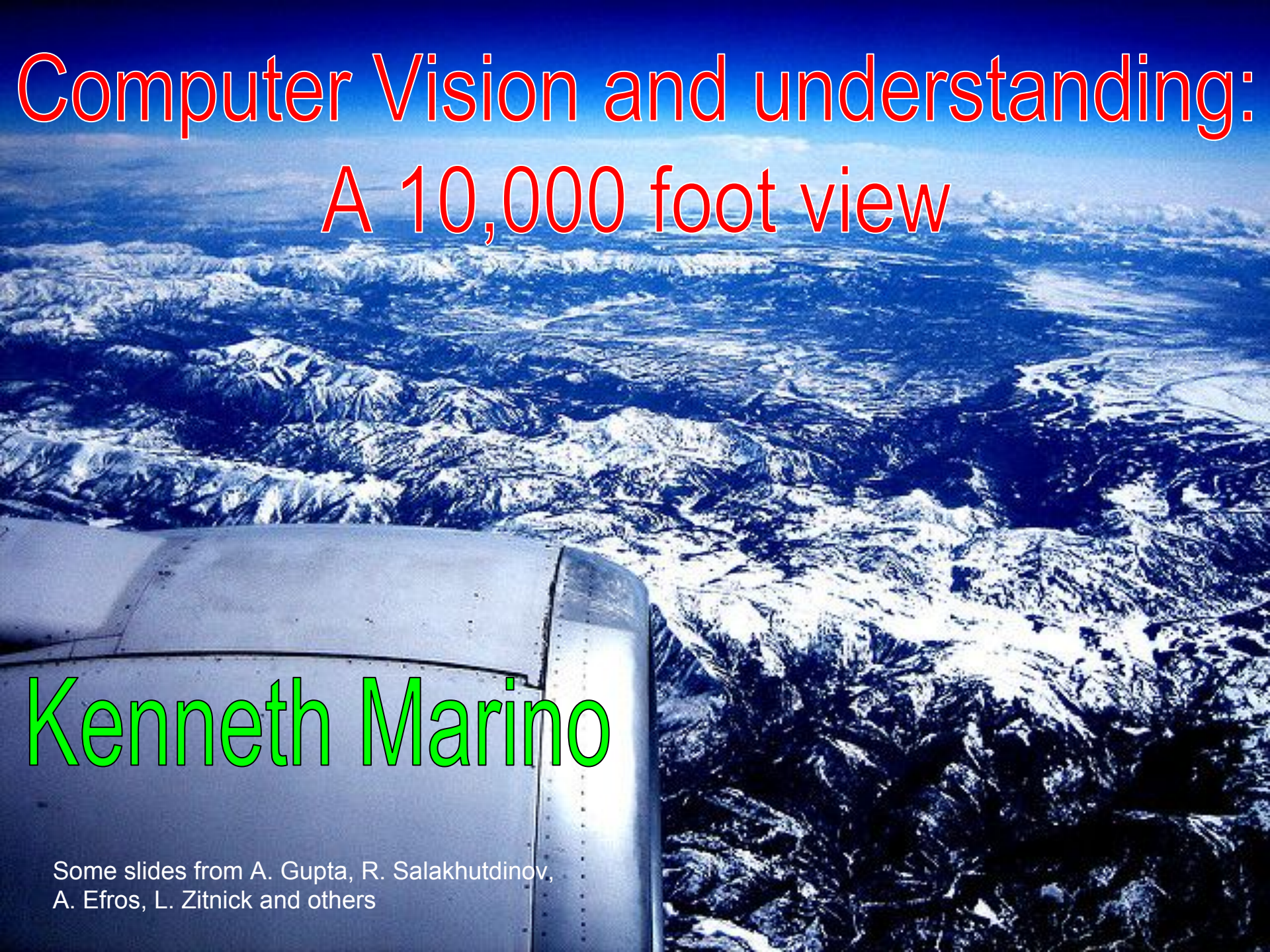

Computer Vision and understanding:
A 10,000 foot view
Kenneth Marino
Some slides from A. Gupta, R. Salakhutdinov,
A. Efros, L. Zitnick and others

Why should we care about Computer Vision

Computer Vision / Deep Learning is Everywhere

- Google, Facebook, Uber, Apple
  - Strong deep learning / computer vision groups hiring everywhere..
  - Beyond Research: Development
    - Image Search
    - Automated Driving

Startups Sold Everyday

- Vision Factory, EuVision, Flutter….

Computer Vision Works!

- Surprisingly recent development


Source: http://www.youtube.com/watch?v=YGO2IwAgmg

What is the goal of Computer Vision

What is the goal of Computer Vision

To create autonomous systems that "understand" visual data

What does it mean to understand?

Early days of Computer Vision

Early days of Computer Vision

“What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking.”

-- David Marr, *Vision (1982)*

Early days of Computer Vision

"What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking."

-- David Marr, *Vision (1982)*

In other words, vision is the process of discovering from images what is present in the world, and where it is."

Early days of Computer Vision

Slide Credit: Abhinav Gupta

Early days of Computer Vision

Answer #1: pixel of brightness 43
at position (124,54) …and depth .7
meters

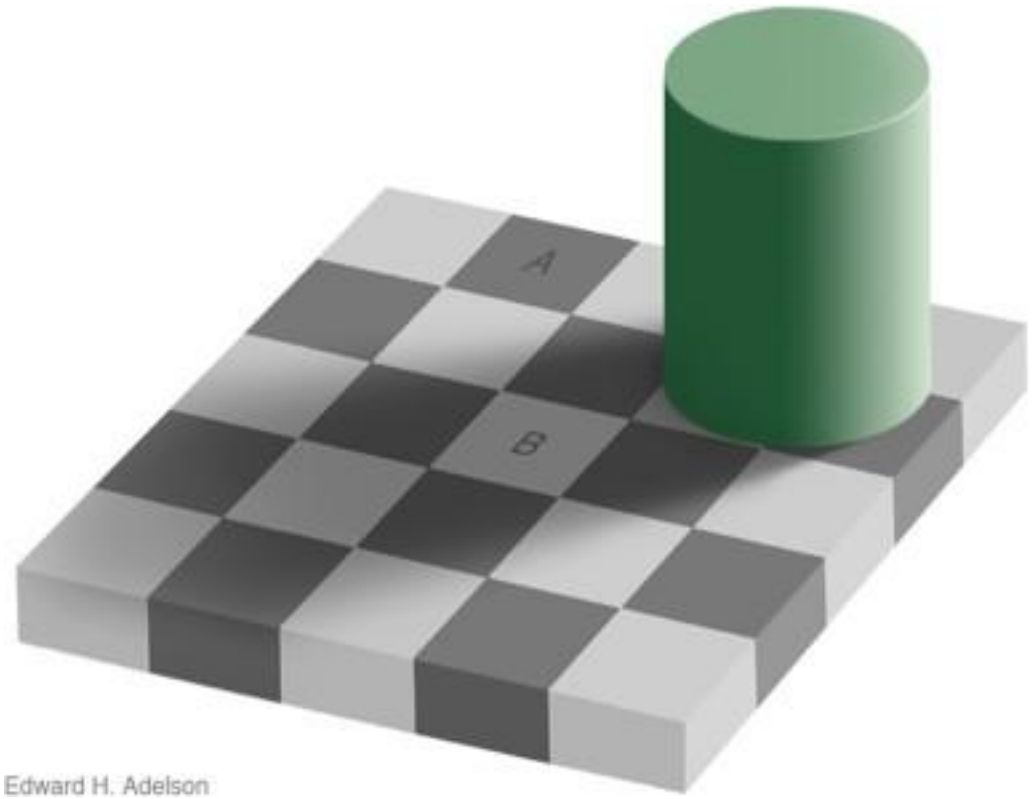
So we’re done?

So we’re done?

No!

