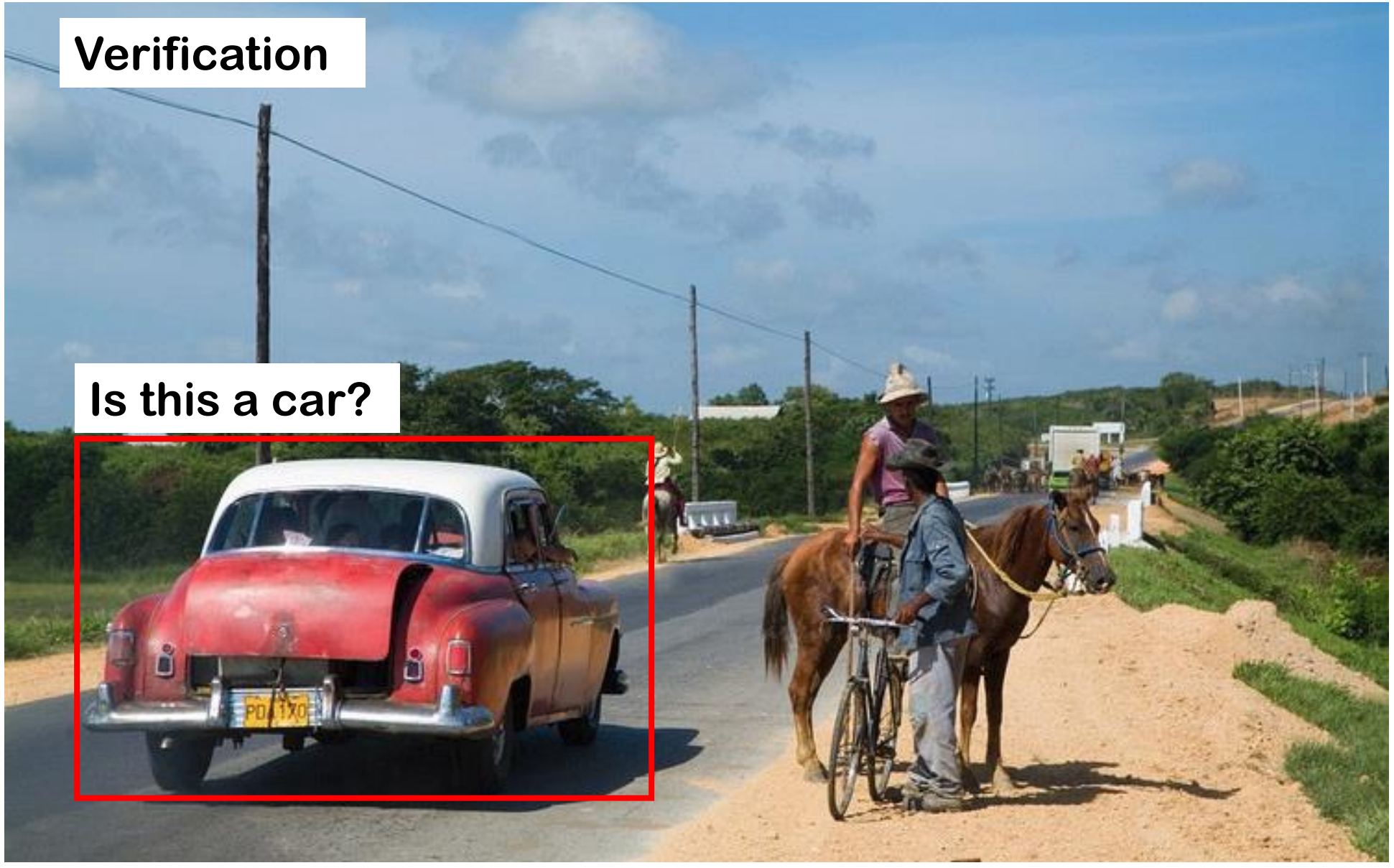

Deep Learning Models
Trained on
BigEarthNet

The issue of data imbalance is prominent



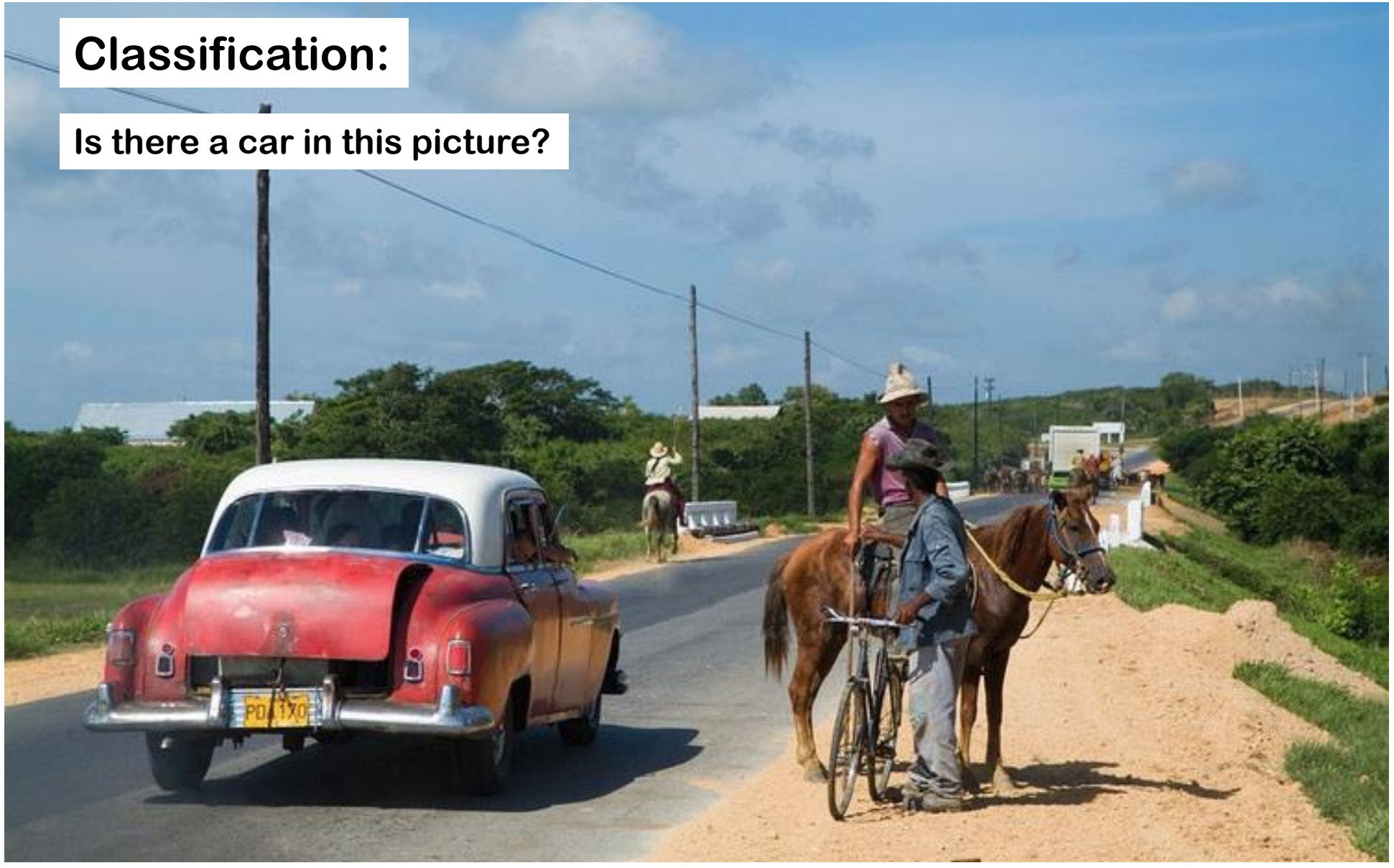
What are the tasks we care about?

- What are the vision problems that can best be solved in a generalized manner using ML/DL techniques?


Verification
Is this a car?

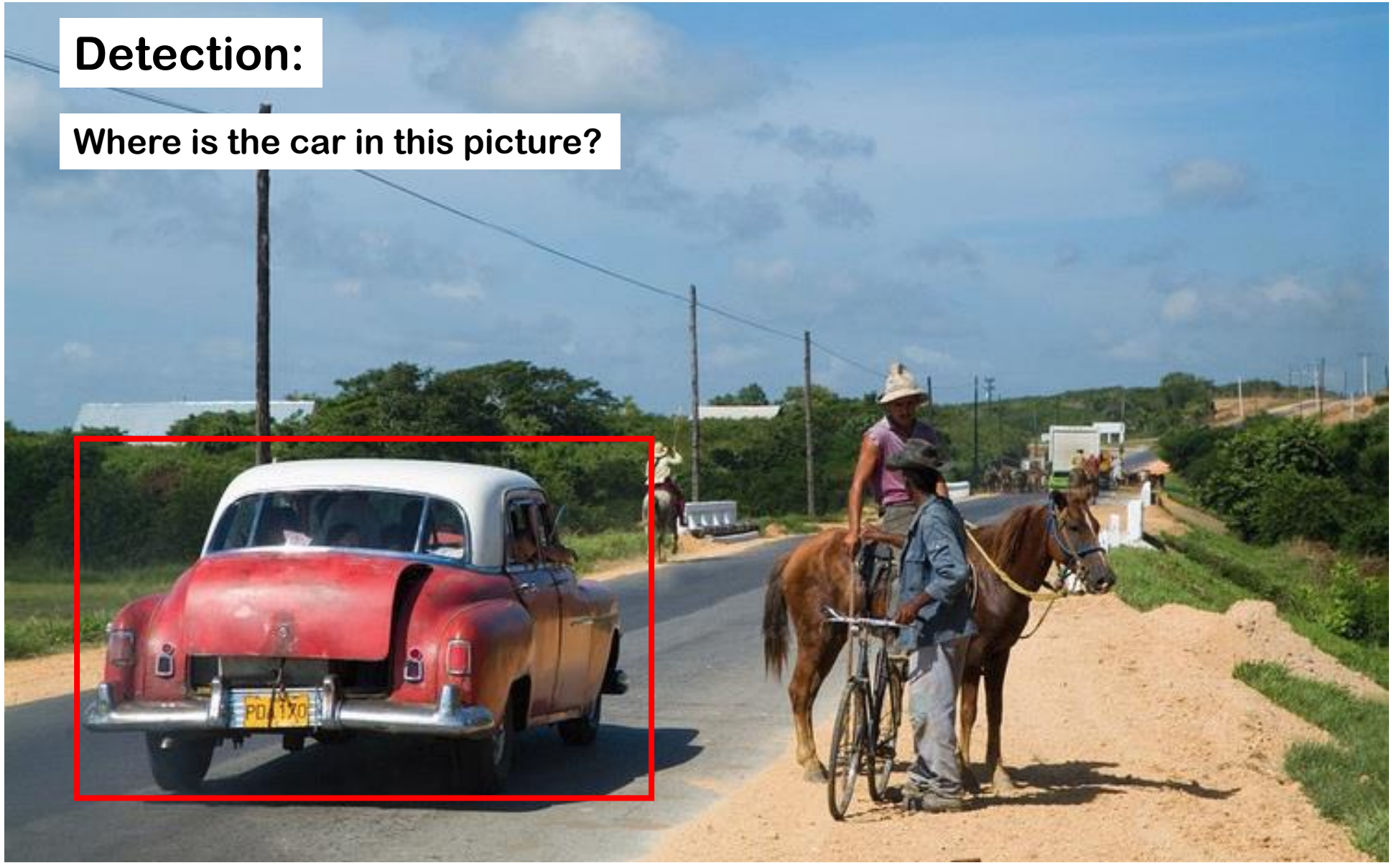
Classification:

Is there a car in this picture?

