Computer Vision
Introduction

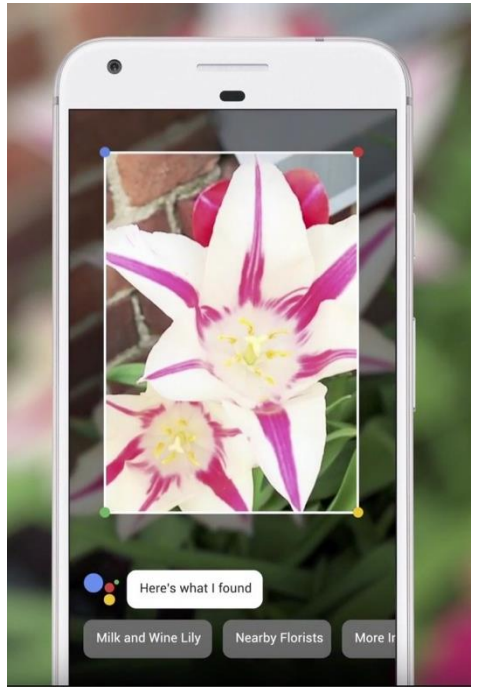
Dr. A U G Sankararao

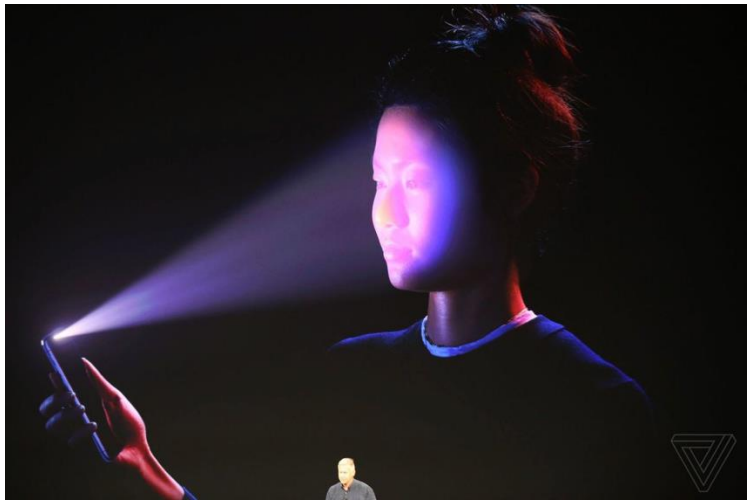
Indian Institute of Information Technology
Sri City, Chittoor

Today’s Class

- Who am I ?
- What is Computer Vision ?
- Real world relavance
- Course Content and Logistics
- Questions

Have you ever used these ?





What is Computer Vision ?

- Make computers understand images and videos
- A field that seeks to automate and endow a computing framework with the ability to interpret images the way humans do.
- A sub-topic of Artificial Intelligence.

Other Definitions

- “the construction of explicit, meaningful descriptions of physical objects from images” (Ballard & Brown, 1982)
- “computing properties of the 3D world from one or more digital images” (Trucco & Verri, 1998)
- “to make useful decisions about real physical objects and scenes based on sensed images” (Sockman & Shapiro, 2001)

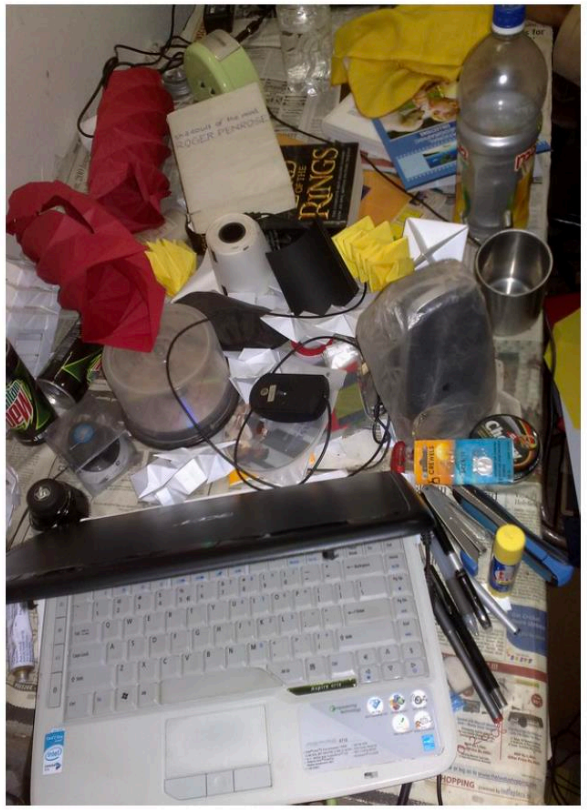
What is Computer Vision ?

- Make computers understand images and videos
- A field that seeks to automate and endow a computing framework with the ability to interpret images the way humans do.
- A sub-topic of Artificial Intelligence.

Other Definitions

- “the construction of explicit, meaningful descriptions of physical objects from images” (Ballard & Brown, 1982)
- “computing properties of the 3D world from one or more digital images” (Trucco & Verri, 1998)
- “to make useful decisions about real physical objects and scenes based on sensed images” (Sockman & Shapiro, 2001)

What is Computer Vision ?



Where is the gluestick? Find the book - what's its full title?

Credit: Bharath Kishore, Flickr CC License

What is wrong with this image?

Credit: Erik Johansson

Can a machine answer the above questions?