Measurement vs. Perception

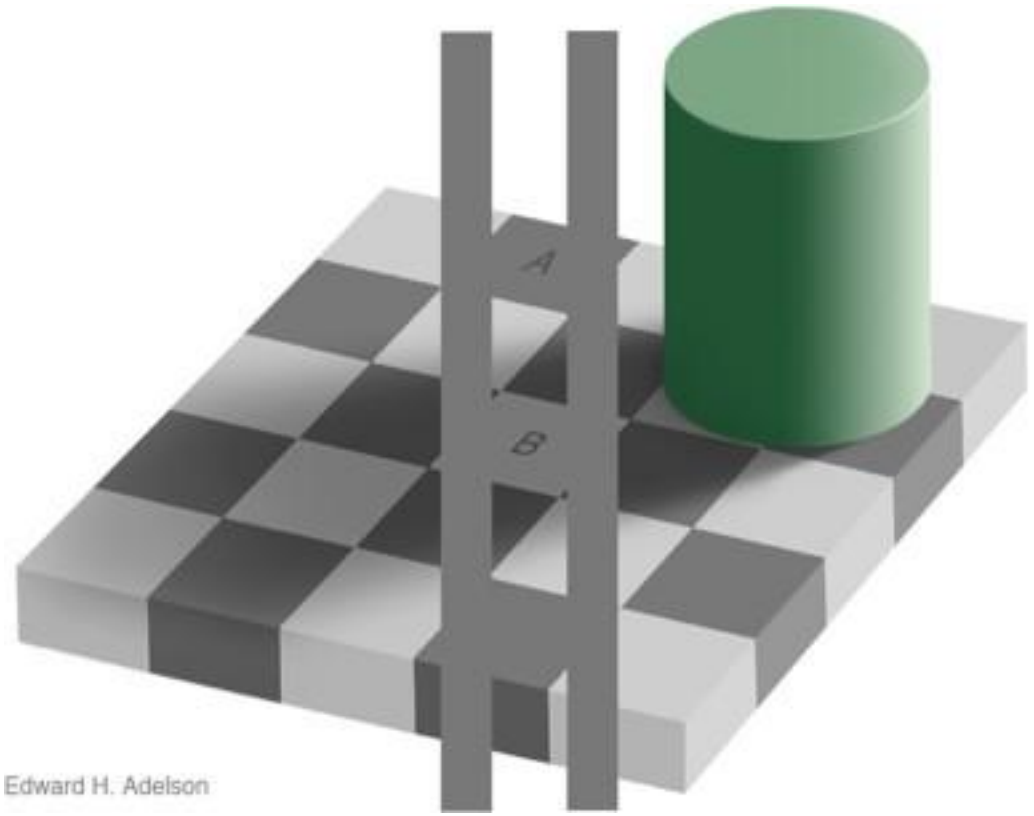
Brightness: Measurement vs. Perception



Edward H. Adelson

Slide Credit: Alyosha Efros

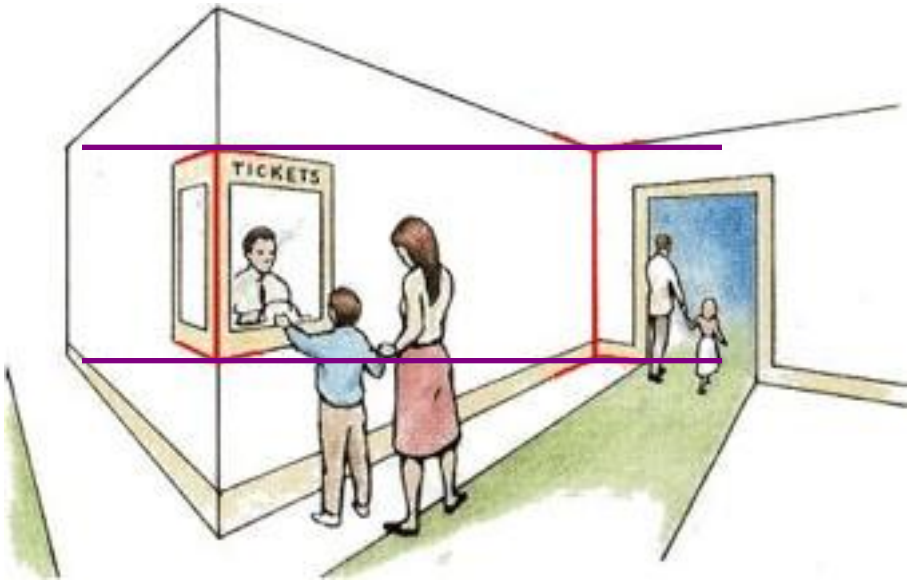
Brightness: Measurement vs. Perception



Proof!

Measurement

Length



Müller-Lyer Illusion

http://www.michaelbach.de/ot/sze_muelue/index.html

Measurement

- Capturing physical quantities like pixel brightness, depth, etc.

Perception/Understanding

- a high-level representation that captures the semantic structure of the scene and its constituent objects.

- Subjective – Depends on Task and Agent

- Intersection of what you see and what you believe (prior knowledge)

Slide Credit: Abhinav Gupta

…but why do we care about perception?

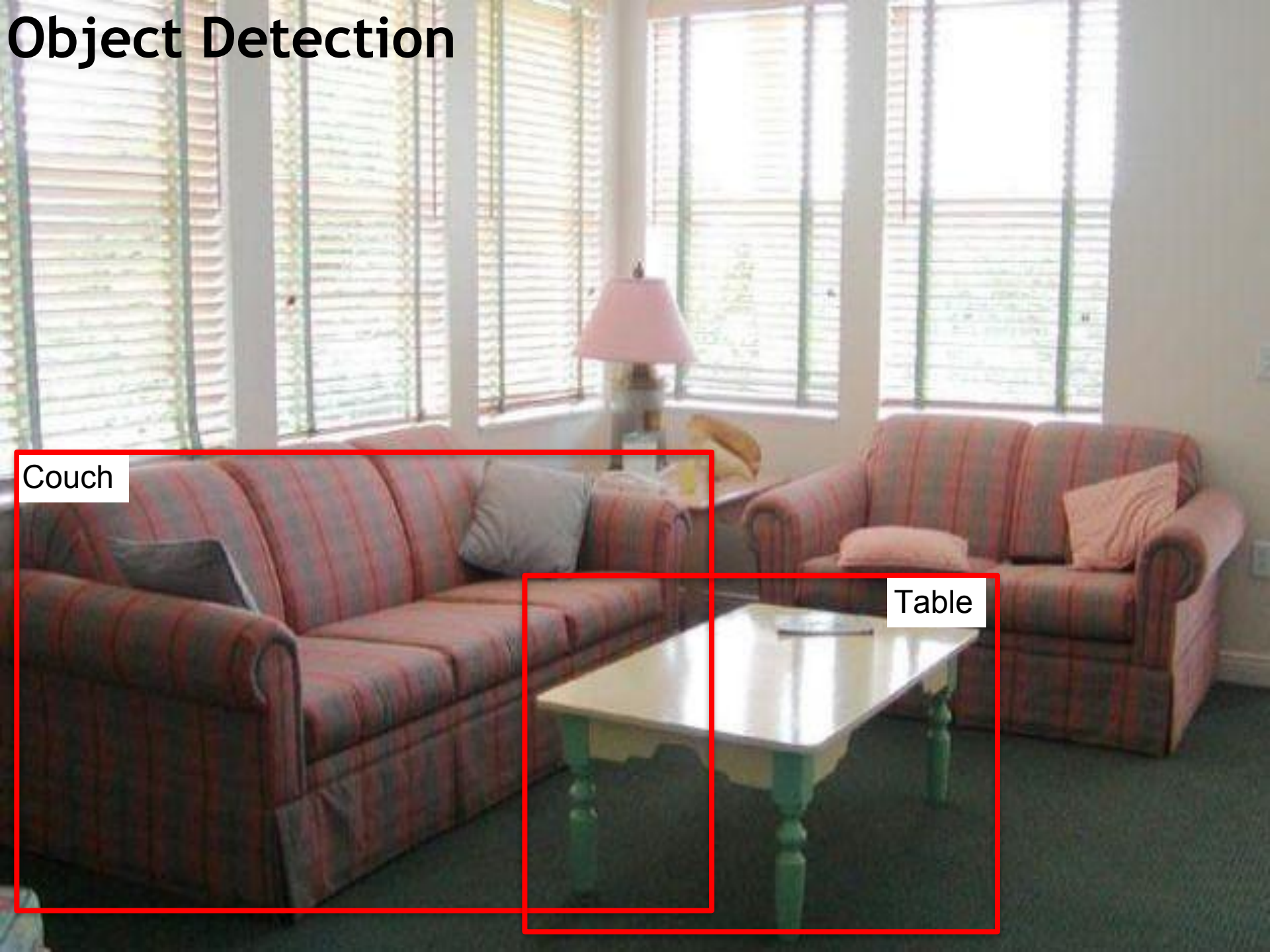
The goals of computer vision (**what + where**) are in terms of what humans care about.

So what do humans care about?