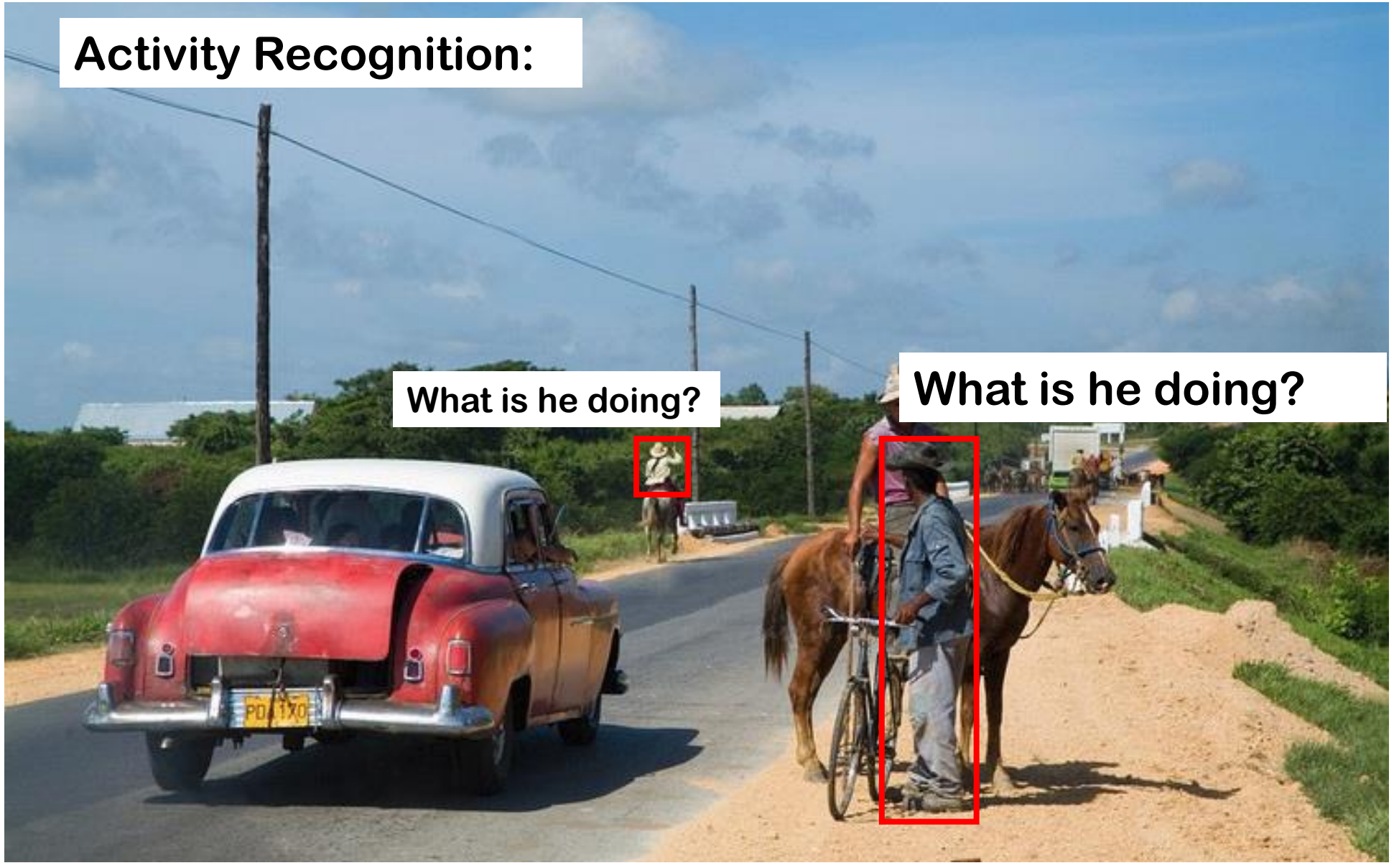


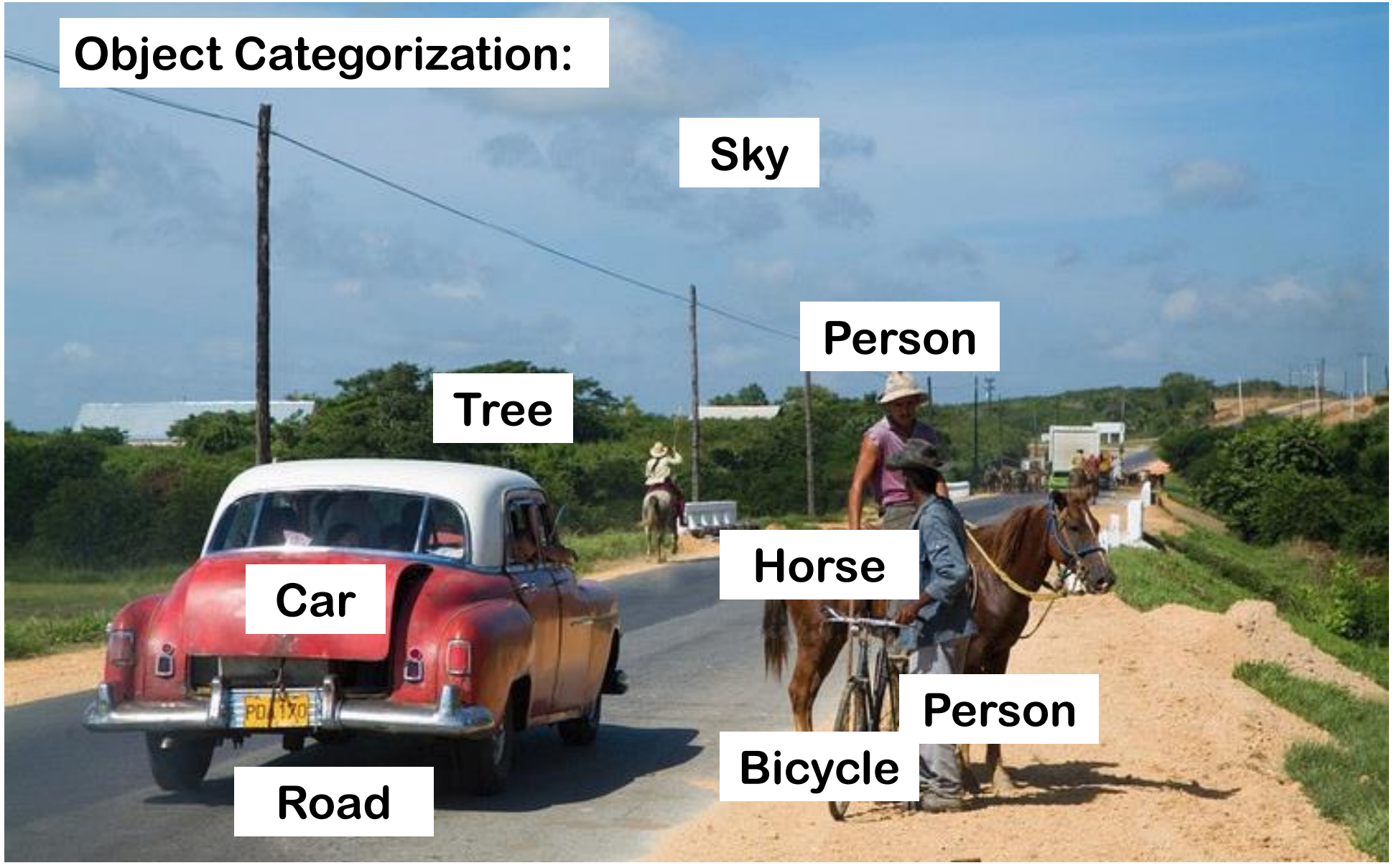
Detection:

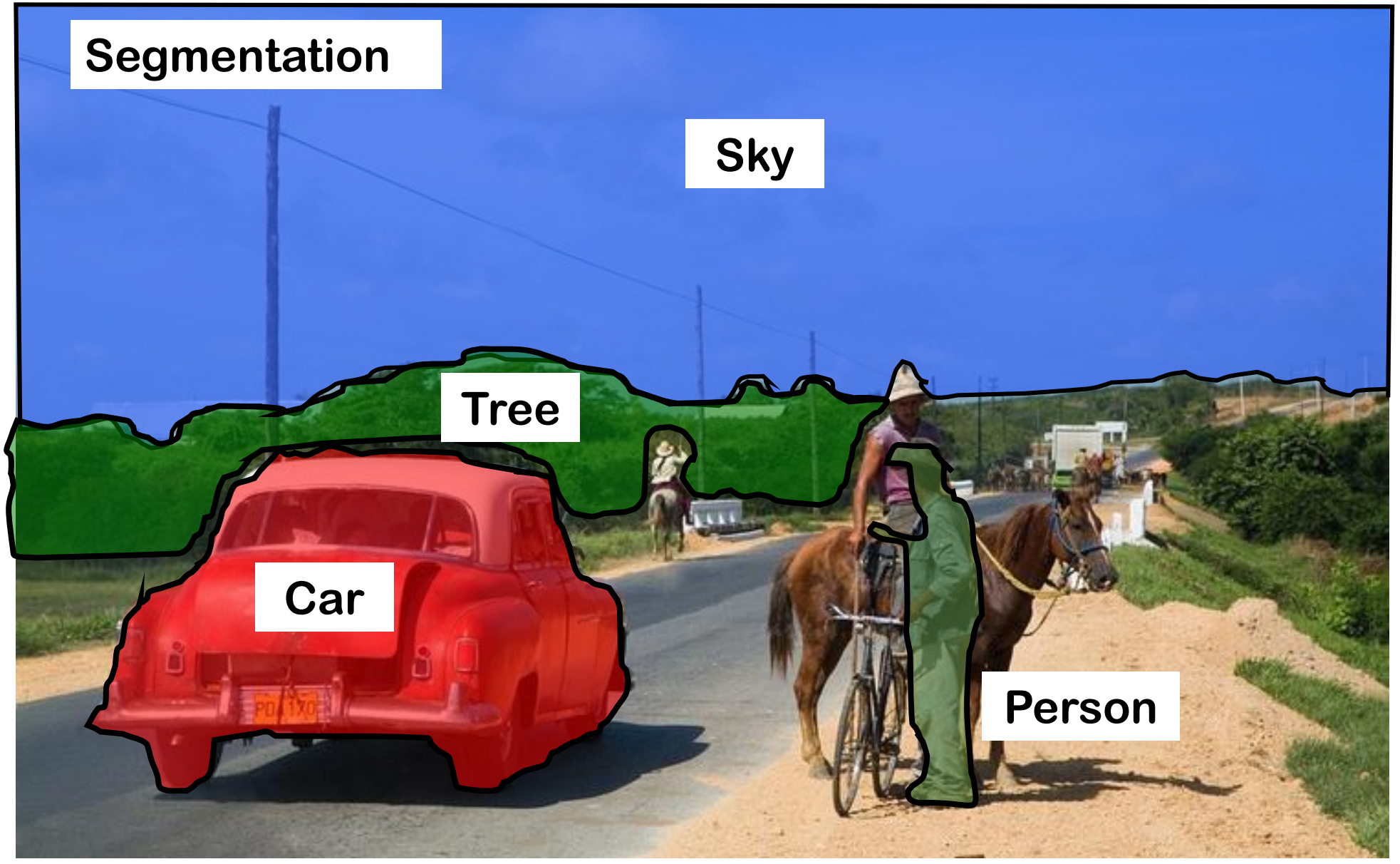
Where is the car in this picture?



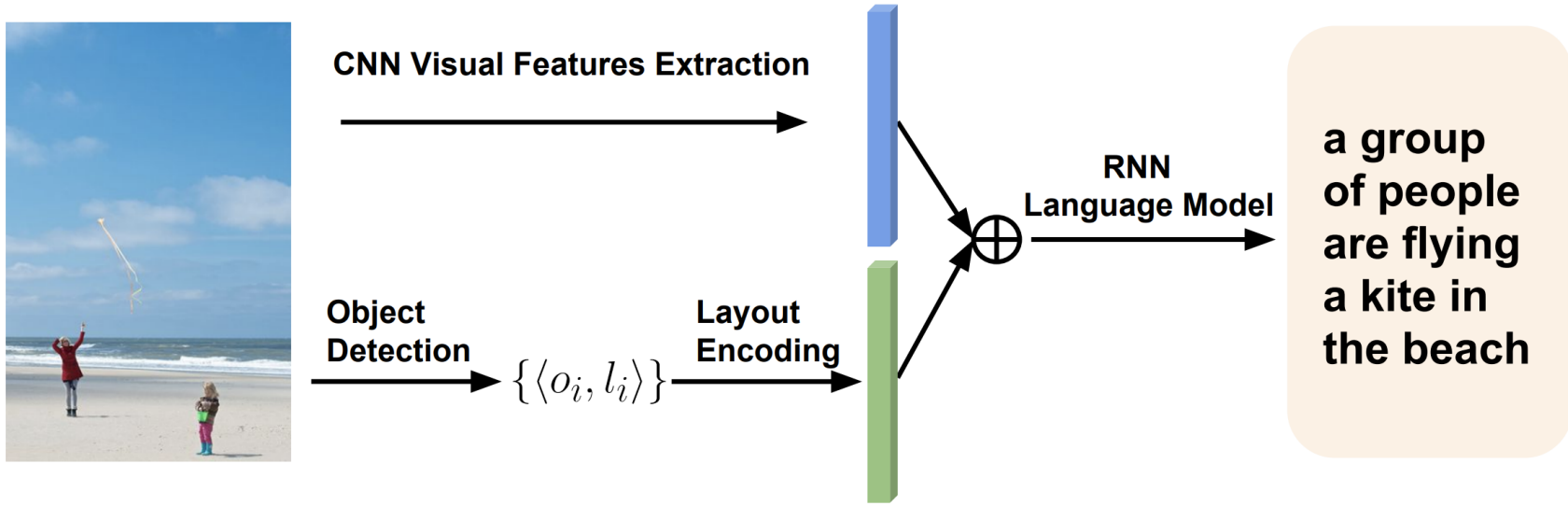

Pose Estimation:


Activity Recognition:
What is he doing?
What is he doing?

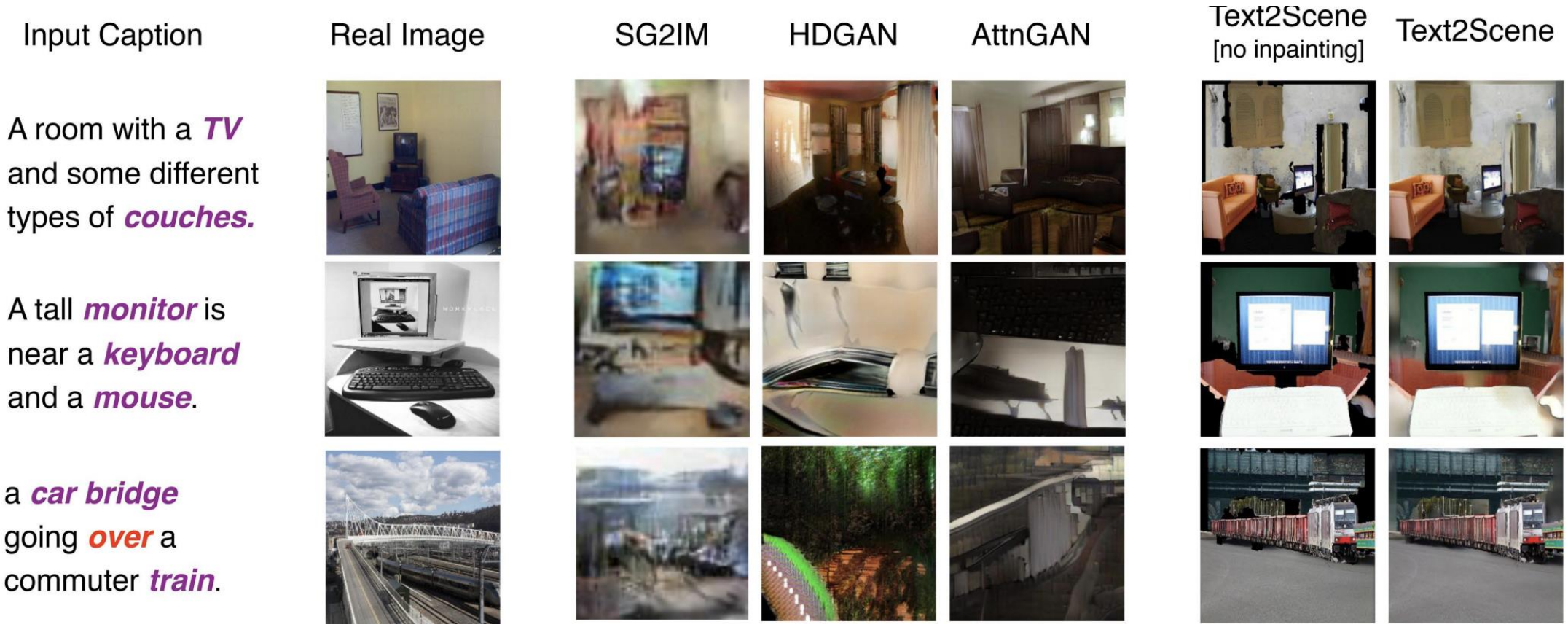

Object Categorization:
Sky
Person
Tree
Horse
Car
Person
Bicycle
Road


Segmentation
Sky
Tree
Car
Person

Describing Images with Language



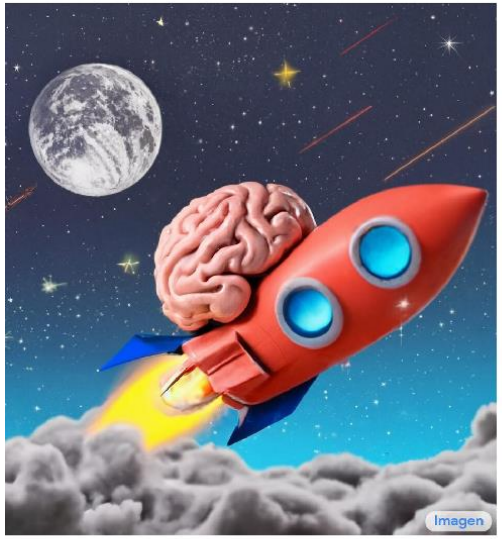
Text-to-Image Synthesis: Text2Scene



dy bears swimming at the Olympics 400m Butter-
event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright
golden glow is coming from the chest.

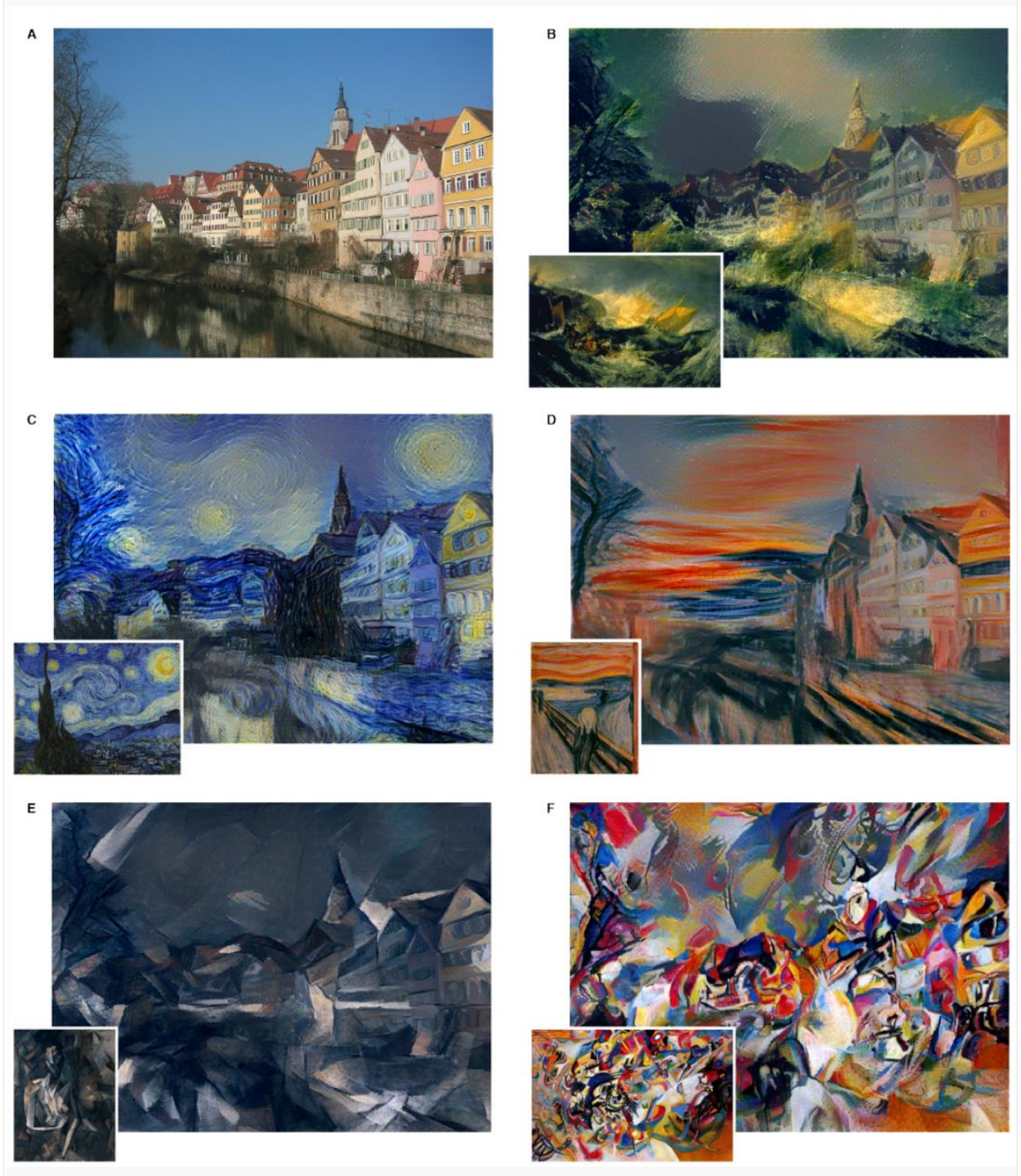


rain riding a rocketship heading towards the moon.