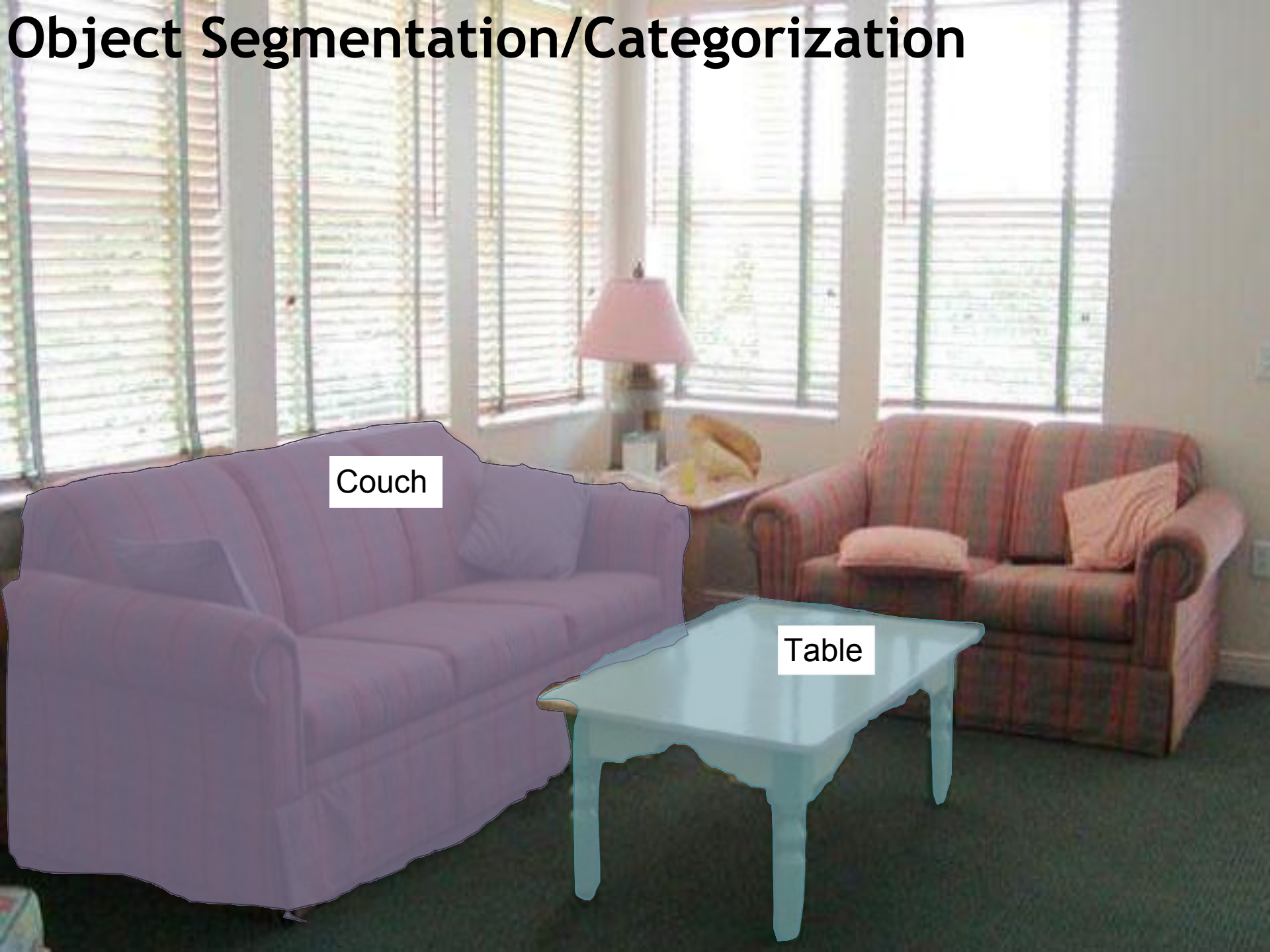
Slide Credit: Abhinav Gupta

Image Classification/ Scene Recognition

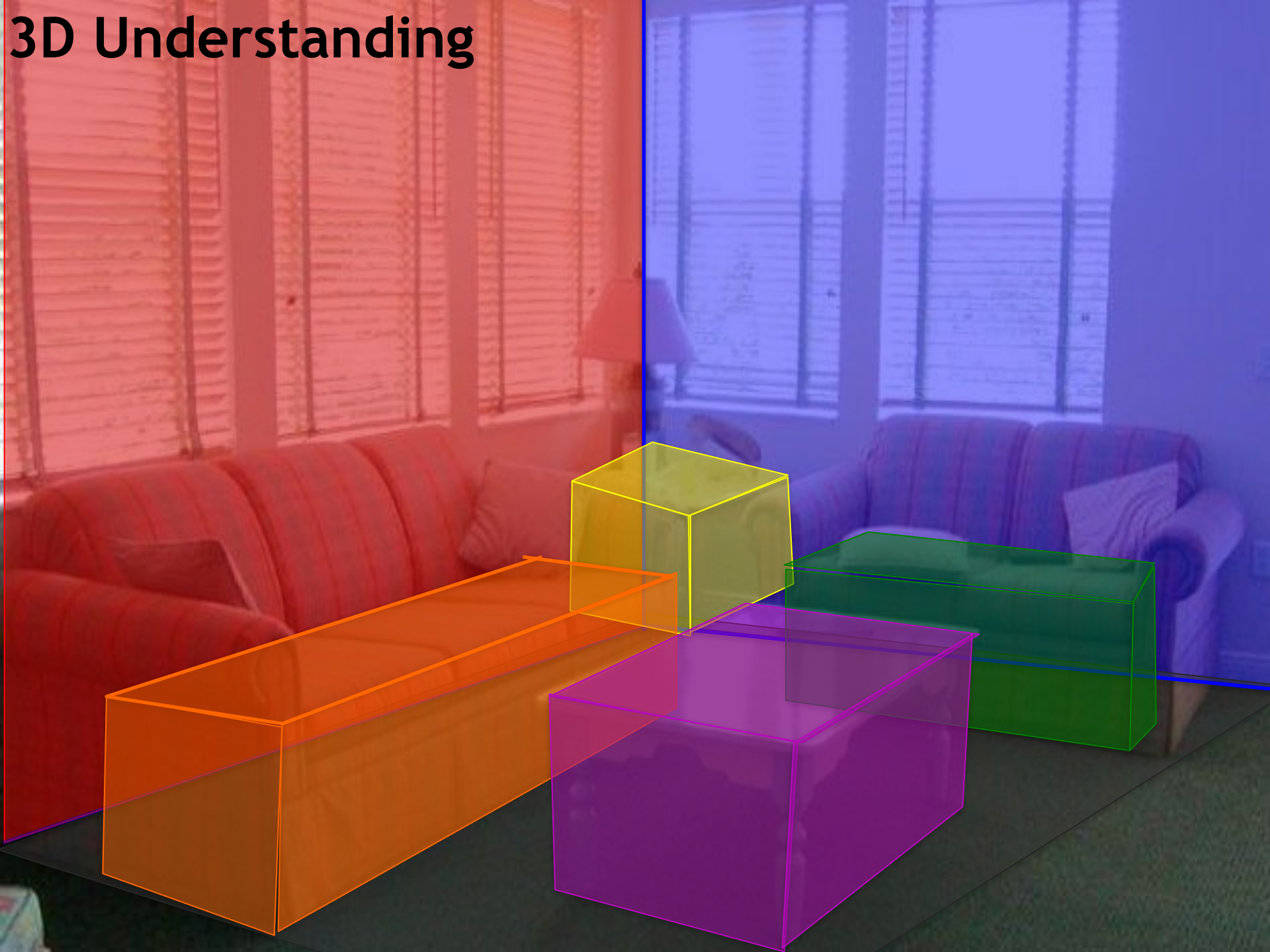
Object Detection

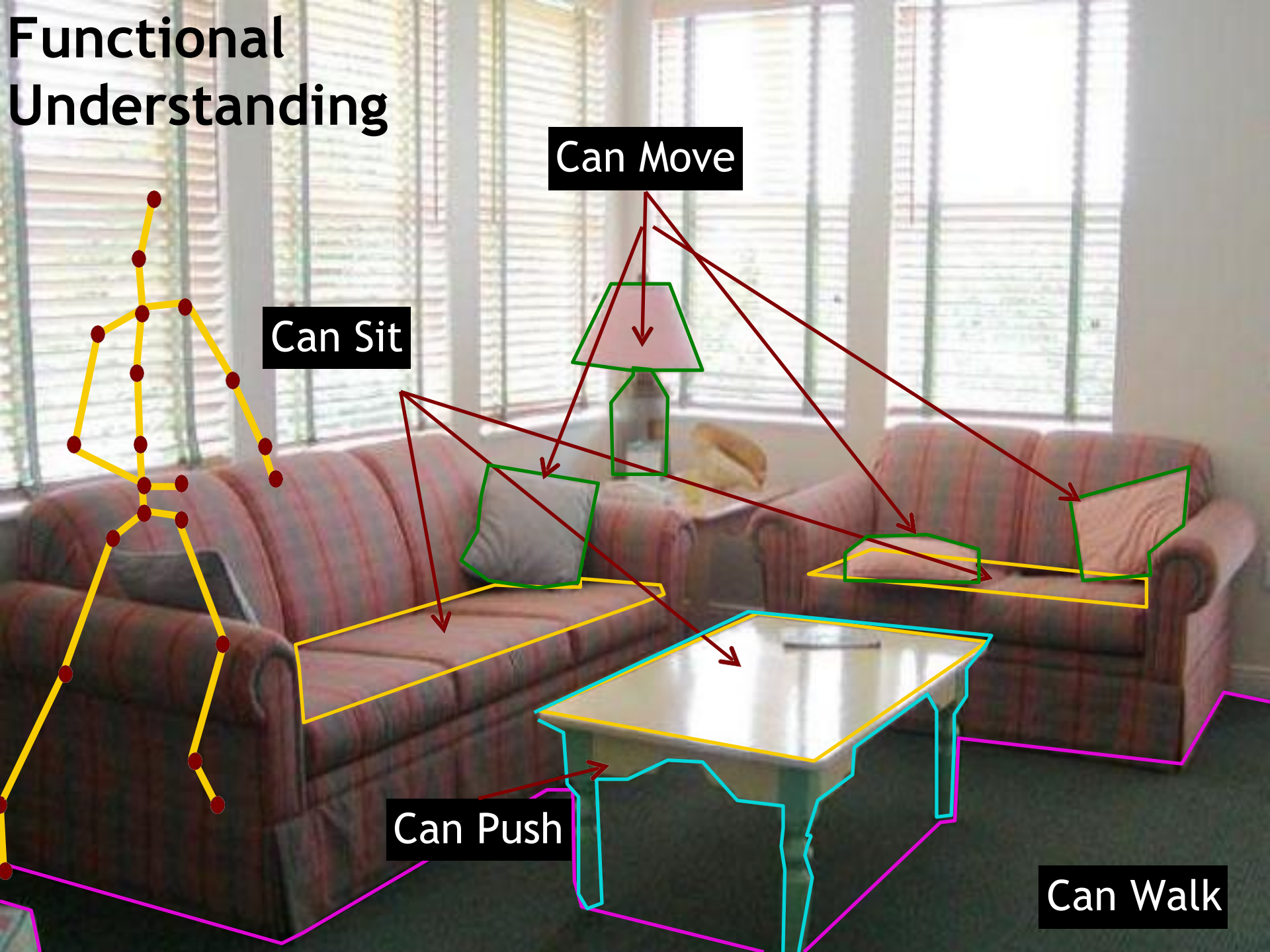


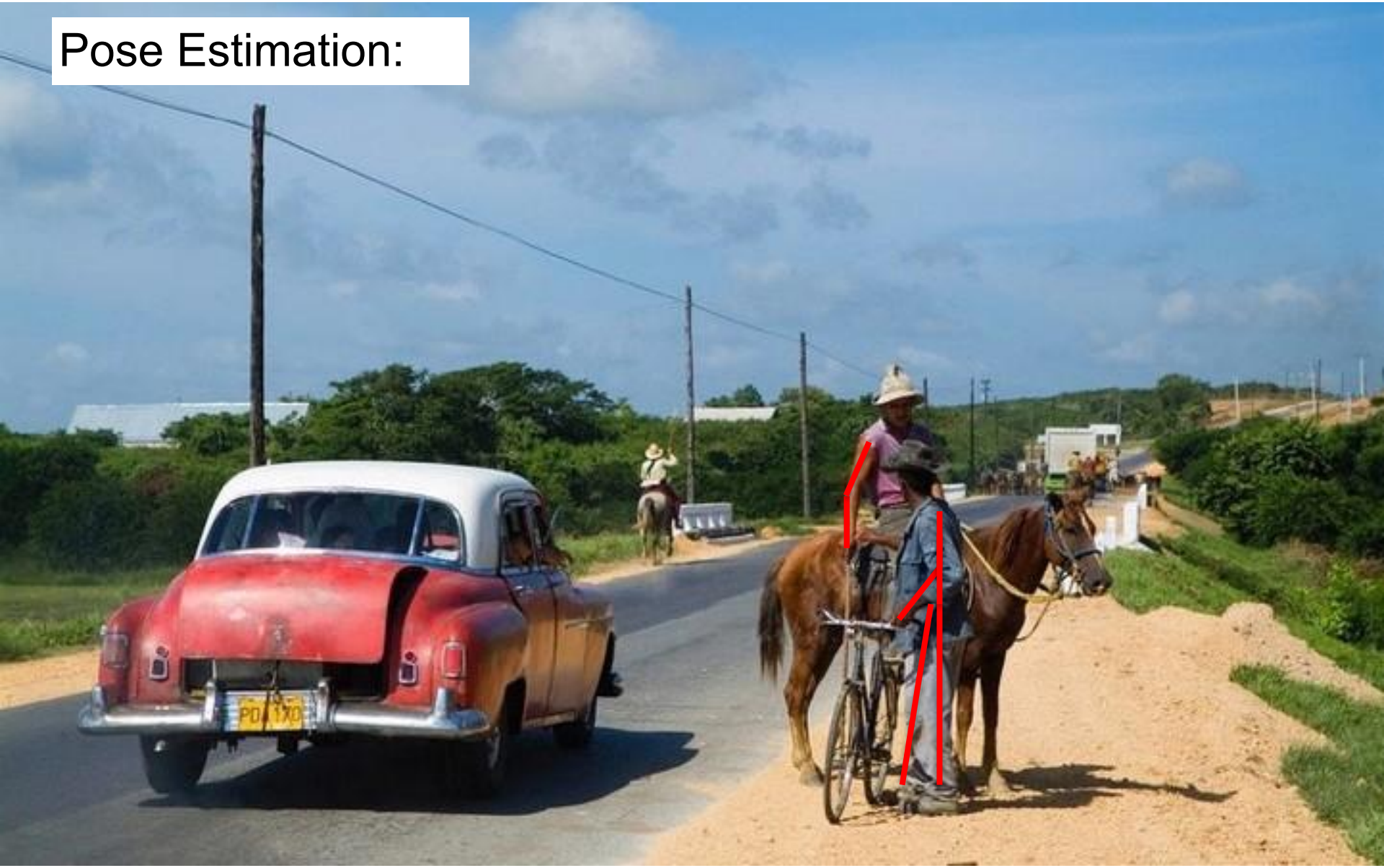
Object Segmentation/Categorization



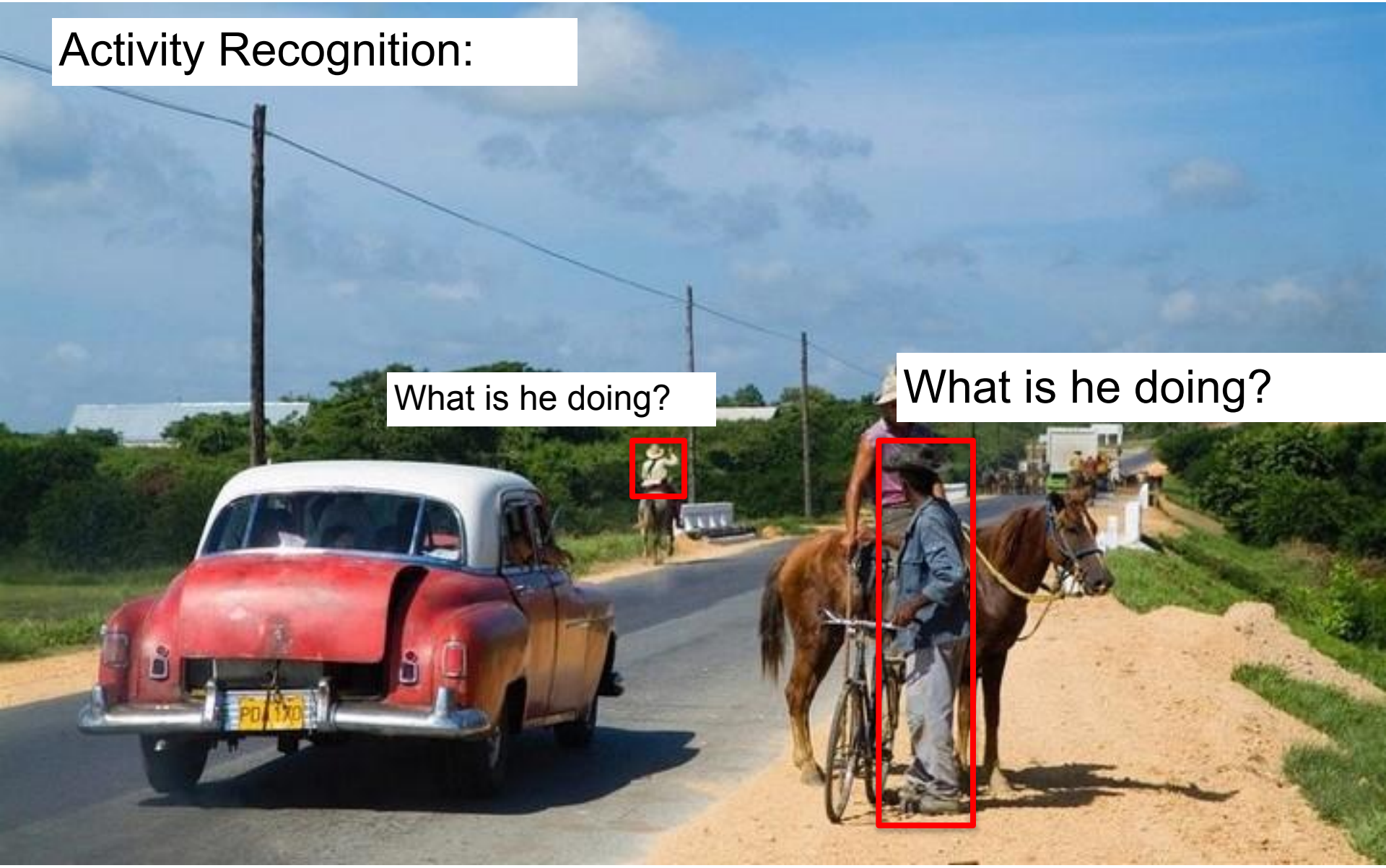
3D Understanding




Functional Understanding
Can Move
Can Sit
Can Push
Can Walk



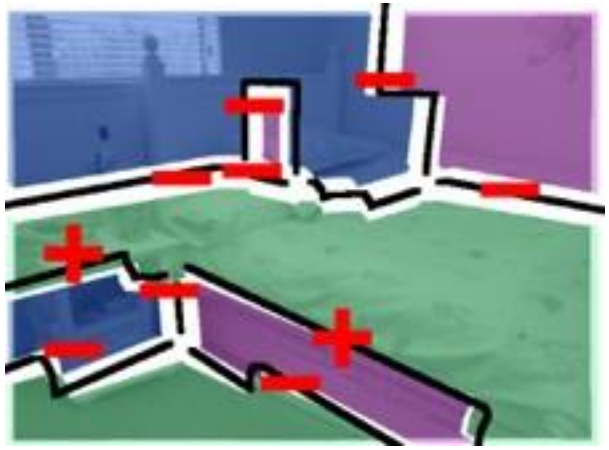
Slide Credit: Abhinav Gupta

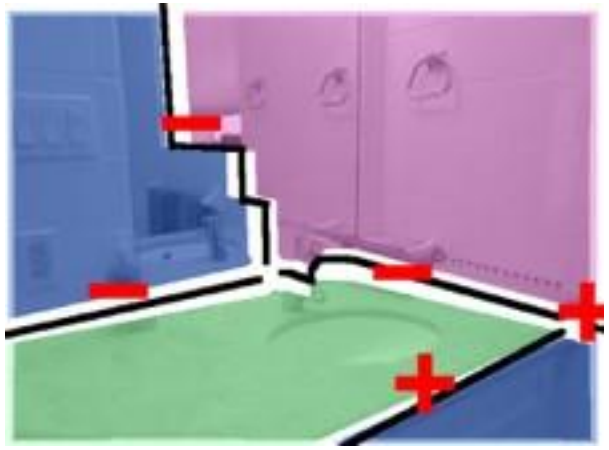

Activity Recognition:
What is he doing?
What is he doing?

Surface Normal Segmentation

Input Image

Surface Connection Graph





Surface Normal Estimation

Slide Credit: Abhinav Gupta

Why are these problems hard?