A dragon fruit wearing karate belt in the snow.

A strawberry mug filled with white sesame seeds. The
mug is floating in a dark chocolate sea.

Neural style transfer



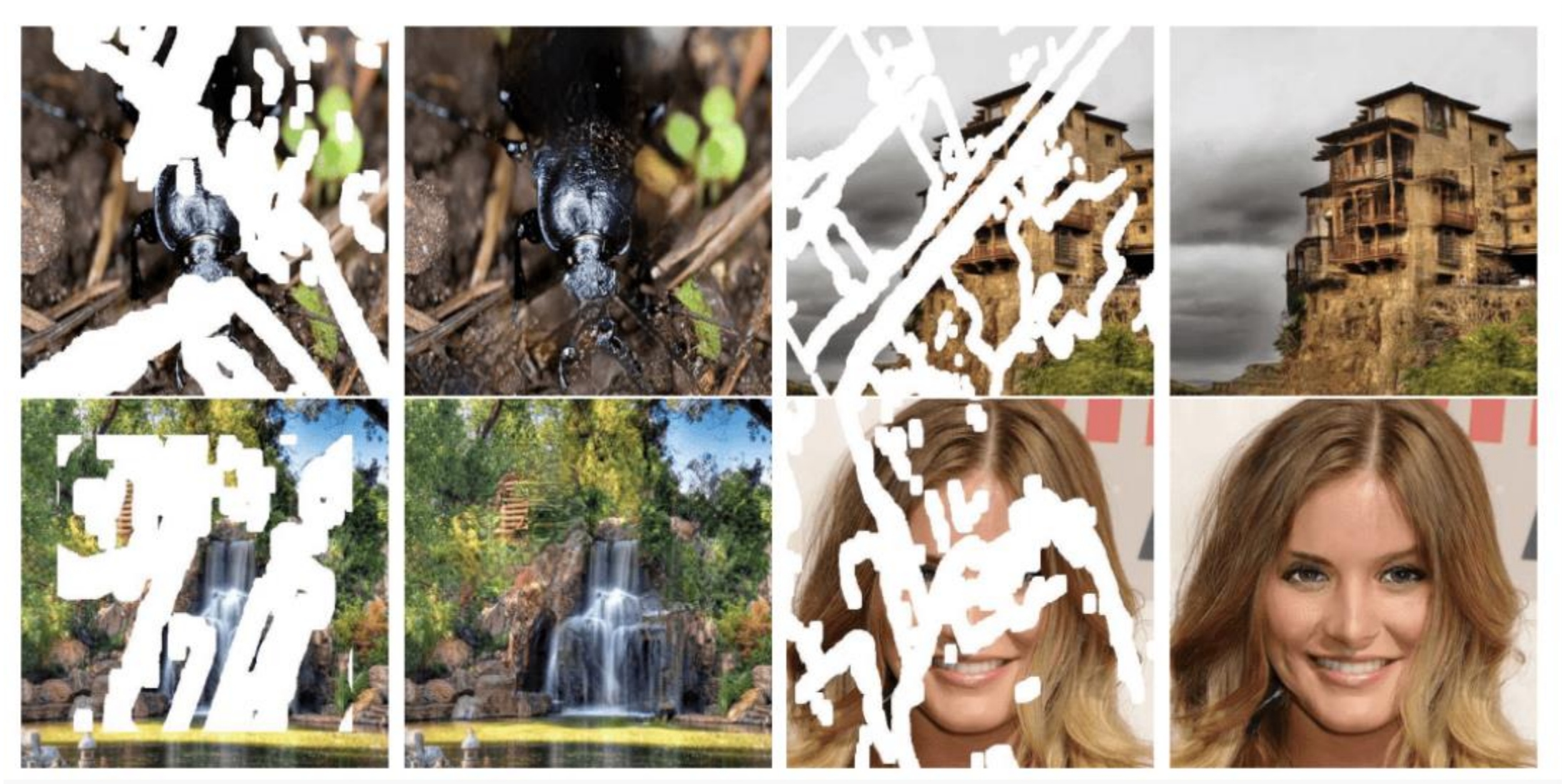


Image inpainting

Zebras ⇄ Horses

zebra → horse

horse → zebra

However, our main tasks to consider at the beginning of the course

Is this a dog?

Image Classification

What is there in image and where?

Object Detection

Which pixels belong to which object?

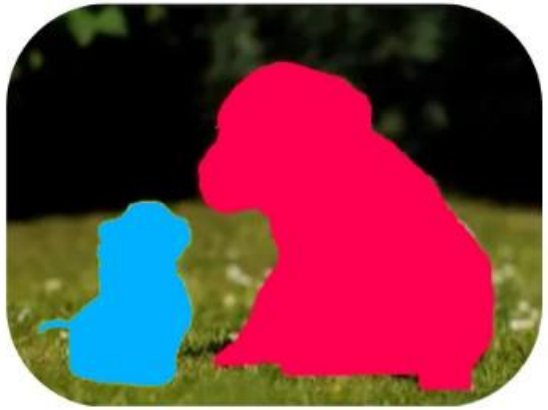


Image Segmentation

Why is recognition a difficult problem?