Challenges 1: view point variation

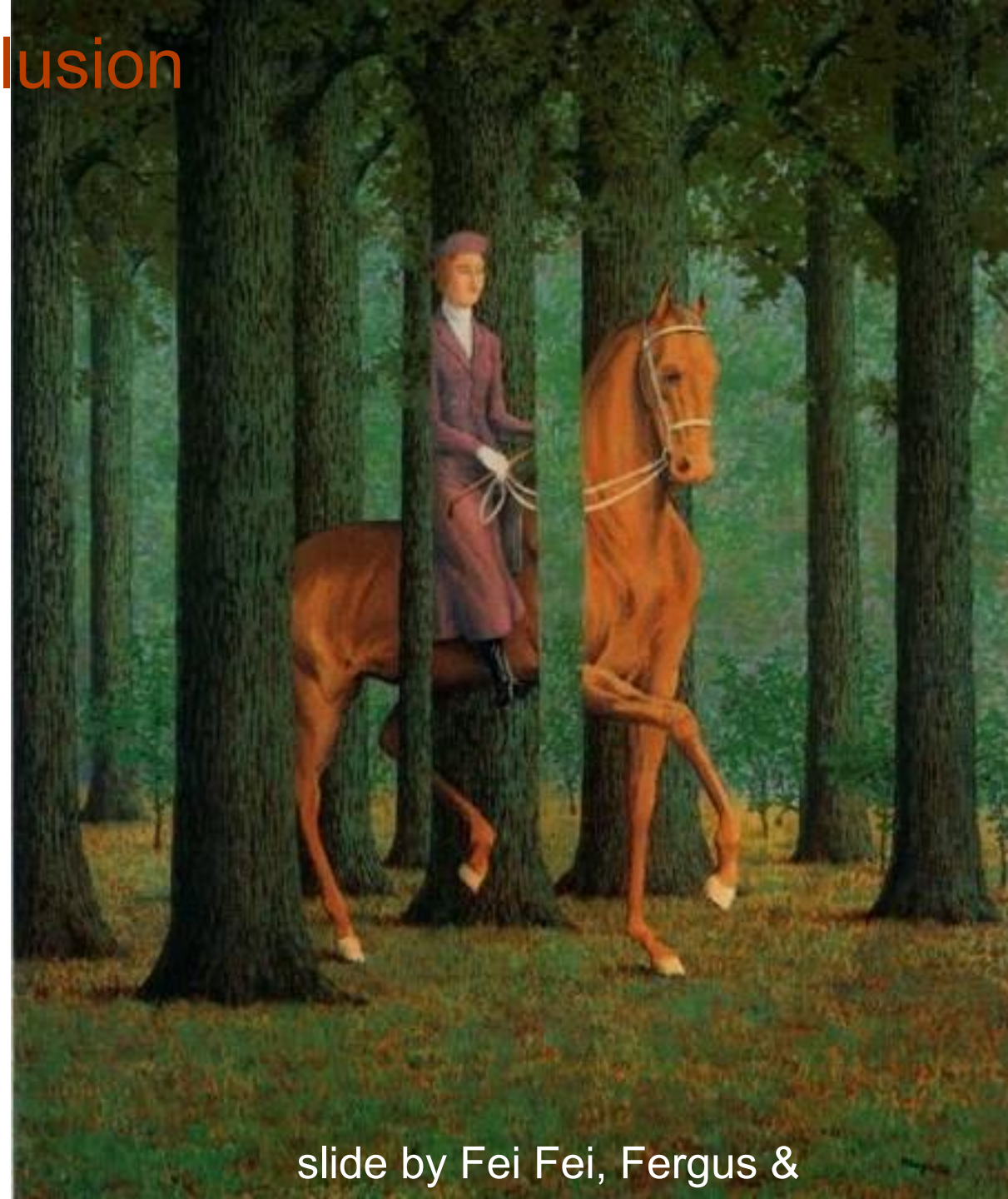
Michelangelo 1475-1564

slide by Fei Fei, Fergus &

Challenges 2: illumination

slide credit: S. Ullman

Challenges 3: occlusion



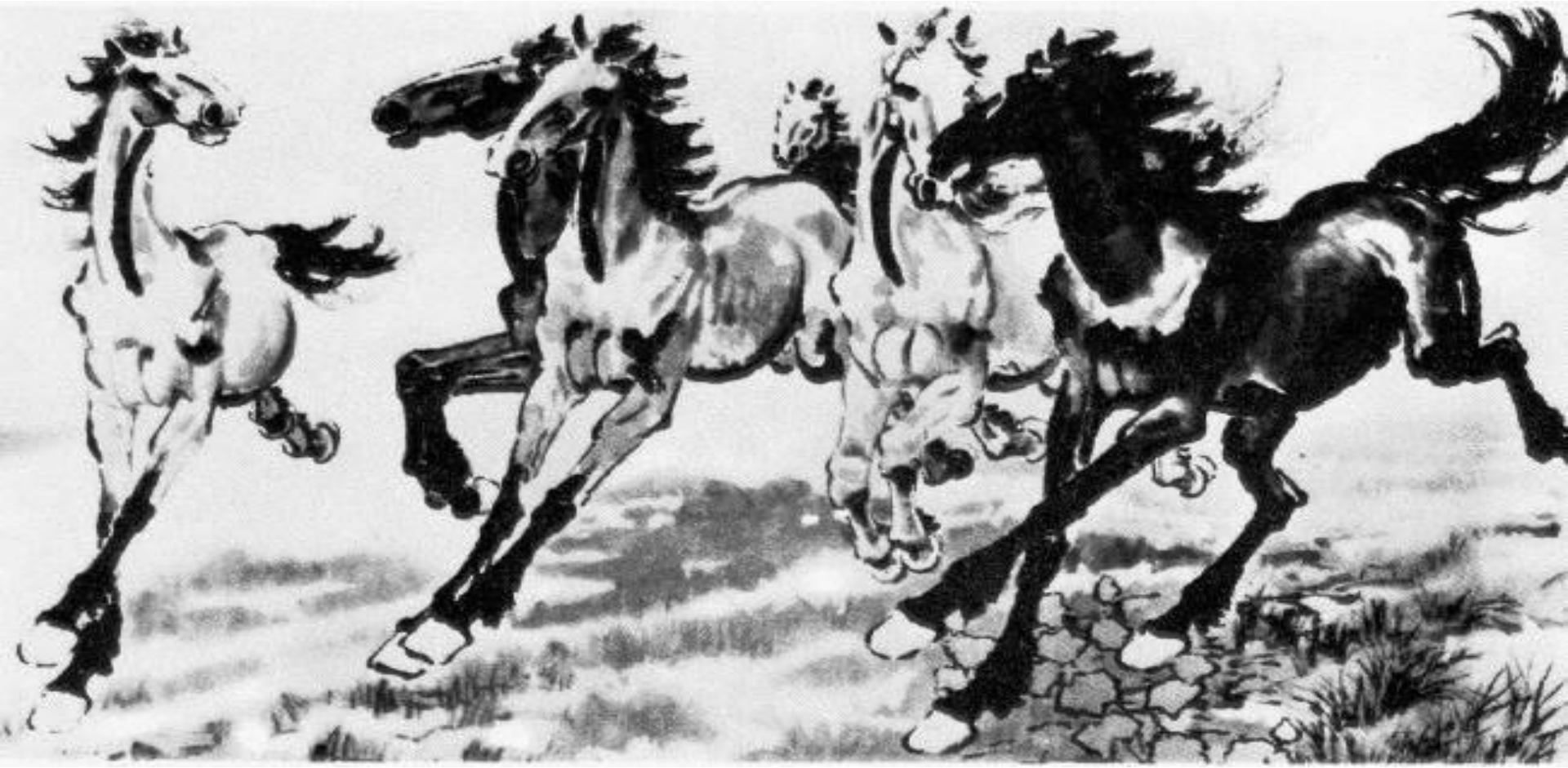
Magritte, 1957

slide by Fei Fei, Fergus &

Challenges 4: scale

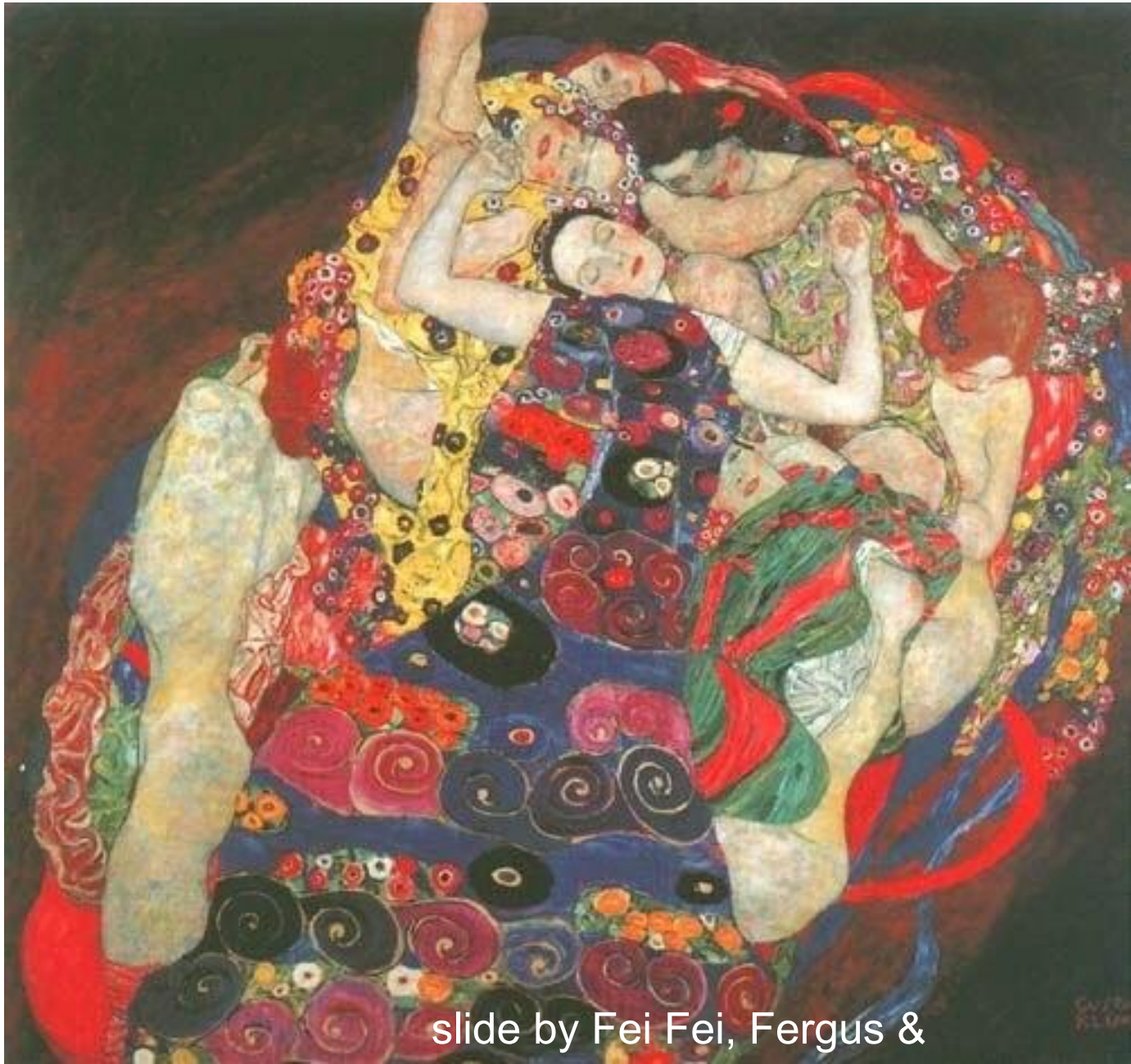
slide by Fei Fei, Fergus &

Challenges 5: deformation



Xu, Beihong 1943

Challenges 6: background clutter



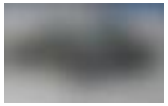
Klimt, 1913

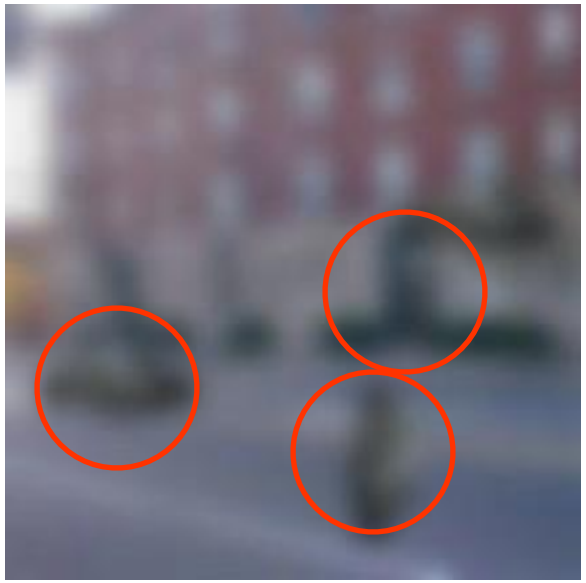
slide by Fei Fei, Fergus &

Challenges 7: object intra-class variation

slide by Fei-Fei, Fergus &

Challenges 8: local ambiguity





slide by Fei-Fei, Fergus &

Challenges 9: the world behind the image

Slide Credit: Alyosha Efros

How do we solve it?

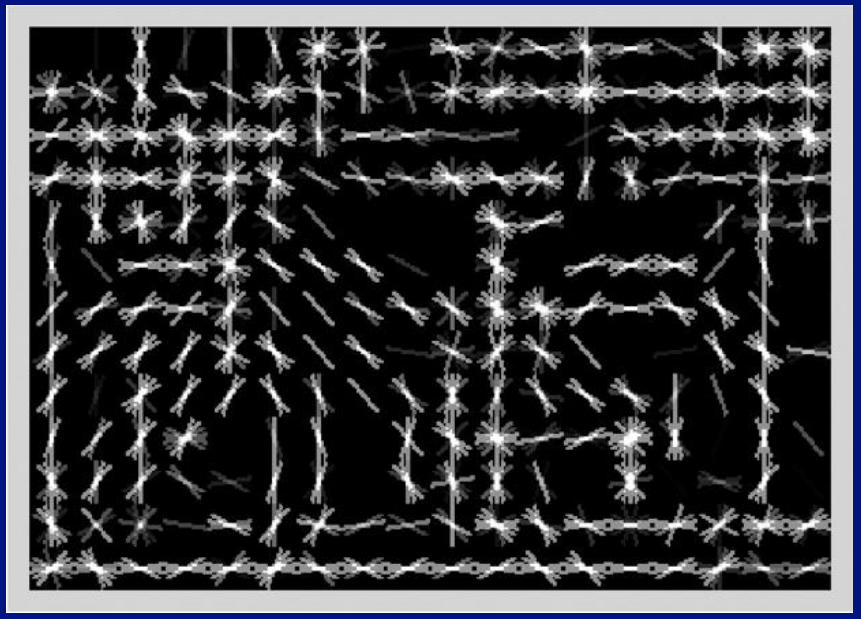
In the days of old

About 4 years ago...

In the days of old

- Take original image
- Do some feature preprocessing
  - Histogram of Oriented Gradients (HOG)
  - SIFT
  - SURF
- Run through some classifier
  - Often SVMs or Decision Trees (Specifically random forests)
    - We'll cover SVMs later in class (similar to perceptron)

Histograms of oriented gradients



From Deva Ramanan's lake Como slides

Lowe's SIFT features

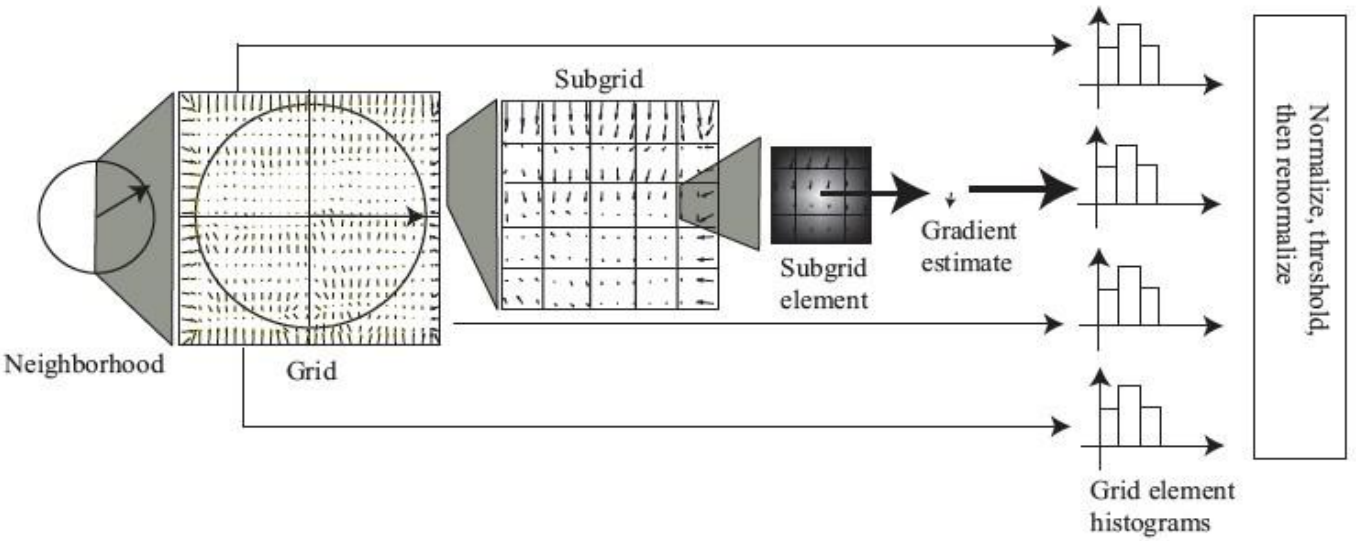


FIGURE 5.14: To construct a SIFT descriptor for a neighborhood, we place a grid over the rectified neighborhood. Each grid is divided into a subgrid, and a gradient estimate is computed at the center of each subgrid element. This gradient estimate is a weighted average of nearby gradients, with weights chosen so that gradients outside the subgrid cell contribute. The gradient estimates in each subgrid element are accumulated into an orientation histogram. Each gradient votes for its orientation, with a vote weighted by its magnitude and by its distance to the center of the neighborhood. The resulting orientation histograms are stacked to give a single feature vector. This is normalized to have unit norm; then terms in the normalized feature vector are thresholded, and the vector is normalized again.

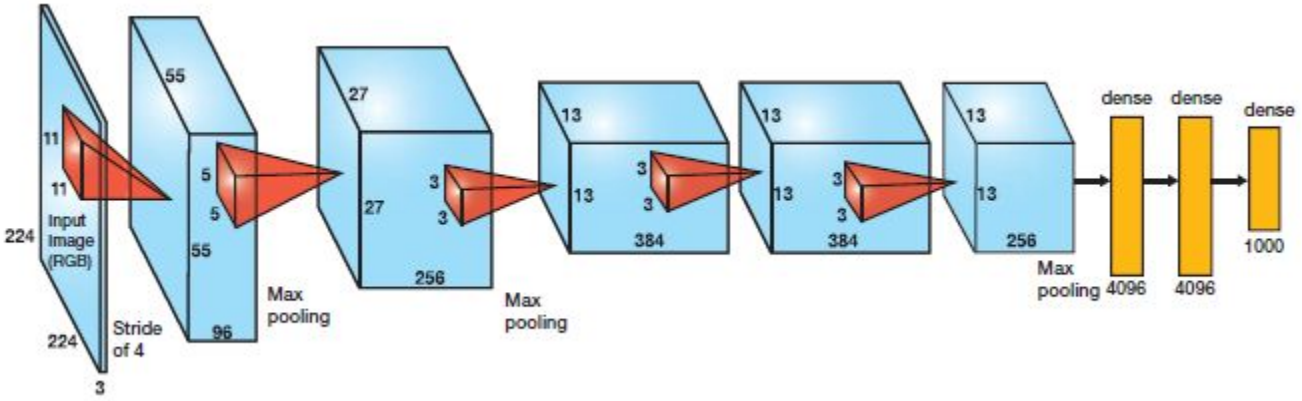
Deep Learning Era

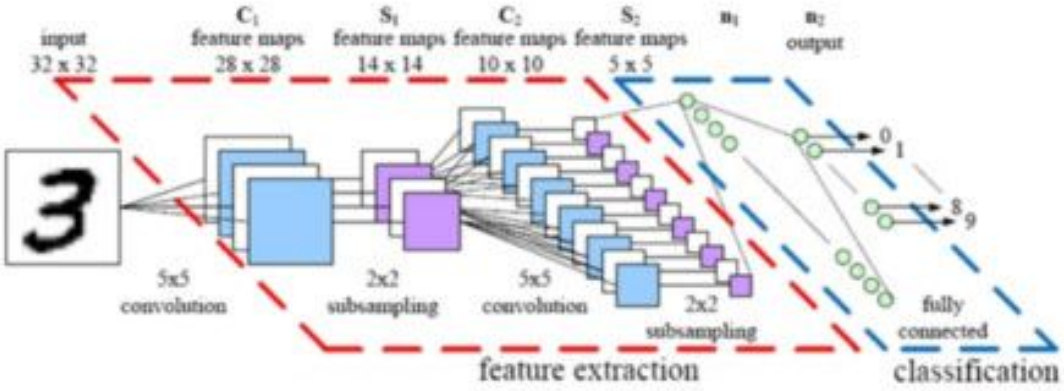
- We can actually learn better representations directly from raw pixel values!
- Run through Convolutional Neural Networks, and other types of Neural Networks
  - You'll probably cover neural networks and CNNs later in class

What Changed?

More Computational Power

Better Algorithms





Most Importantly

More Data!!!

Revisiting “Understanding”

- Is it actually enough to just know what’s in an image and where?

Q: Do you see a fruit that Gallagher would likely smash with the Sledge-O-Matic?

Clearly more here

Idea 1: Caption Generation

a car is parked in
the middle of nowhere .

a wooden table and chairs
arranged in a room .

there is a cat sitting on a shelf .

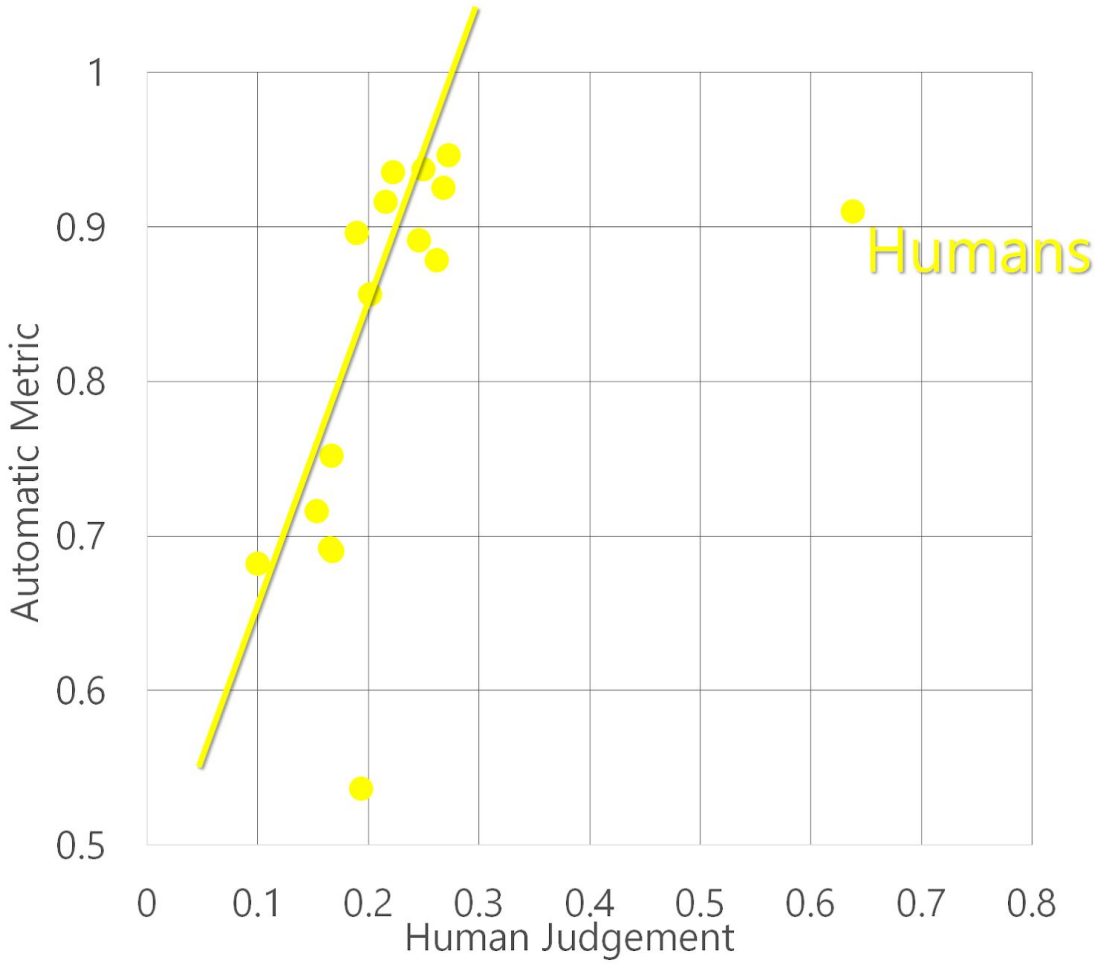
a ferry boat on a marina
with a group of people .

a little boy with a bunch
of friends on the street .