Challenges 1: view point variation

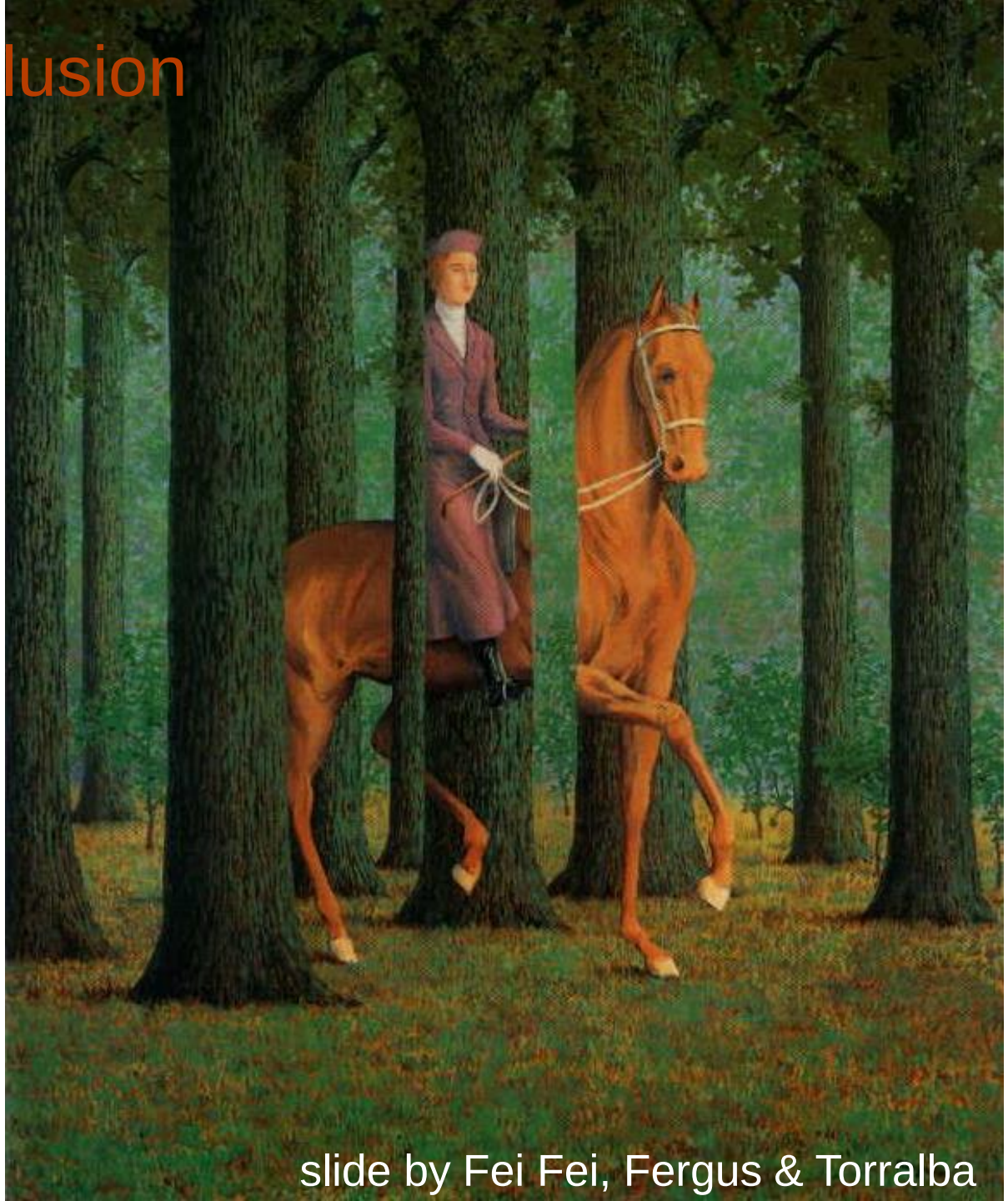
Michelangelo 1475-1564

slide by F

Challenges 2: illumination

slide credit: S. Ullman

Challenges 3: occlusion



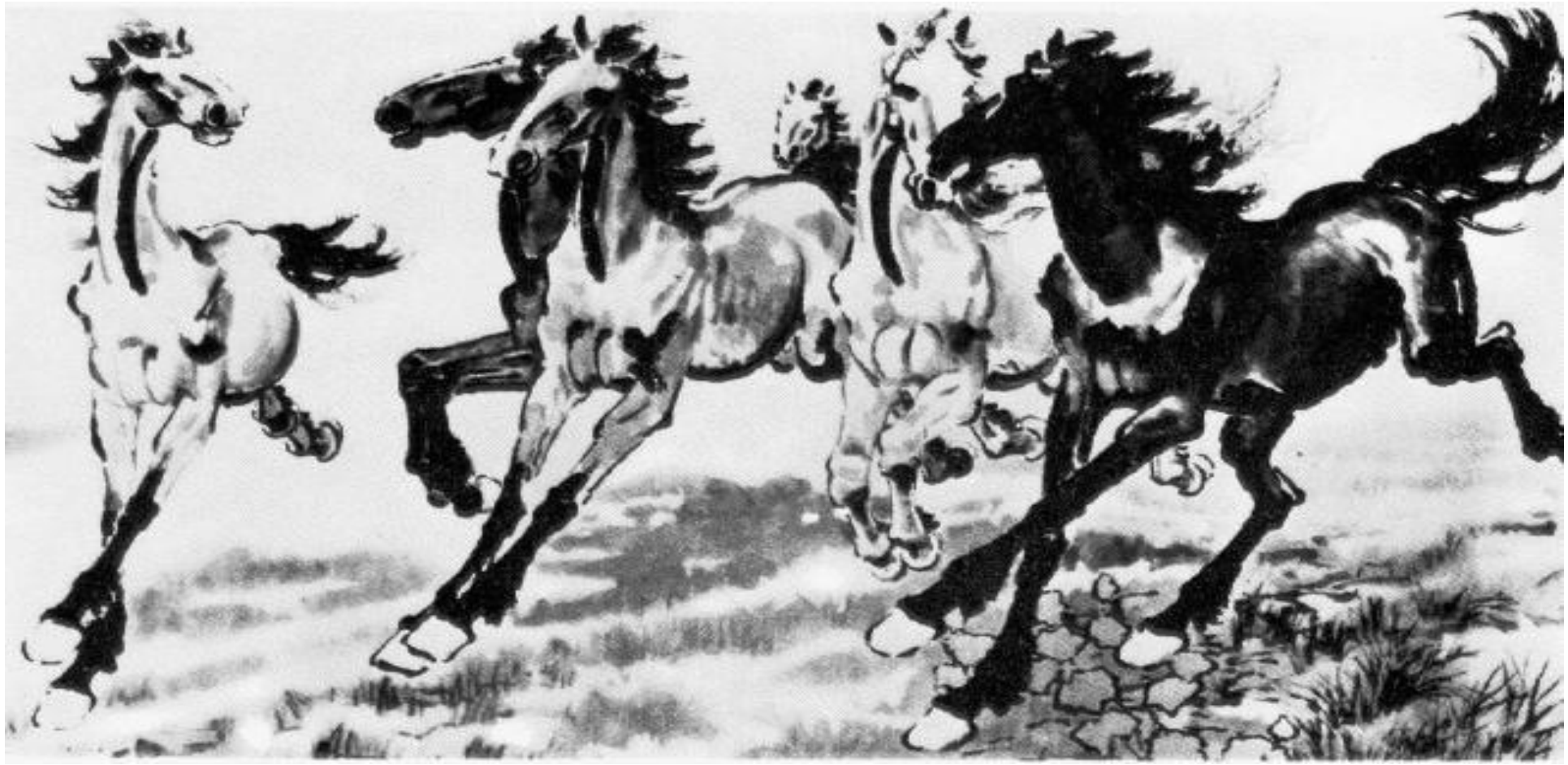
Magritte, 1957

slide by Fei Fei, Fergus & Torralba

Challenges 4: scale

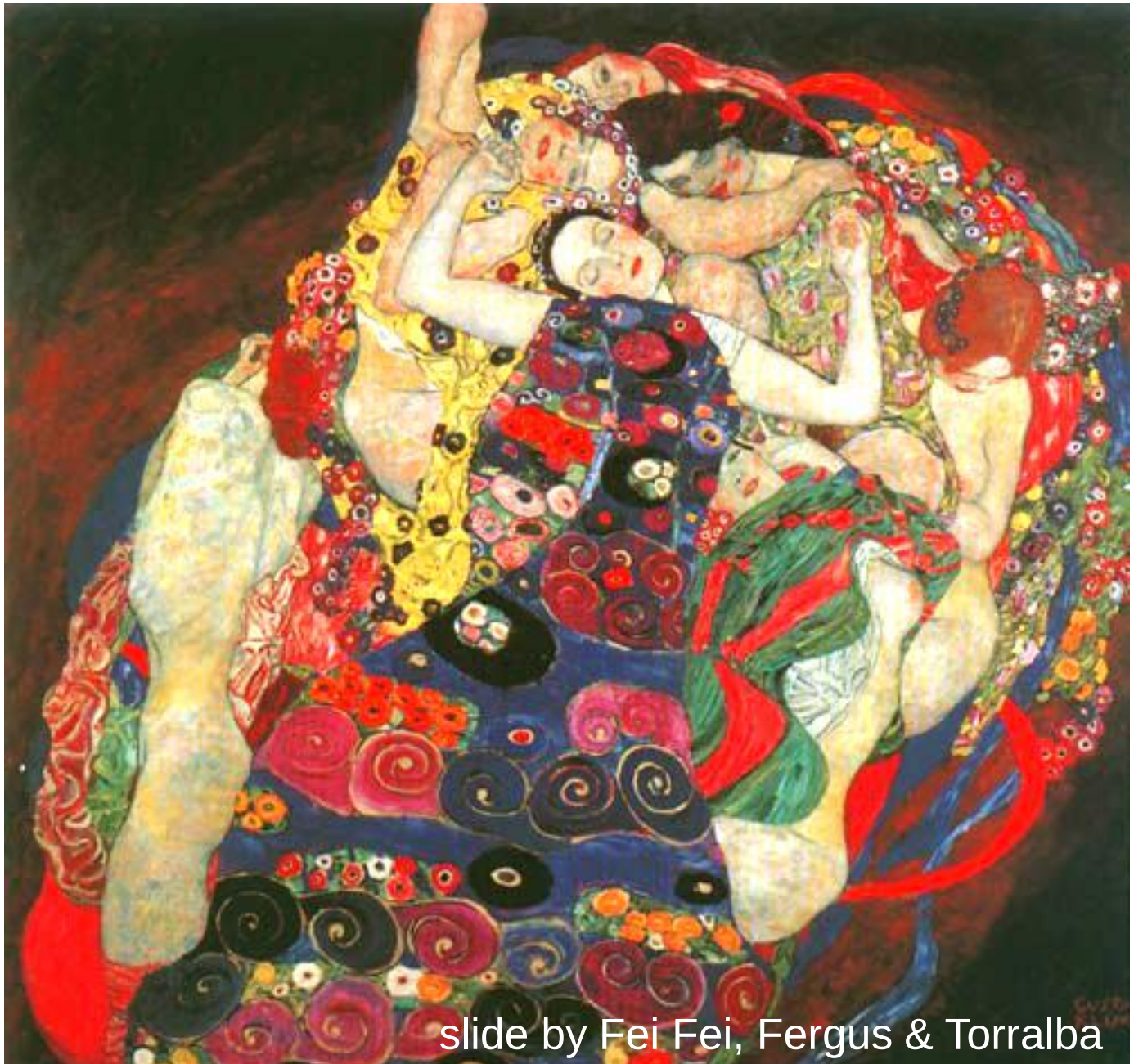
slide by Fei Fei, Fergus & Torralba

Challenges 5: deformation



Xu, Beihong 1943

Challenges 6: background clutter

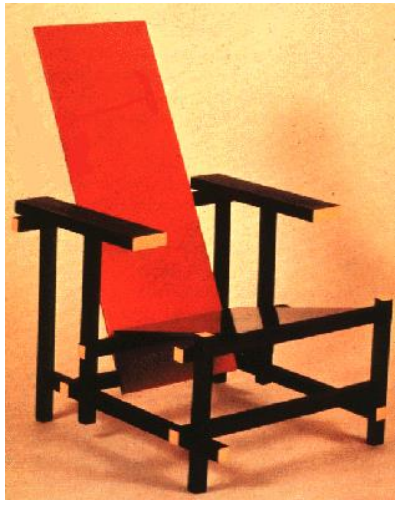


Klimt, 1913

slide by Fei Fei, Fergus & Torralba

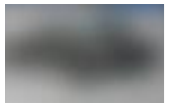
Challenges 7: object intra-class variation

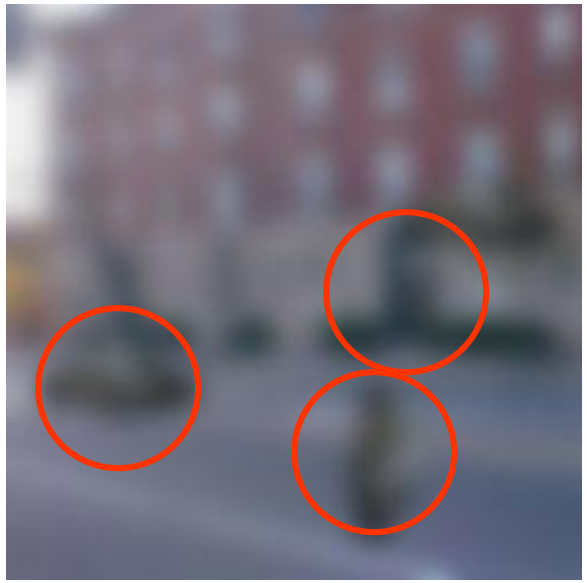


slide by Fei-Fei, Fergus & Torralba

Challenges 8: local ambiguity

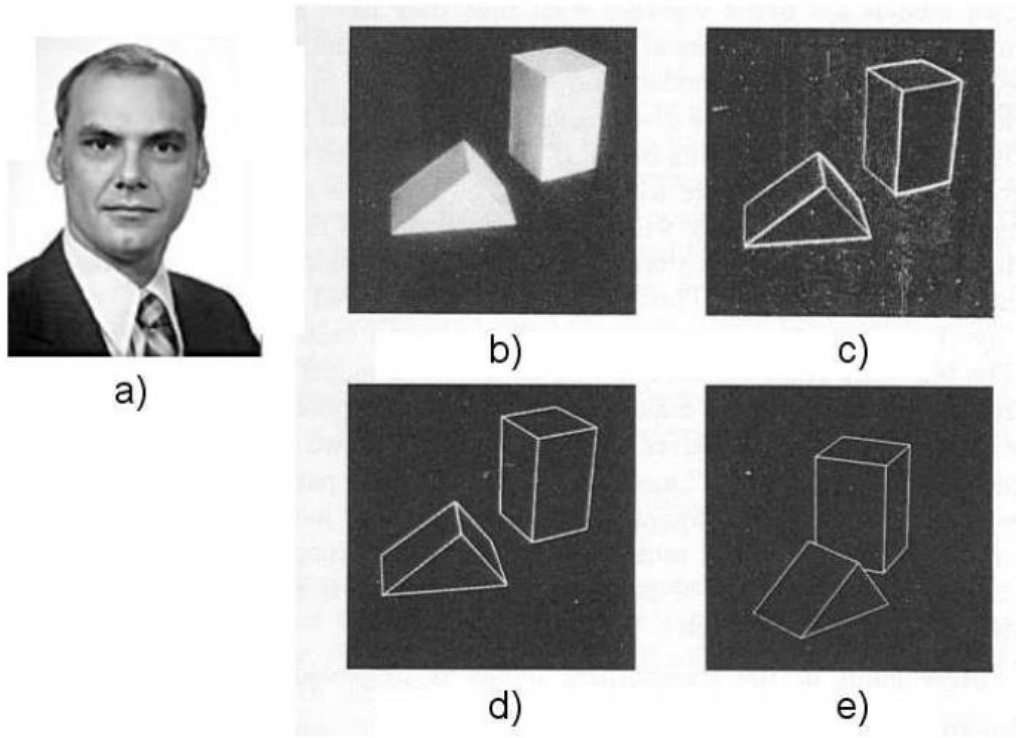




slide by Fei-Fei, Fergus & Torralba

What are the progresses so far?

Recognition as an alignment problem: Block world



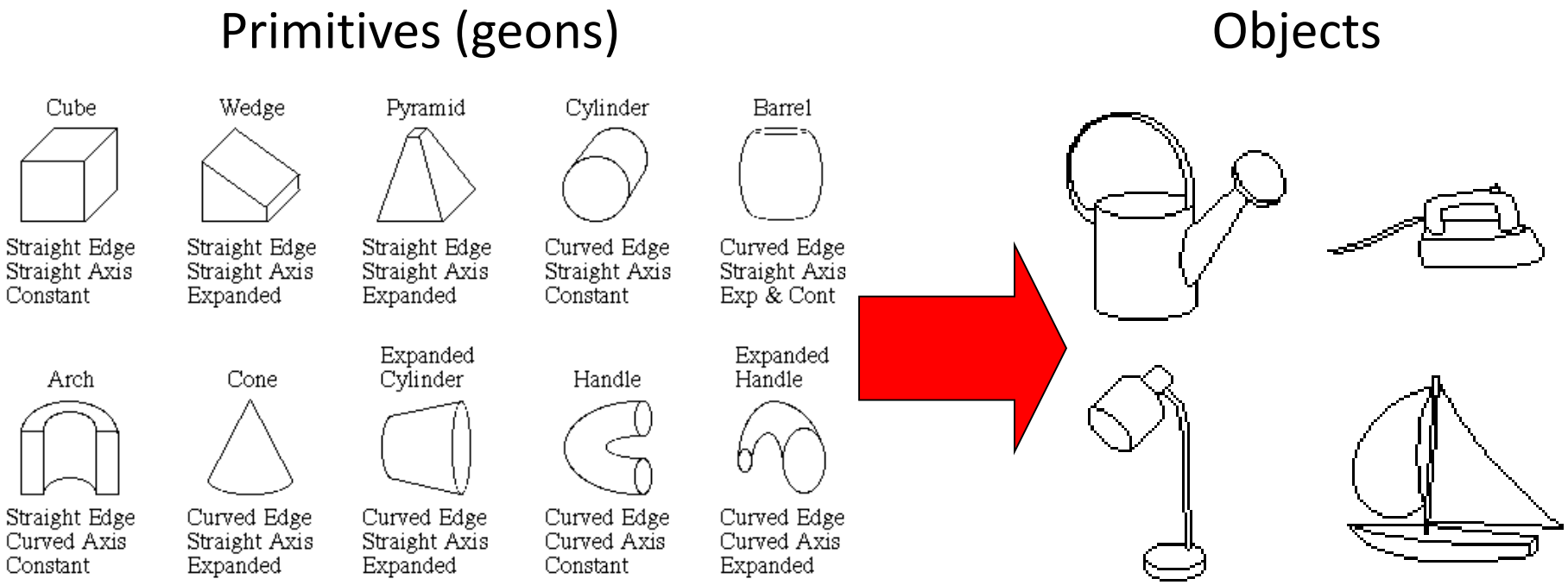
L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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

Recognition by components

Biederman (1987)



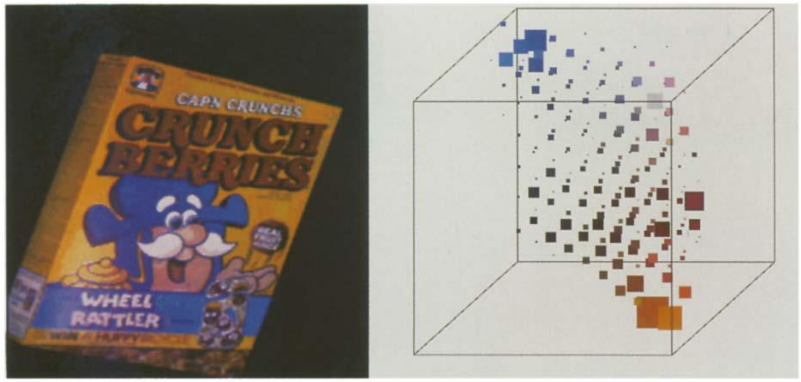
http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

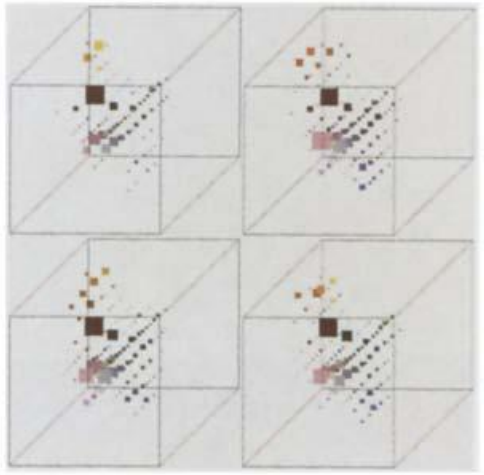
History of ideas in recognition

- 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: combination of local and global methods, data-driven methods, context

Eigenfaces (Turk & Pentland, 1991)

Color Histograms



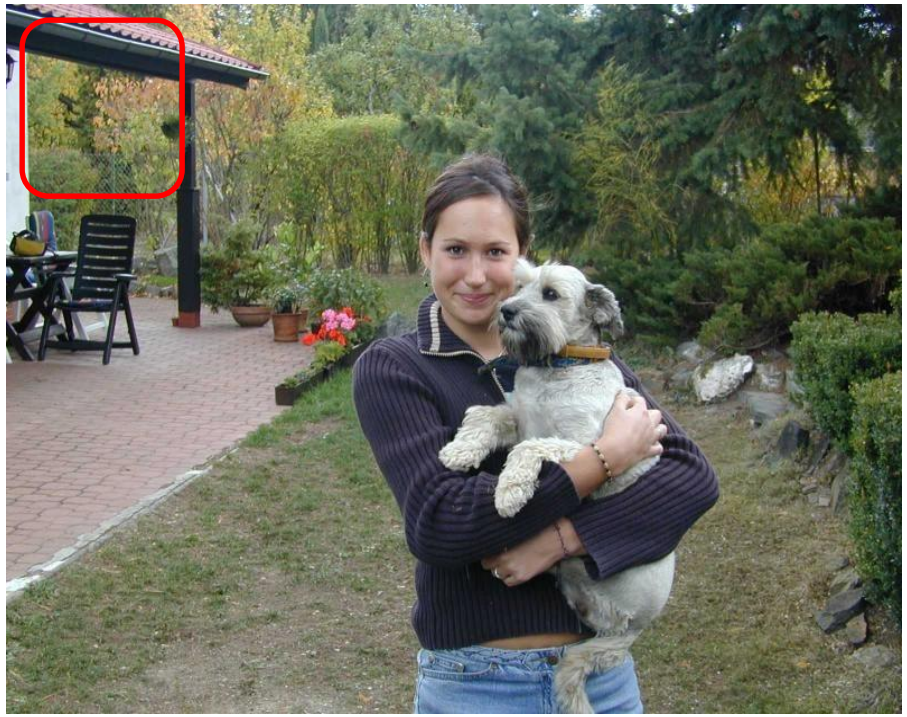


Swain and Ballard, Color Indexing, IJCV 1991.

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: combination of local and global methods, data-driven methods, context

Sliding window approaches



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: combination of local and global methods, data-driven methods, context

Local features for object instance recognition

D. Lowe (1999, 2004)

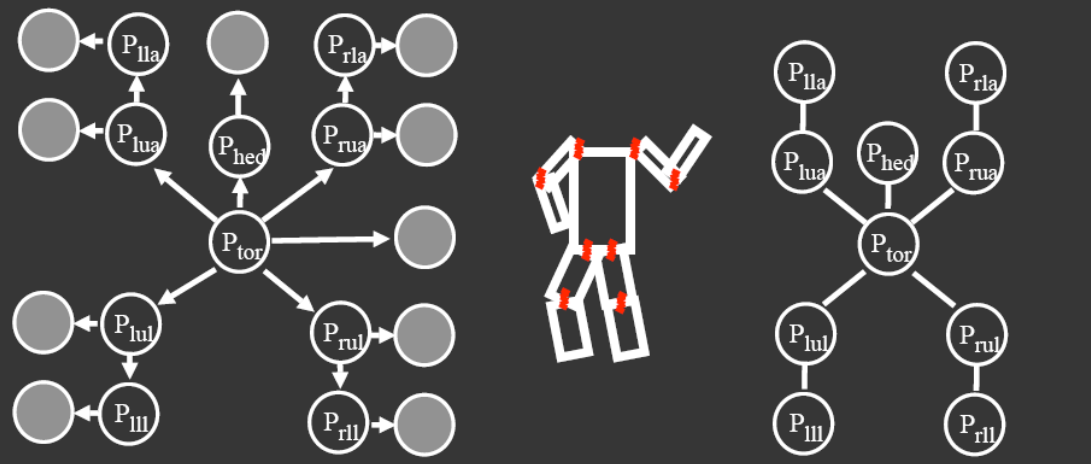
History of ideas in recognition

- 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: combination of local and global methods, data-driven methods, context

Representing people

Pictorial structure model

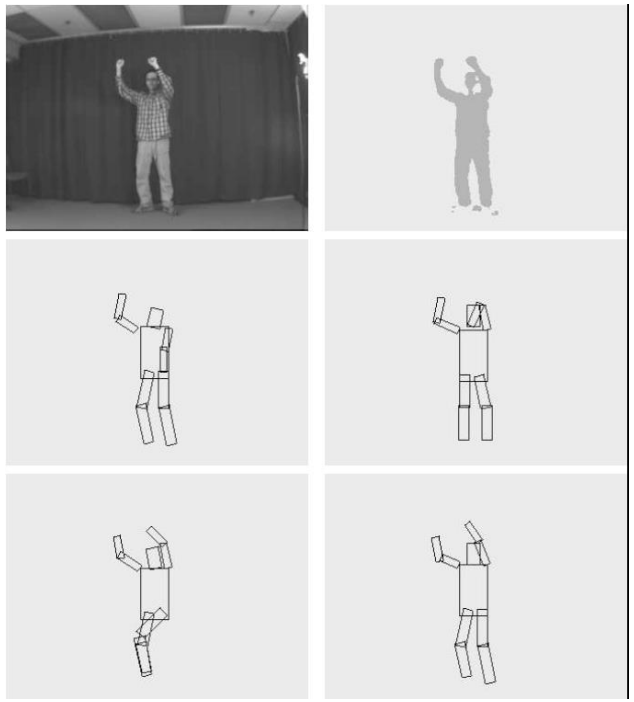
Fischler and Elschlager(73), Felzenszwalb and Huttenlocher(00)



$$\Pr(P_{tor}, P_{arm}, \ldots | \mathrm{Im}) \propto \prod_{i,j} \Pr(P_i \mid P_j) \prod_i \Pr(\mathrm{Im}(P_i))$$

part geometry

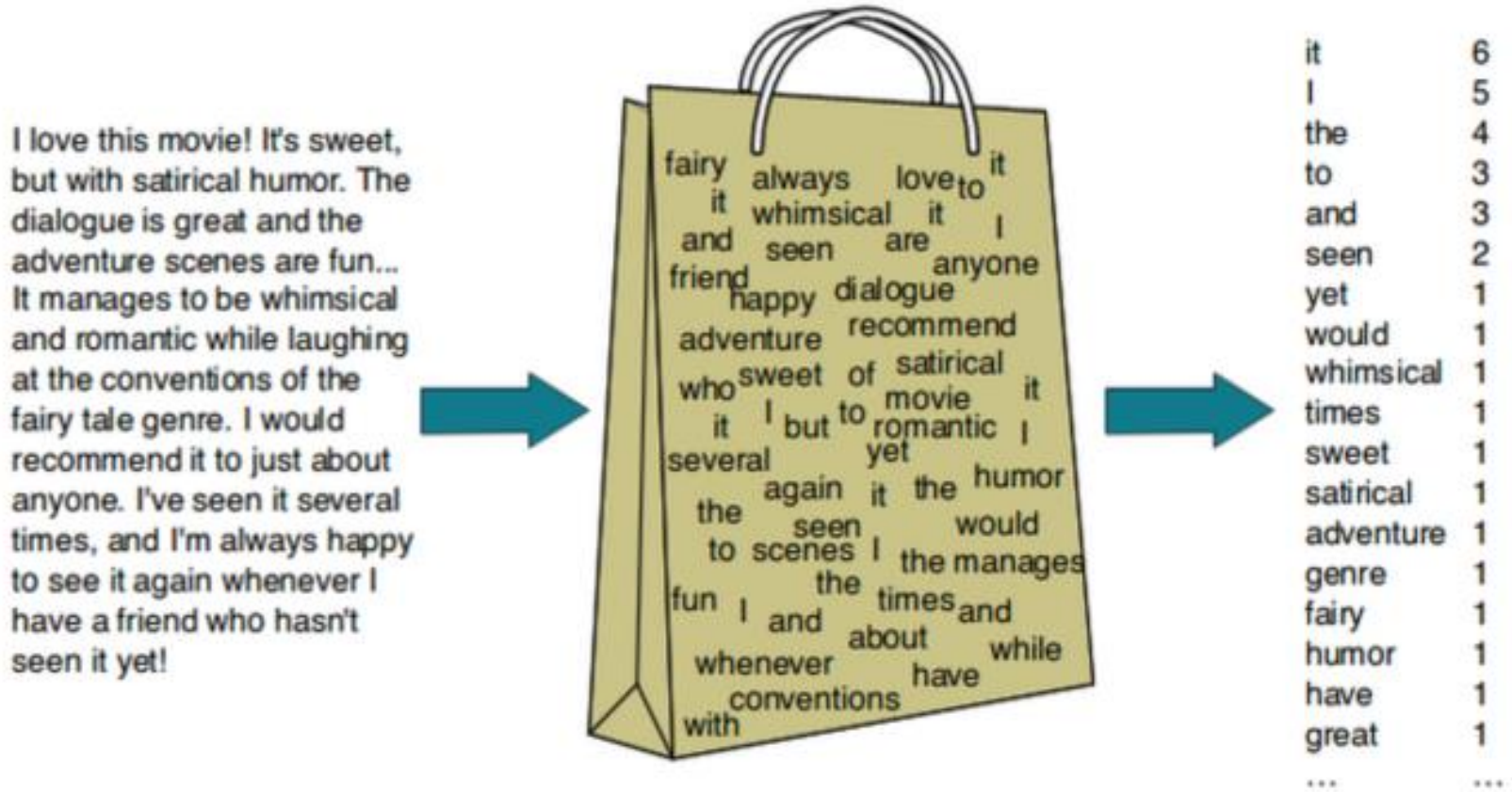
part appearance



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: combination of local and global methods, data-driven methods, context

Bag of words in NLP

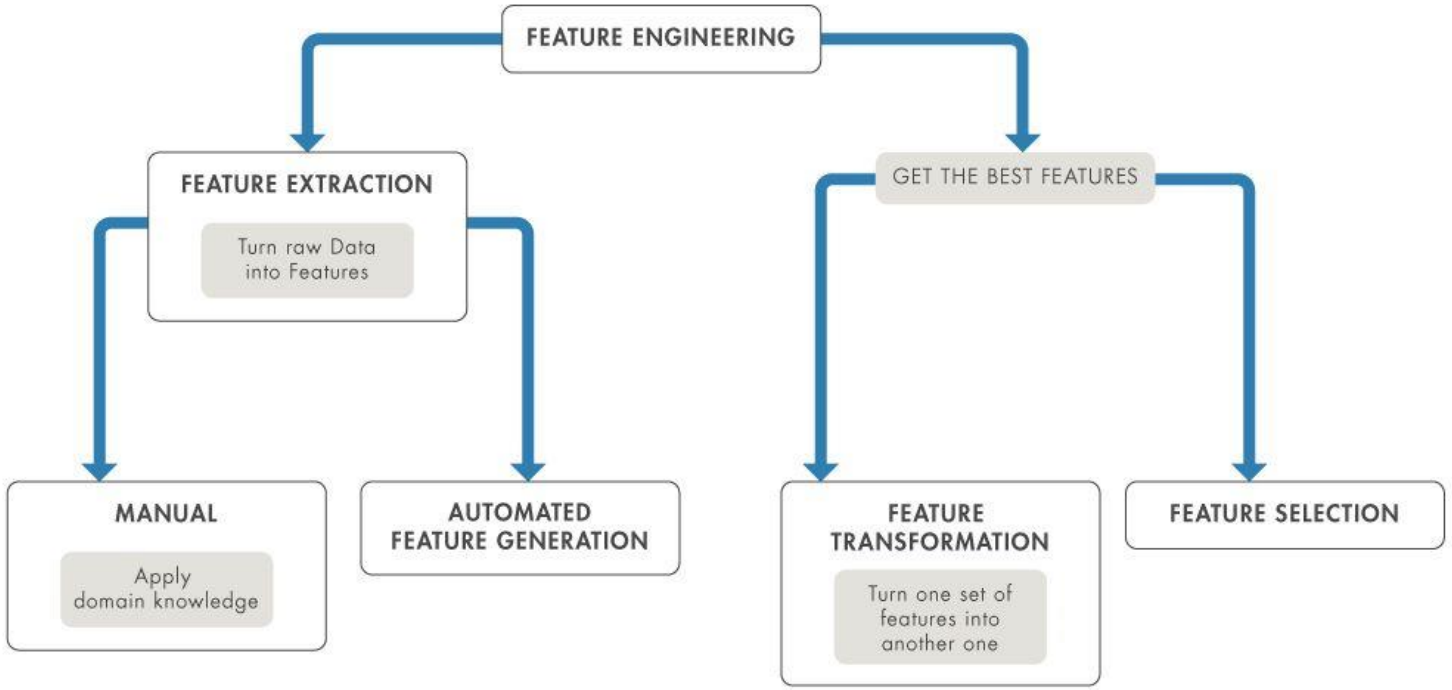


For each theme of paragraph, the bag contains theme-specific words

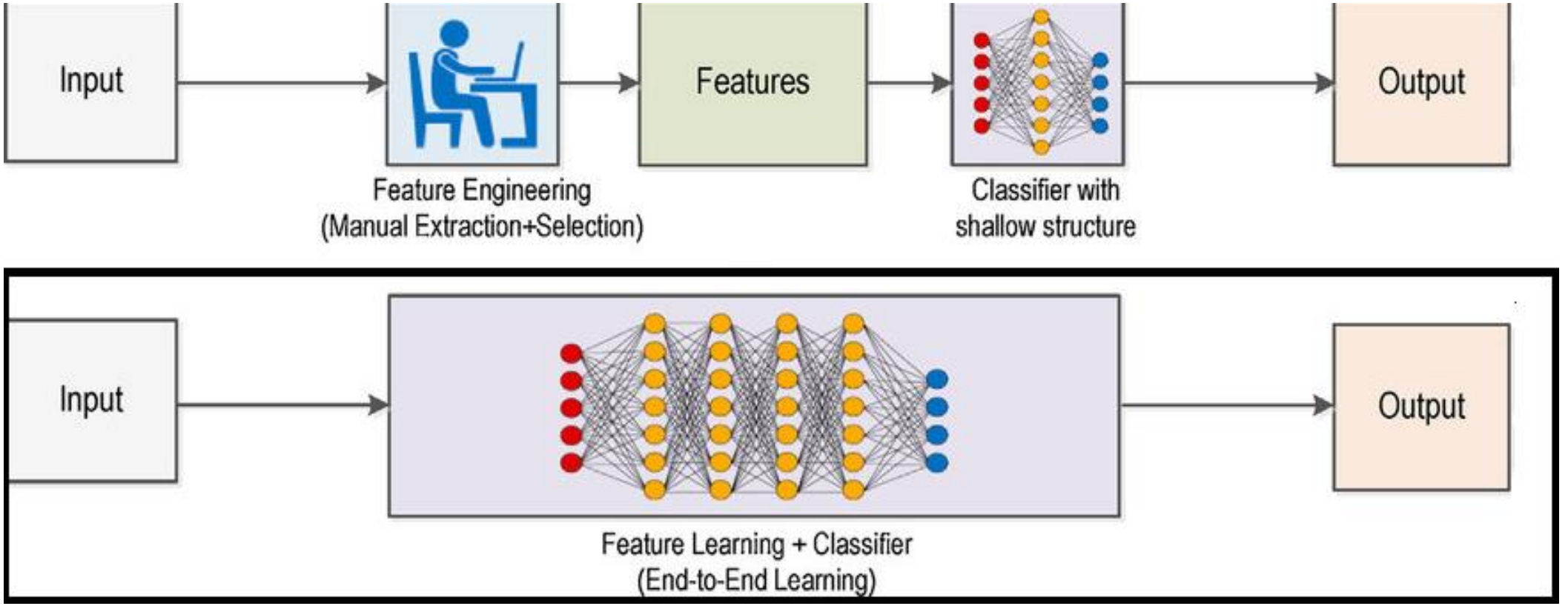
Can we do the same with images? What are visual words?

Object → **Bag of ‘words’**

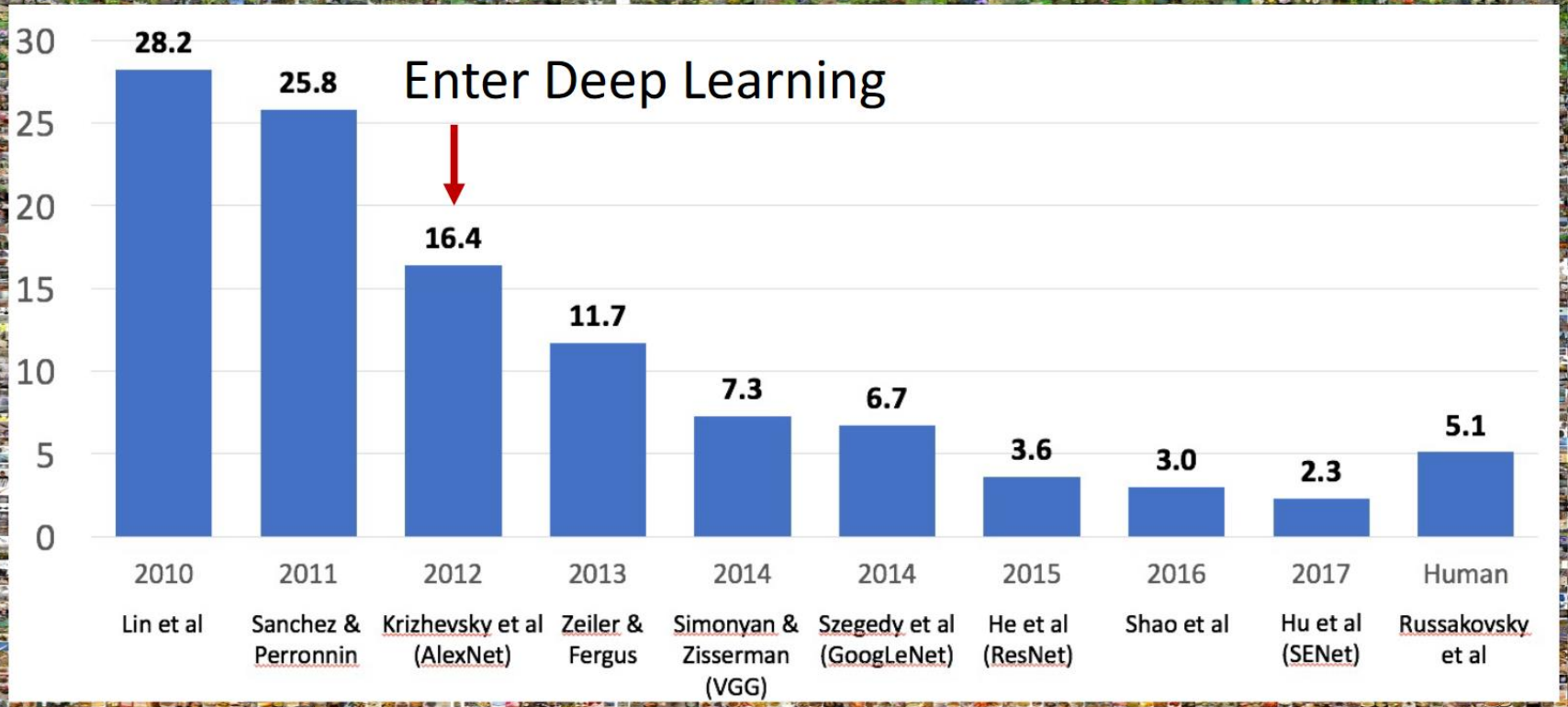
- ✓ All these techniques are based on manually defining features from the images
- ✓ Features – fixed length representations of the images/objects which are discriminative
- ✓ Manual feature extractions are ambiguous – which features are to be considered or discarded are not clear
- ✓ Many ad-hoc techniques follow to refine the features