Problem: No good evaluations

Evaluation

COCO Caption Challenge

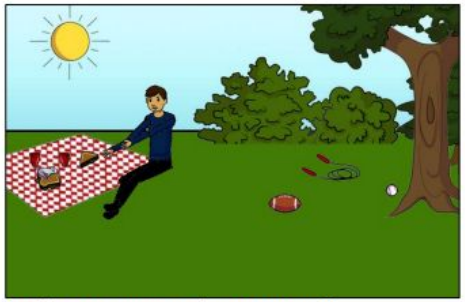


Slide Credit: Larry Zitnick

Next try: Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

- Input
  - I: An image (MS COCO)
  - Q: A question about the image
- Output
  - A: The answer to the question

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

How many other people are in the image?

I think 3. They are occluded

How many are men?

Do we need “embodiment”

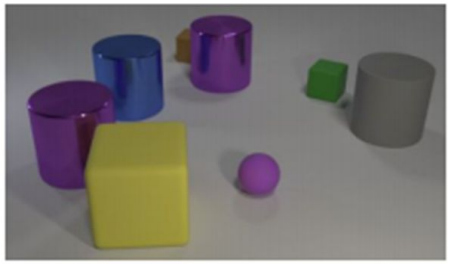
- Perhaps we can only judge how good perception / language understanding is in the context of an agent

Other Interesting Directions

Challenges: learning with minimal supervision

Kristen Grauman

Visual Reasoning

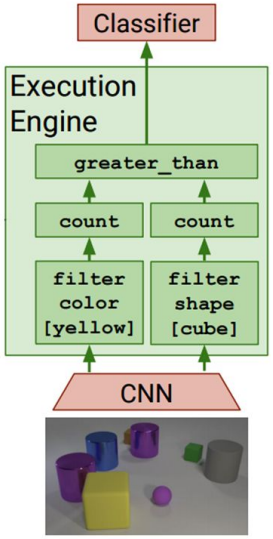


Are there more cubes than yellow things?

1. Predict program

```
greater than
count
filter color [yellow]
<SCENE>
count
filter shape [cube]
<SCENE>
```

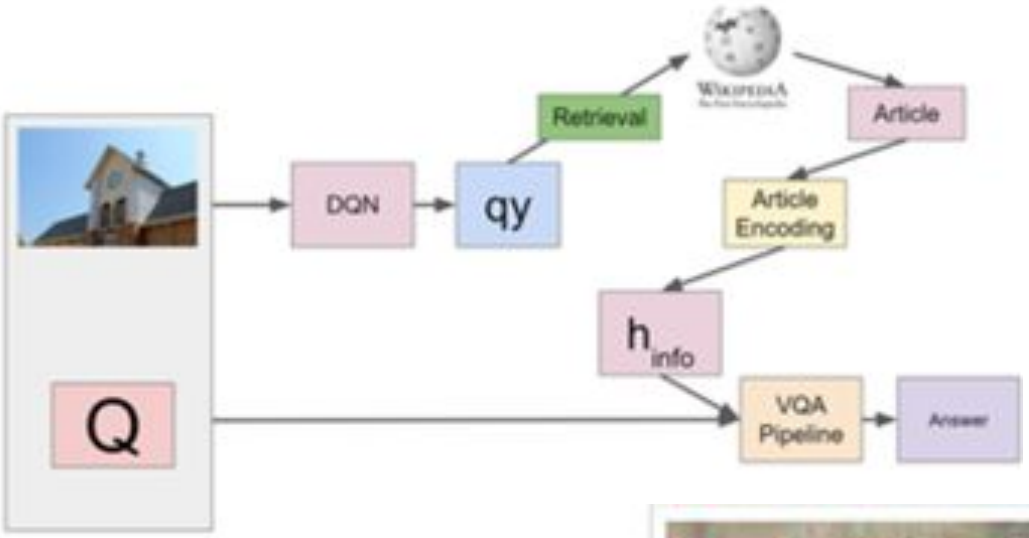
2. Execute

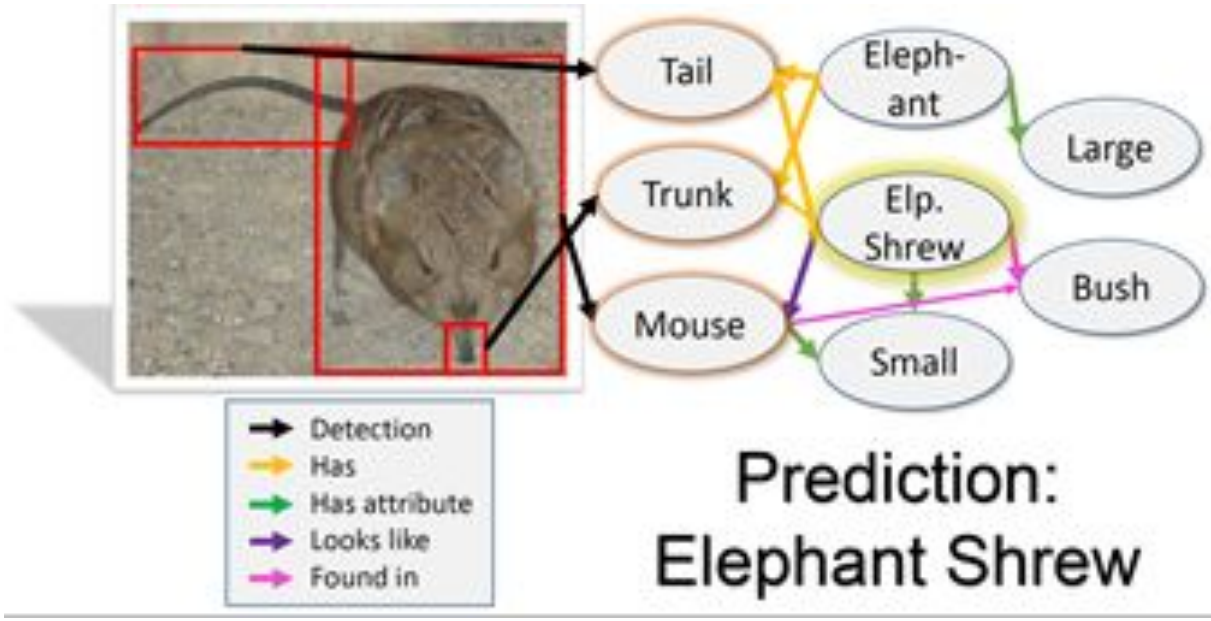


Inferring and Executing Programs for Visual Reasoning, Johnson et al., ICCV 2017

Slide Credit: Larry Zitnick

Incorporating Outside Knowledge





The more you know: Using knowledge graphs for image classification, Marino, Salakhutdinov, Gupta, CVPR17

Using Knowledge Graphs

Elephant Shrew

Elephant Shrew

- Looks like a mouse

Elephant Shrew

- Looks like a mouse
- Has a trunk

Elephant Shrew

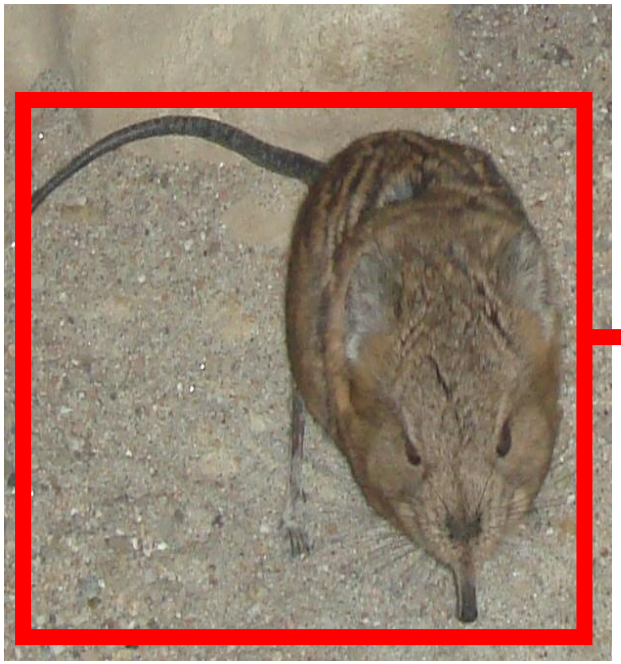
- Looks like a mouse
- Has a trunk
- Has a tail