The main idea of doing deep learning



IMAGENET Large Scale Visual Recognition Challenge



Perceptron

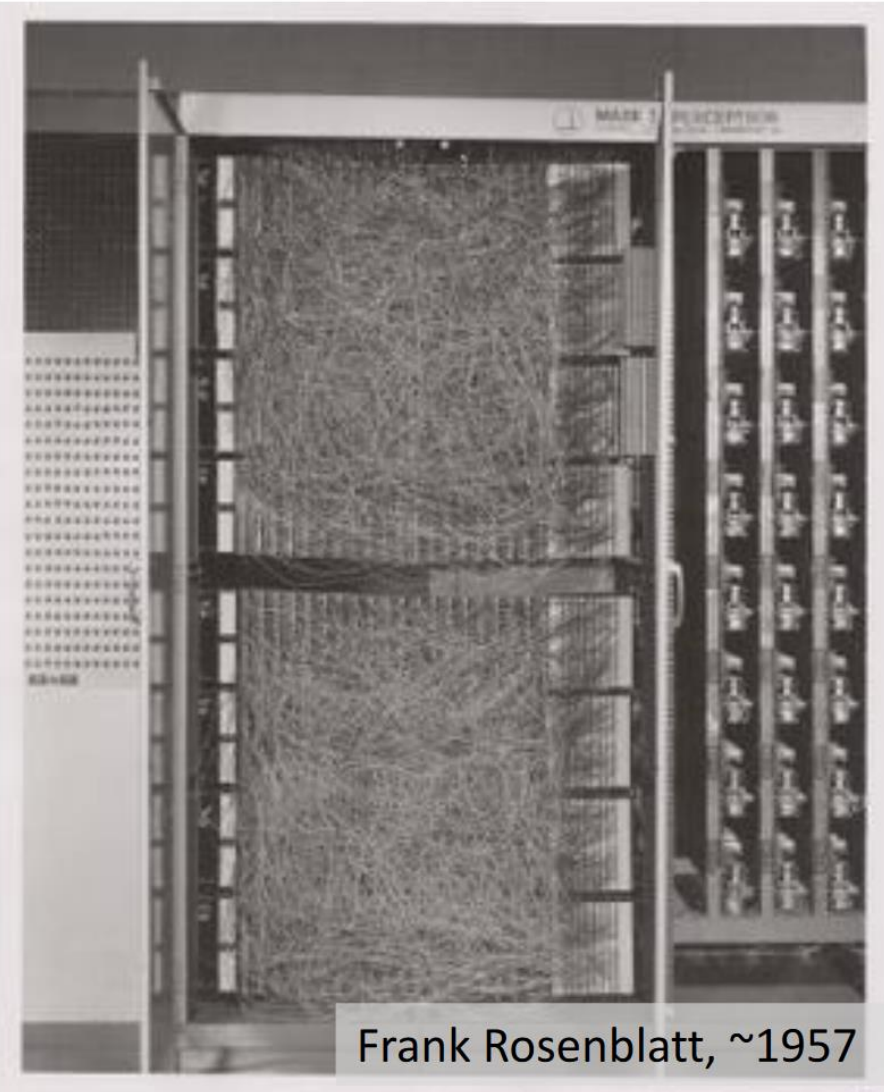
One of the earliest algorithms that could learn from data

Implemented in hardware! Weights stored in potentiometers, updated with electric motors during learning

Connected to a camera that used 20x20 cadmium sulfide photocells to make a 400-pixel image

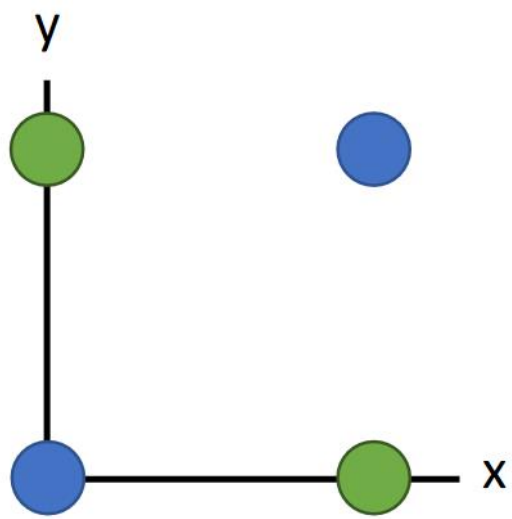
Could learn to recognize letters of the alphabet

Today we would recognize it as a **linear classifier**

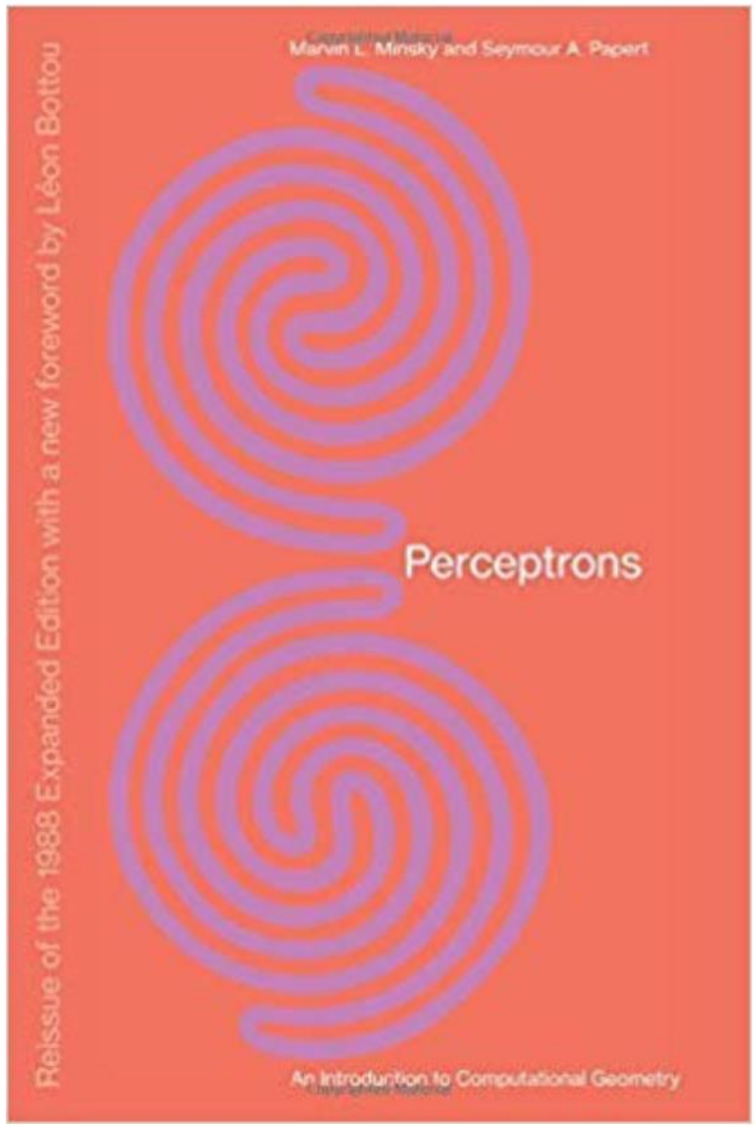

Frank Rosenblatt, ~1957

Minsky and Papert, 1969

X	Y	F(x,y)
0	0	0
0	1	1
1	0	1
1	1	0



Showed that Perceptrons could not learn the XOR function
Caused a lot of disillusionment in the field

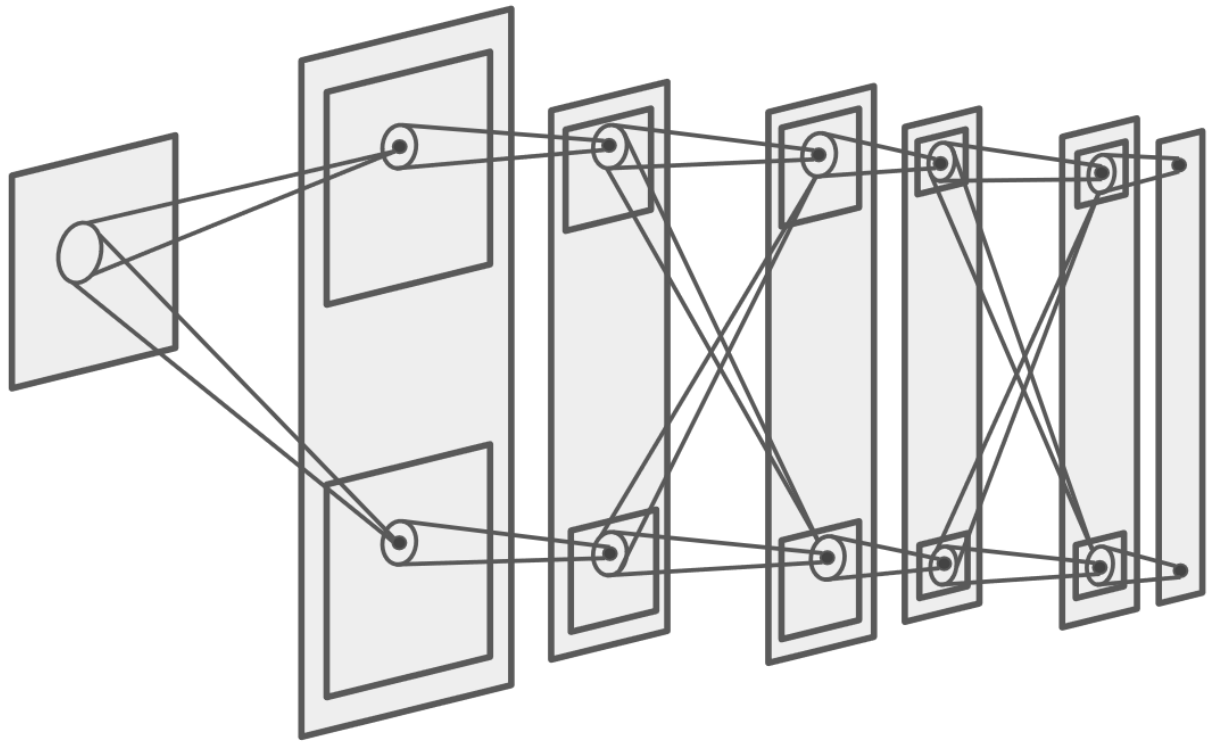


Neocognitron: Fukushima, 1980

Computational model the visual system, directly inspired by Hubel and Wiesel's hierarchy of complex and simple cells

Interleaved simple cells (convolution) and complex cells (pooling)

No practical training algorithm



Backprop: Rumelhart, Hinton, and Williams, 1986

Introduced backpropagation for computing gradients in neural networks

Successfully trained perceptrons with multiple layers

$$\frac{\partial E_p}{\partial w_{ji}} = \frac{\partial E_p}{\partial o_{pj}} \frac{\partial o_{pj}}{\partial w_{ji}}$$

recognizable math

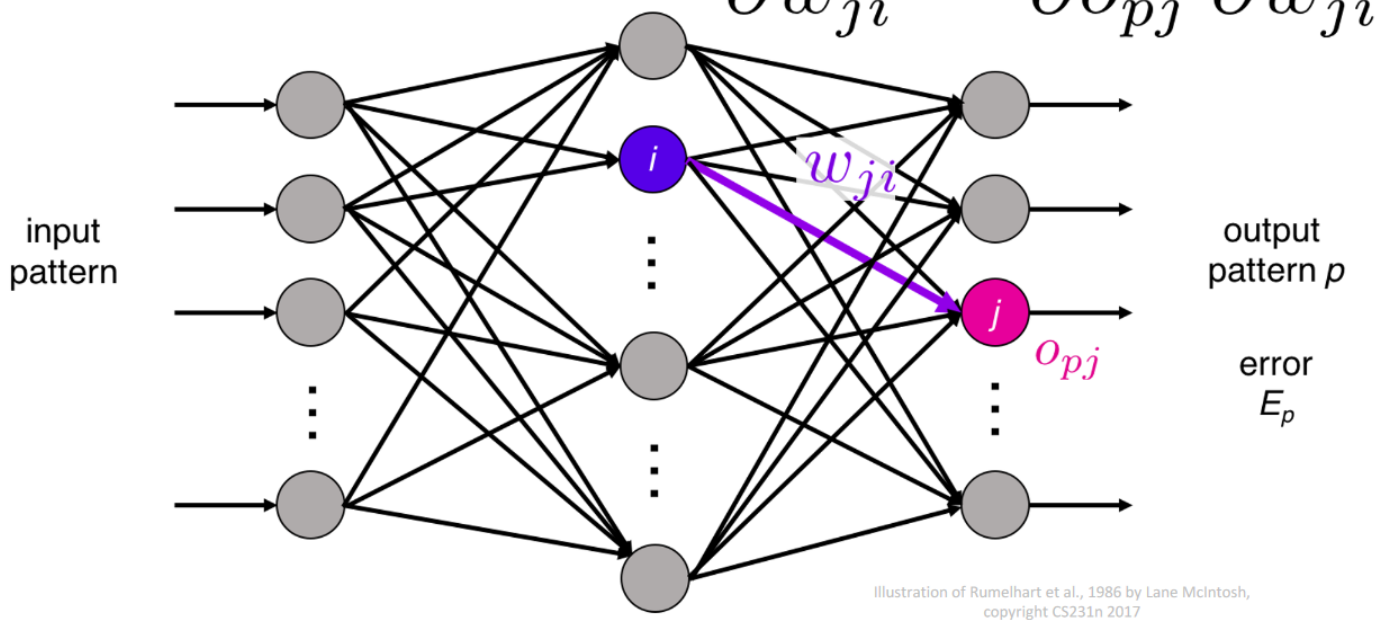
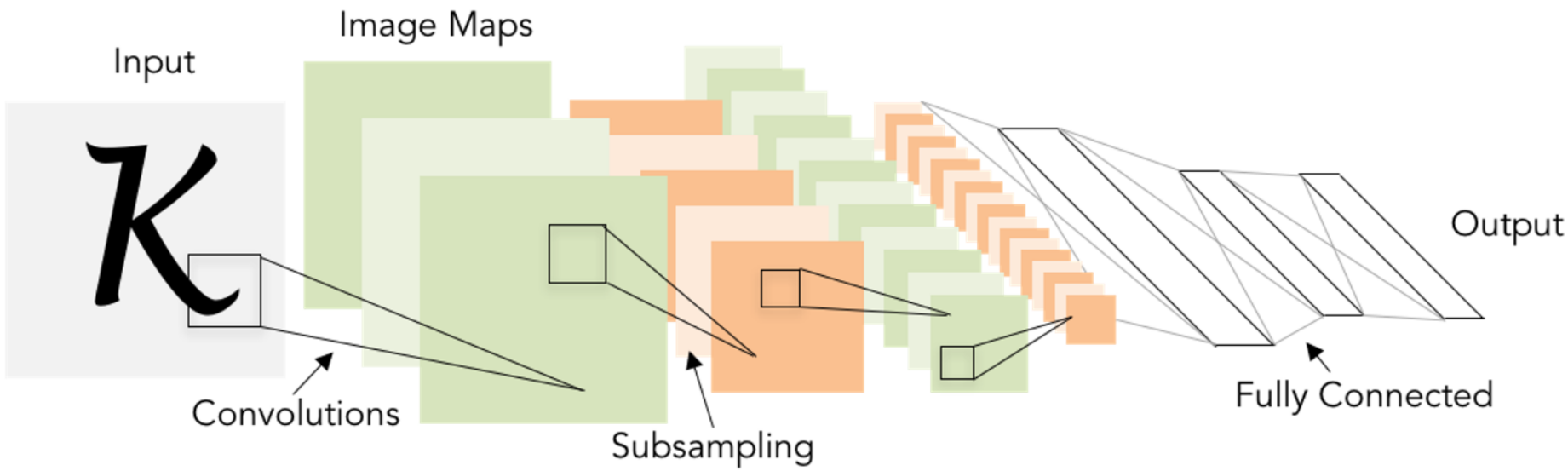


Illustration of Rumelhart et al., 1986 by Lane McIntosh, copyright CS231n 2017

Convolutional Networks: LeCun et al, 1998



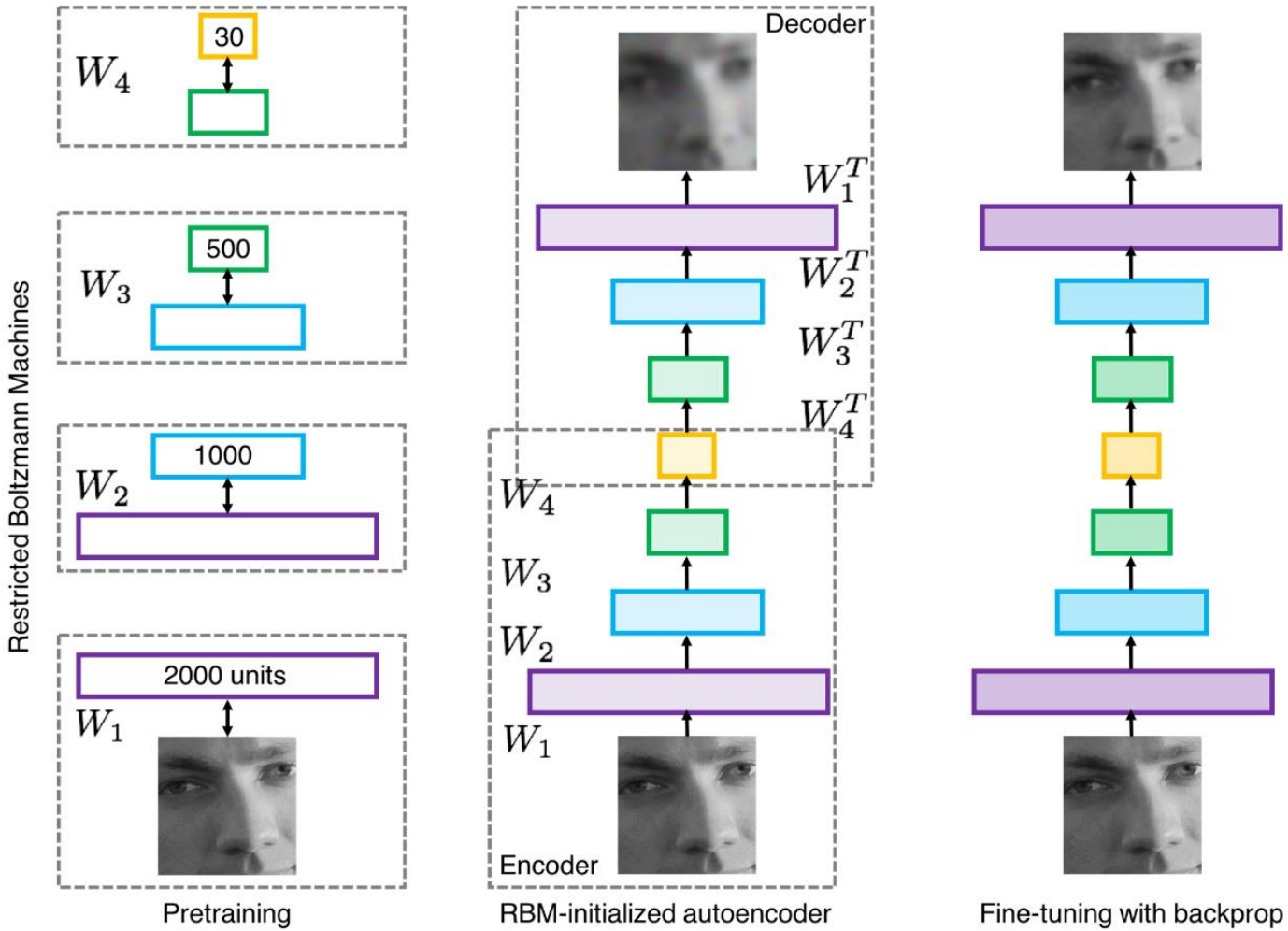
Applied backprop algorithm to a Neocognitron-like architecture
Learned to recognize handwritten digits
Was deployed in a commercial system by NEC, processed handwritten checks
Very similar to our modern convolutional networks!

2000s: “Deep Learning”

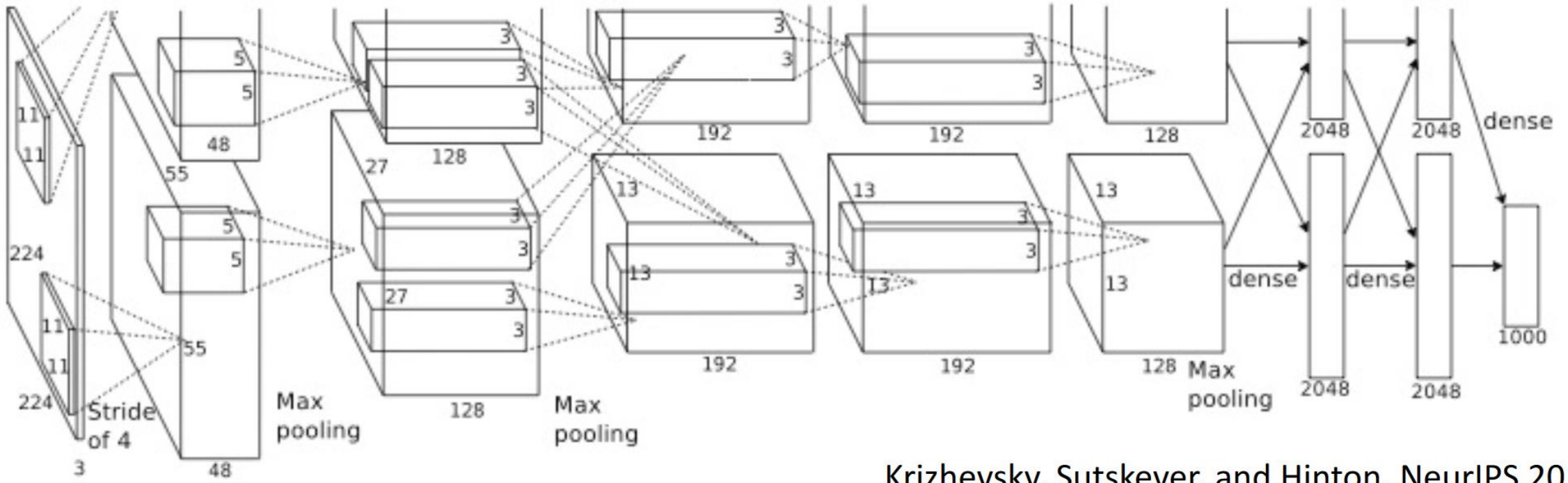
People tried to train neural networks that were deeper and deeper

Not a mainstream research topic at this time

Hinton and Salakhutdinov, 2006
Bengio et al, 2007
Lee et al, 2009
Glorot and Bengio, 2010

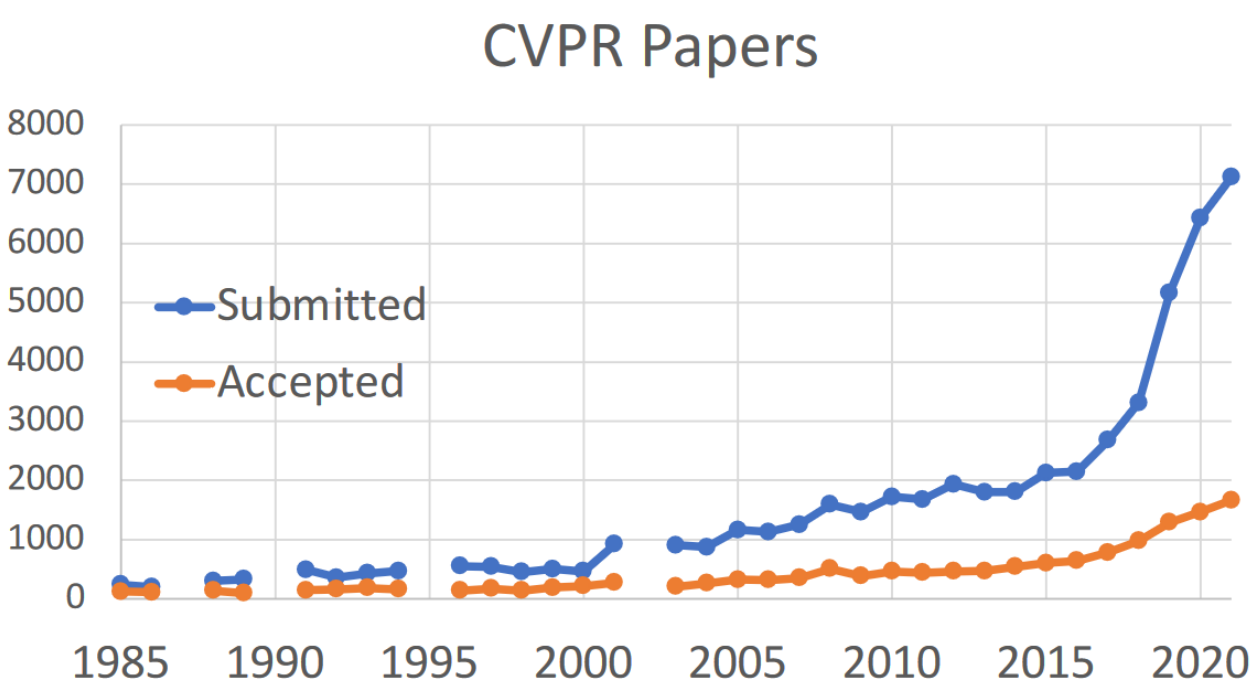


AlexNet: Deep Learning Goes Mainstream

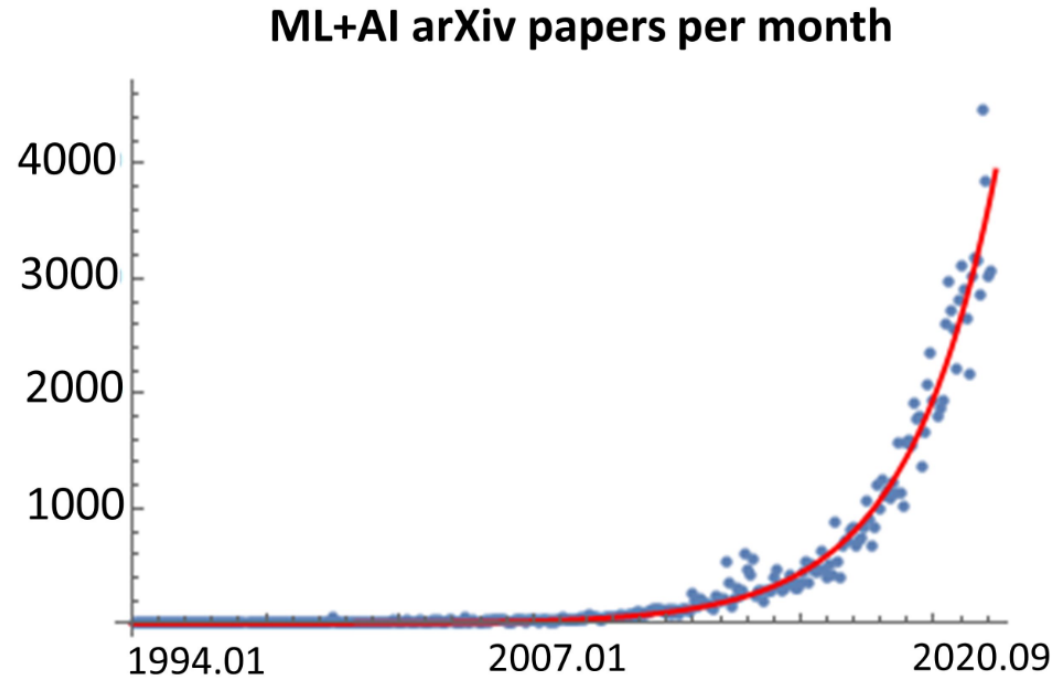


Krizhevsky, Sutskever, and Hinton, NeurIPS 2012

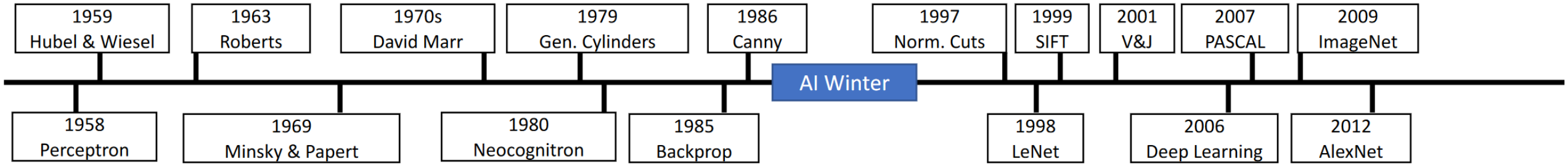
2012 to Present: Deep Learning Explosion

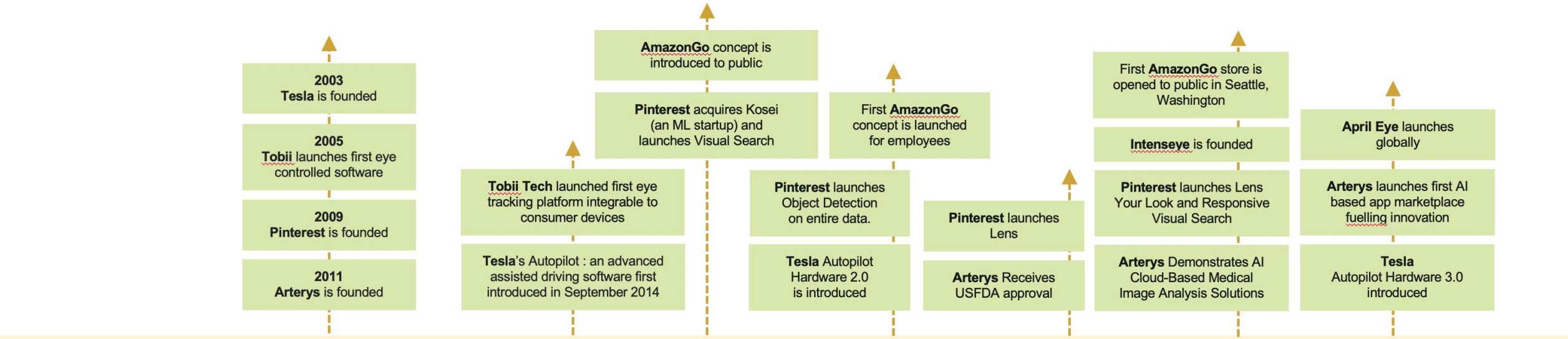


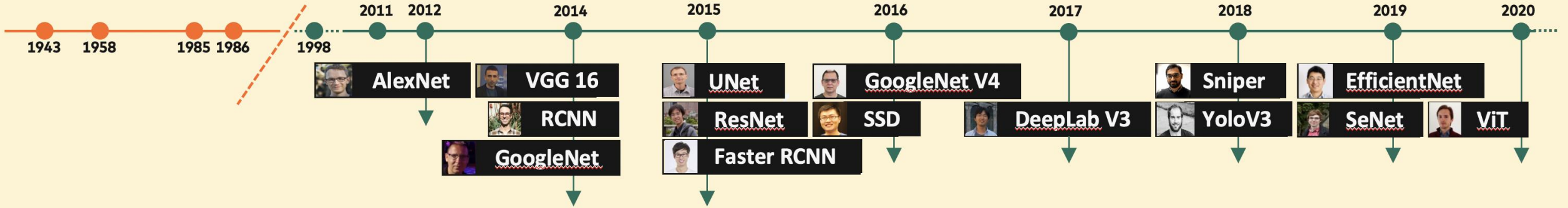
Publications at top Computer Vision conference