Elephant Shrew

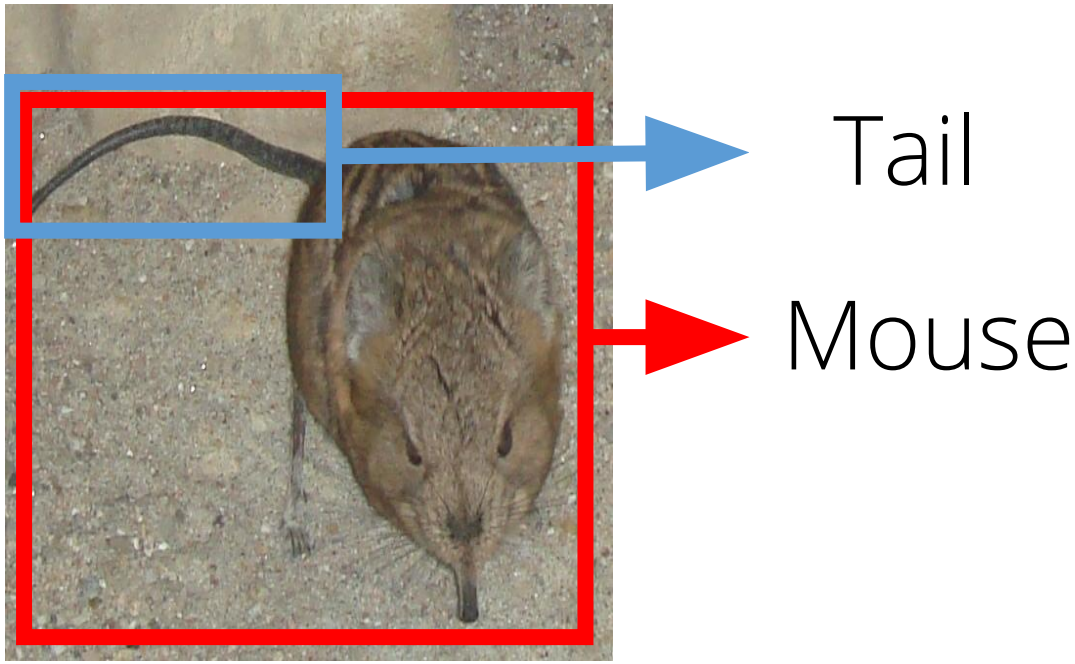
- Looks like a mouse
- Has a trunk
- Has a tail
- Is brown

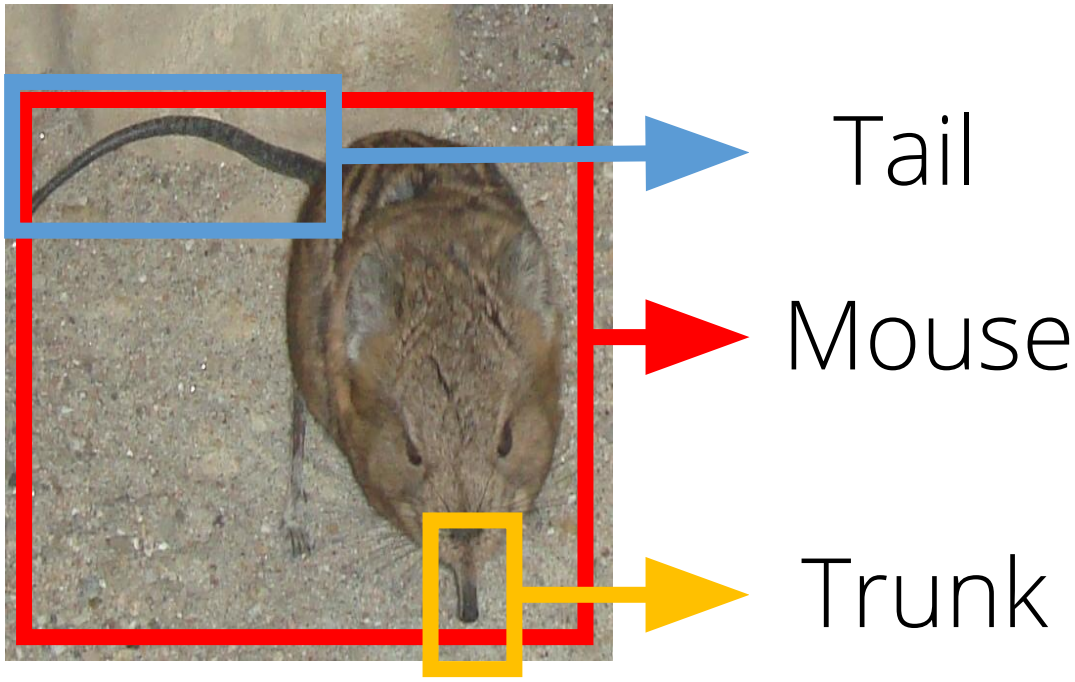
Elephant Shrew

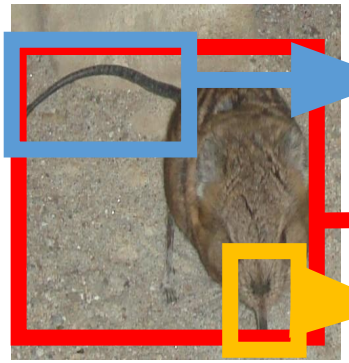
- Looks like a mouse
- Has a trunk
- Has a tail
- Is brown
- Lives in Africa



Mouse


Tail
Mouse


Tail
Mouse
Trunk



Tail

Mouse

Trunk

Elephant Shrew

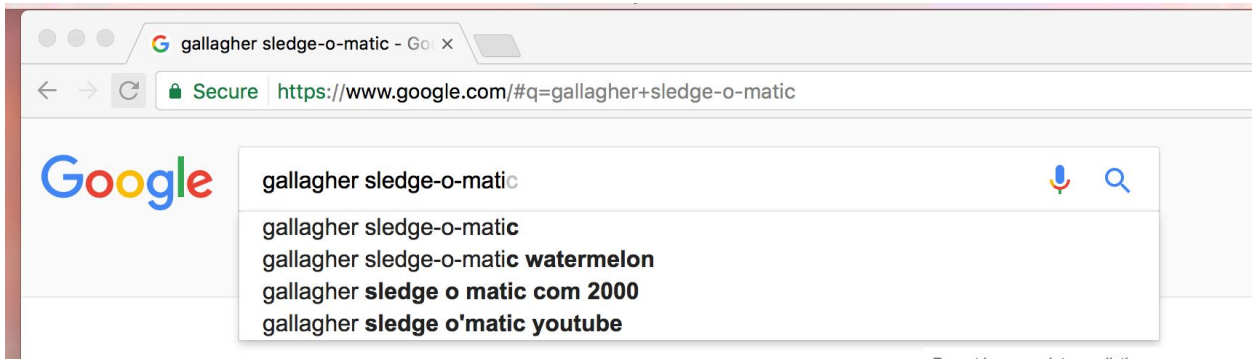
- Looks like a mouse
- Has a trunk
- Has a tail
- Is brown
- Lives in Africa

Elephant Shrew

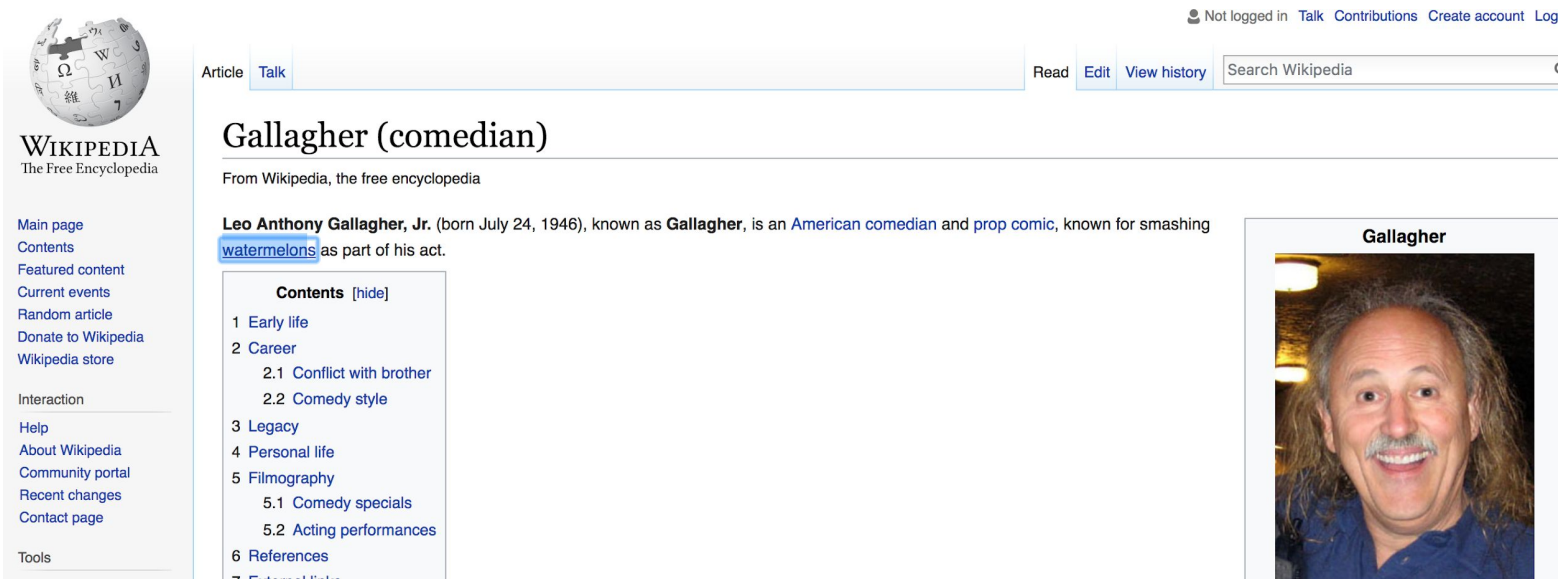
Using Web Search

Q: Do you see a fruit that Gallagher would likely smash with the Sledge-O-Matic?

How I would solve the question



How I would solve the question



Not logged in Talk Contributions Create account Log

Article Talk | Read Edit View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Gallagher (comedian)

From Wikipedia, the free encyclopedia

Leo Anthony Gallagher, Jr. (born July 24, 1946), known as **Gallagher**, is an American comedian and prop comic, known for smashing watermelons as part of his act.

Contents [hide]

1 Early life
2 Career
  2.1 Conflict with brother
  2.2 Comedy style
3 Legacy
4 Personal life
5 Filmography
  5.1 Comedy specials
  5.2 Acting performances
6 References

Gallagher

Q: Do you see a fruit that Gallagher would likely smash with the Sledge-O-Matic?

Q: Do you see a fruit that Gallagher would likely smash with the Sledge-O-Matic?

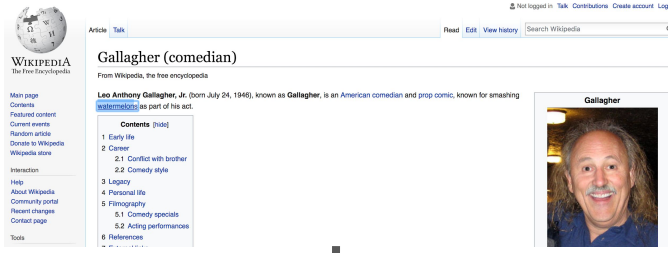
Q: Do you see a fruit that Gallagher would likely smash with the Sledge-O-Matic?

A: Yes

Use Query to get info

qy

Page



Article Encoder

h_{info}