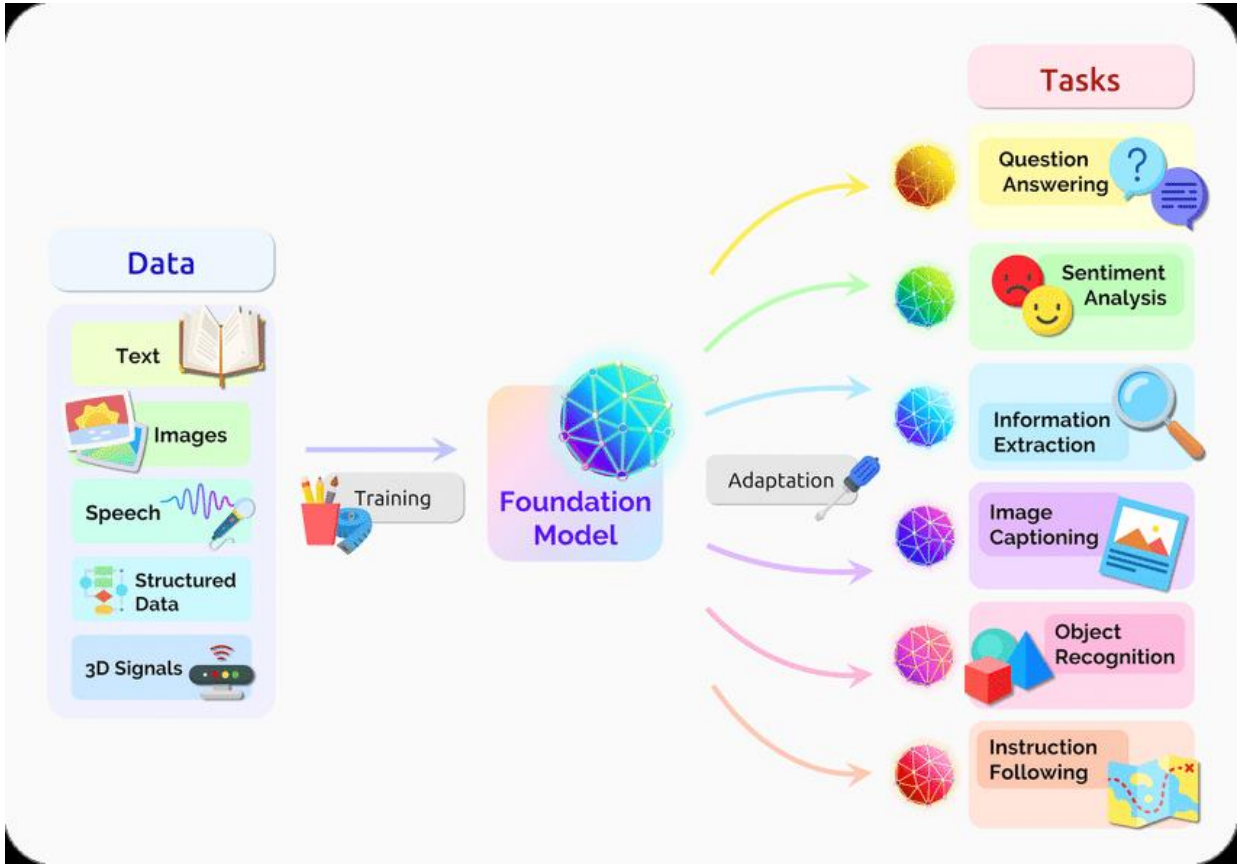
arXiv papers per month (source)


1959
Hubel & Wiesel
1963
Roberts
1970s
David Marr
1979
Gen. Cylinders
1986
Canny
1997
Norm. Cuts
1999
SIFT
2001
V&J
2007
PASCAL
2009
ImageNet
AI Winter
1958
Perceptron
1969
Minsky & Papert
1980
Neocognitron
1985
Backprop
1998
LeNet
2006
Deep Learning
2012
AlexNet


2003
Tesla is founded
2005
Tobii launches first eye controlled software
2009
Pinterest is founded
2011
Arterys is founded
Tobii Tech launched first eye tracking platform integrable to consumer devices
Tesla's Autopilot : an advanced assisted driving software first introduced in September 2014
AmazonGo concept is introduced to public
Pinterest acquires Kosei (an ML startup) and launches Visual Search
First AmazonGo concept is launched for employees
Pinterest launches Object Detection on entire data.
Tesla Autopilot Hardware 2.0 is introduced
Pinterest launches Lens
Arterys Receives USFDA approval
First AmazonGo store is opened to public in Seattle, Washington
Intenseye is founded
Pinterest launches Lens Your Look and Responsive Visual Search
Arterys Demonstrates AI Cloud-Based Medical Image Analysis Solutions
April Eye launches globally
Arterys launches first AI based app marketplace fuelling innovation
Tesla Autopilot Hardware 3.0 introduced


1943
1958
1985
1986
1998
2011
2012
2014
2015
2016
2017
2018
2019
2020
AlexNet
VGG 16
RCNN
GoogleNet
UNet
ResNet
Faster RCNN
GoogleNet V4
SSD
DeepLab V3
Sniper
YoloV3
EfficientNet
SeNet
ViT

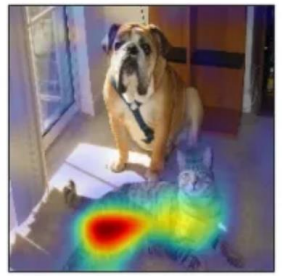
Multi-modal foundation models – current trends



Can we learn a common and generalized model from multiple modalities of data, that can solve different downstream tasks all at once?

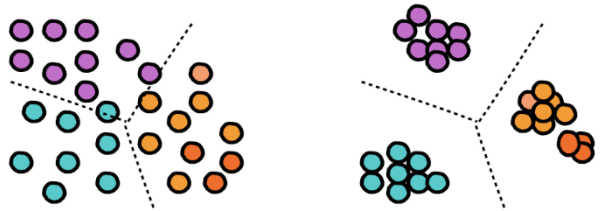
Course syllabus

- Image recognition
  - Bag of words – the state of the art before deep learning
  - Neural Networks and Convolutional Networks – image classification
    - Architecture
    - Learning algorithms
    - Optimizers
    - Efficient architectures
    - Explainability / interpretability
    - Image classification / object detection / retrieval
    - Idea of improved feature learning (metric learning)



(c) Grad-CAM 'Cat'

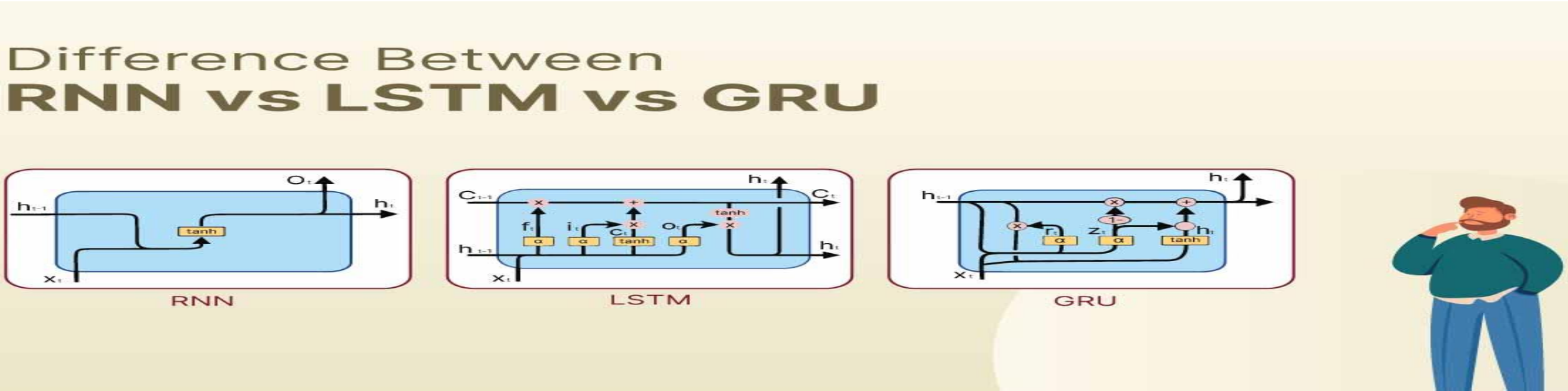
(i) Grad-CAM 'Dog'



Separable Features (e.g. classification) Discriminative Features (e.g. metric learning)

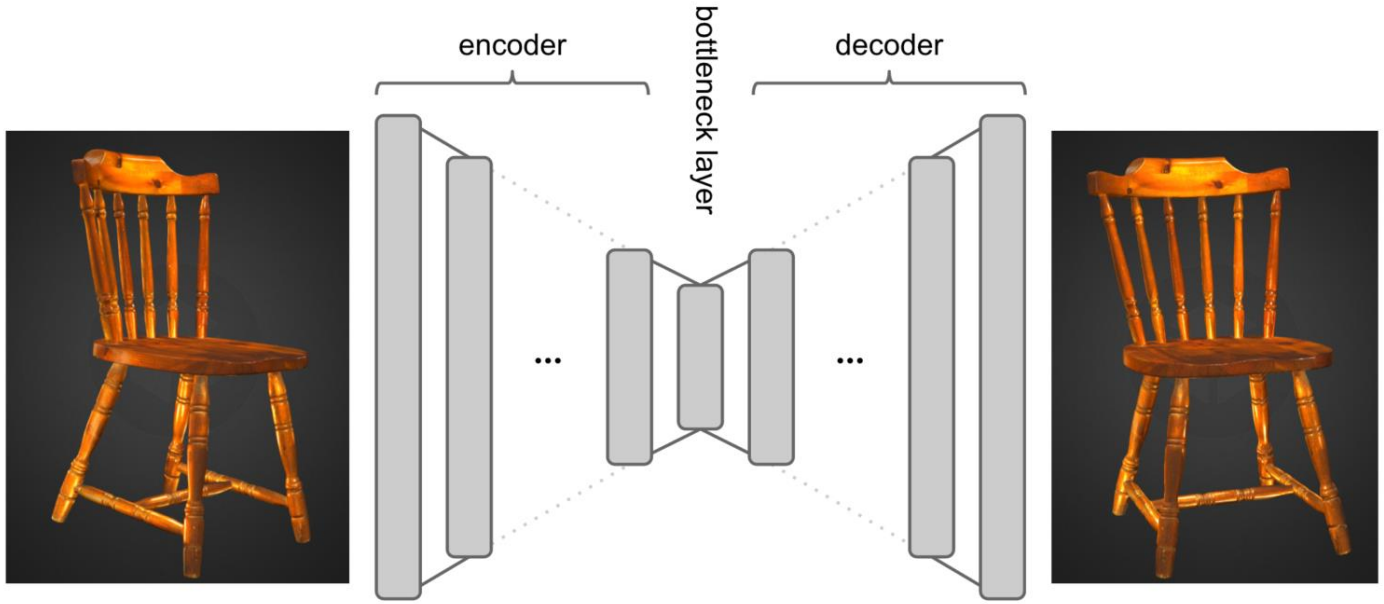
Course syllabus

- Recurrent networks
  - Training RNN – back-propagation through time
  - LSTM
  - GRU



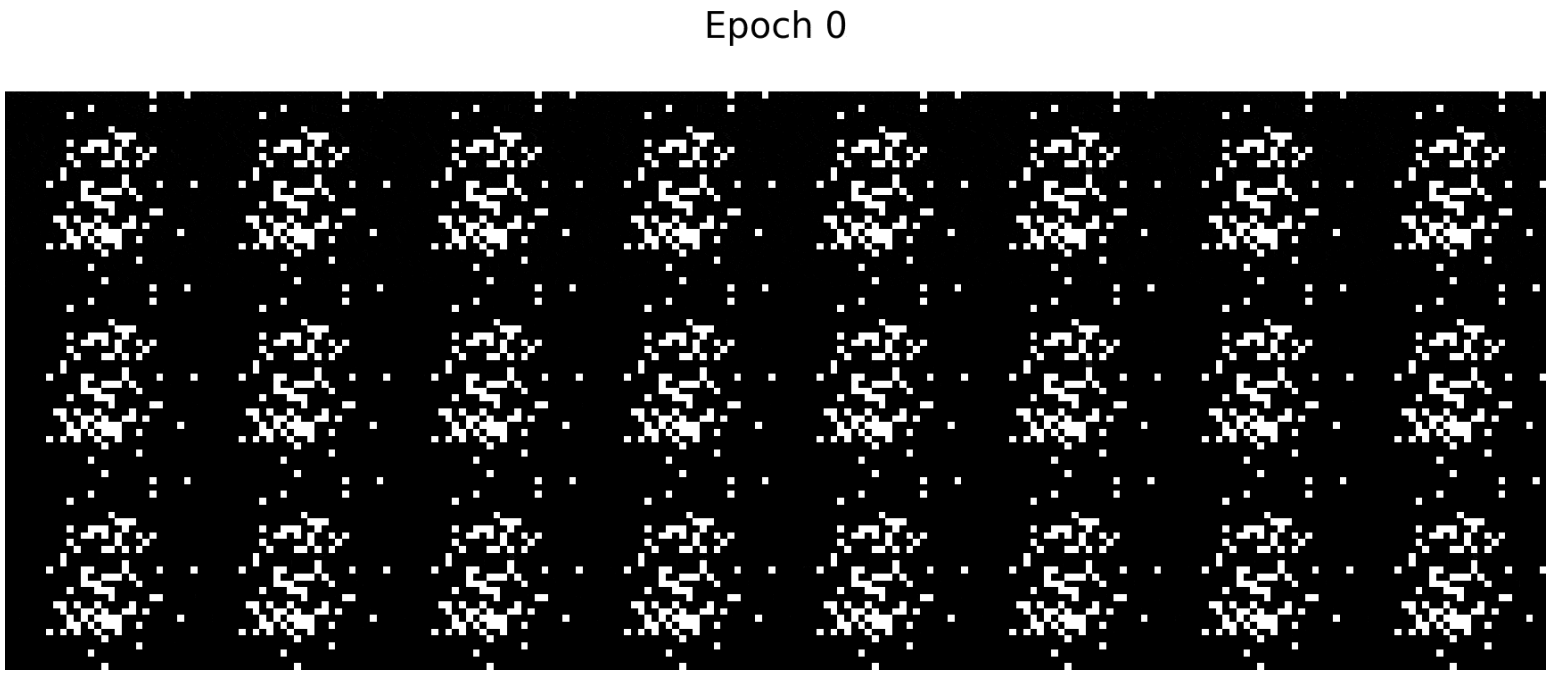
Course syllabus

- Deep Representation learning
  - Regularization in encoder-decoder model
    - Sparse
    - Denoising
    - Contrastive
    - Stacked
    - Idea of U-Net type models



Course syllabus

- Deep generative models
  - Variational auto-encoder
  - GANs
    - Vanilla GAN
    - LS GAN
    - Info GAN
    - W-GAN…



What we won’t cover in this course and are part of GNR 650

- ViT
- Diffusion models
- Self-supervised learning
- Vision-language models / tasks / foundation models

Evaluation

- 4 quizzes– (5 x 4 = **20**%)
- Mid-sem – **20**%
- Multiple coding assignments – **30%** (2 students form a group, and will update codes in a git page and share the git with the TA, the git should include the codes, analysis etc)
- End-sem – **30**%

- **1.5 marks deducted for every 3 missing classes – attendance is important!!**

Requirements for audit

- Appear for 50% evaluation

E - Office hours & TA

- Thursday 5-6PM
- gnr638_2025@googlegroups.com (for mail communication)
- Slack: gnr638_2025
- Moodle will be used to share the reading materials and other resources

Resources

- Online lectures (Michigan, NYU, TUM)

- https://d2l.ai/d2l-en.pdf
- https://www.cs.princeton.edu/courses/archive/fall19/cos597B/lecnotes/bookdraft.pdf
- https://github.com/janishar/mit-deep-learning-book-pdf
- https://tanthiamhuat.files.wordpress.com/2018/03/deeplearningwithpython.pdf

Pytorch, Tensorflow, Keras

- Pytorch – zero to GAN
  - https://www.youtube.com/watch?v=GIsg-ZUy0MY
- Keras by deepLizard
  - https://www.youtube.com/watch?v=qFJeN9V1ZsI
- Tensorflow tutorial
  - https://www.youtube.com/watch?v=tPYj3fFJGjk