Slide adapted from DL for Vision by Prof. Vineeth N (NPTEL)

Computer Vision

- Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. (perception and interpretation)
- Computing properties of the 3D world from visual data (measurement)
- Algorithms to mine, search, and interact with visual data (search and organization)

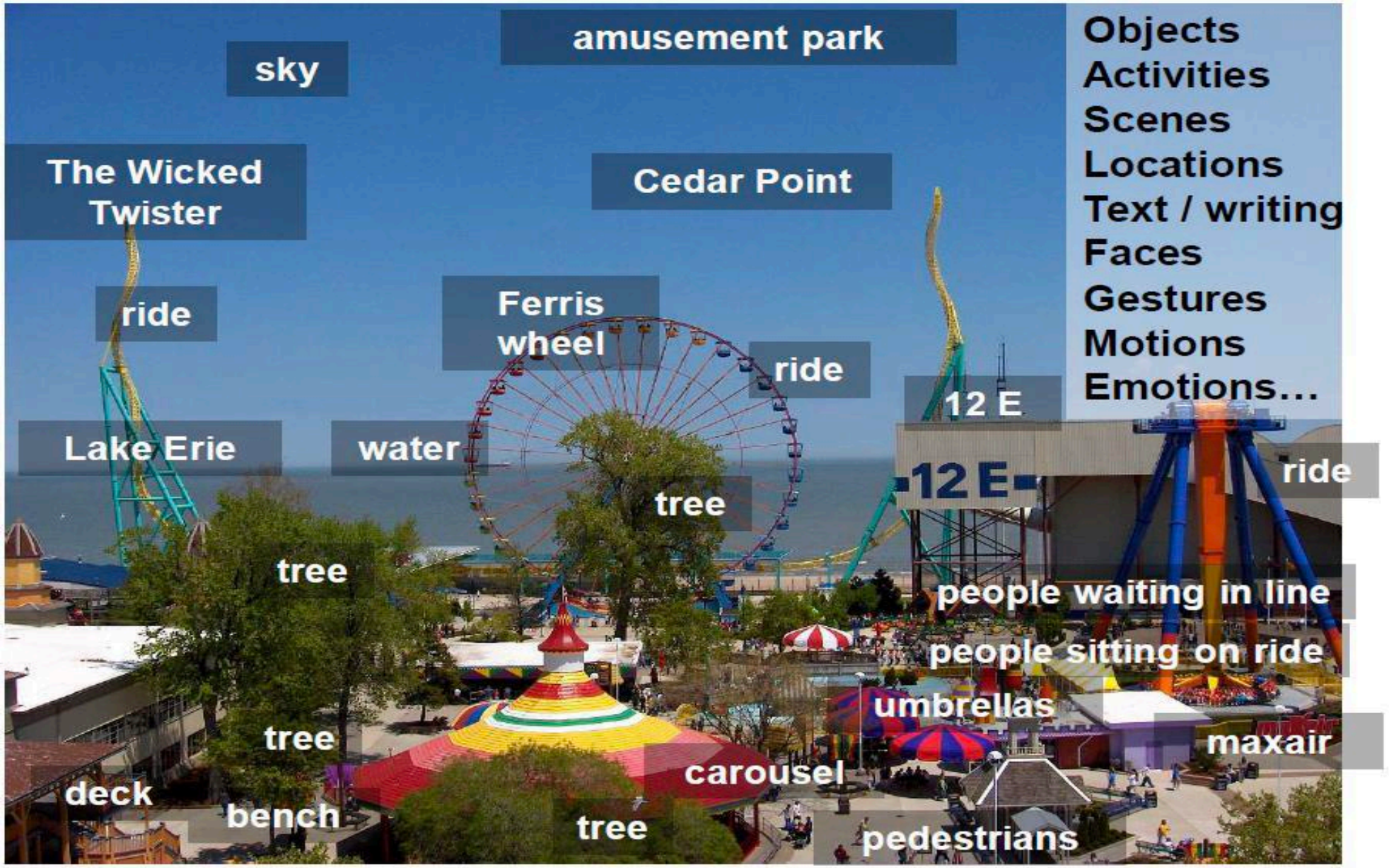
- Watch:

https://www.youtube.com/watch?v=2ioKHHYAQgI

https://www.youtube.com/watch?v=GbzqnEU1Z-Q

Vision for perception and interpretation

- Semantic information
- Metric 3D information



Vision for measurement

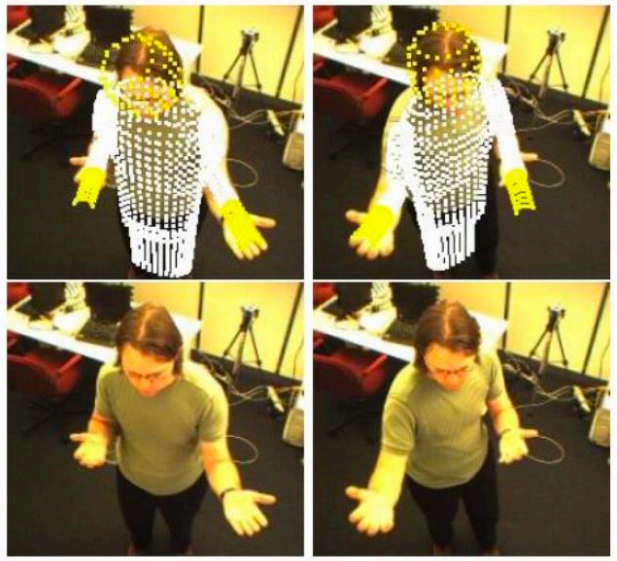
Real-time stereo

Wang et al.

Structure from motion

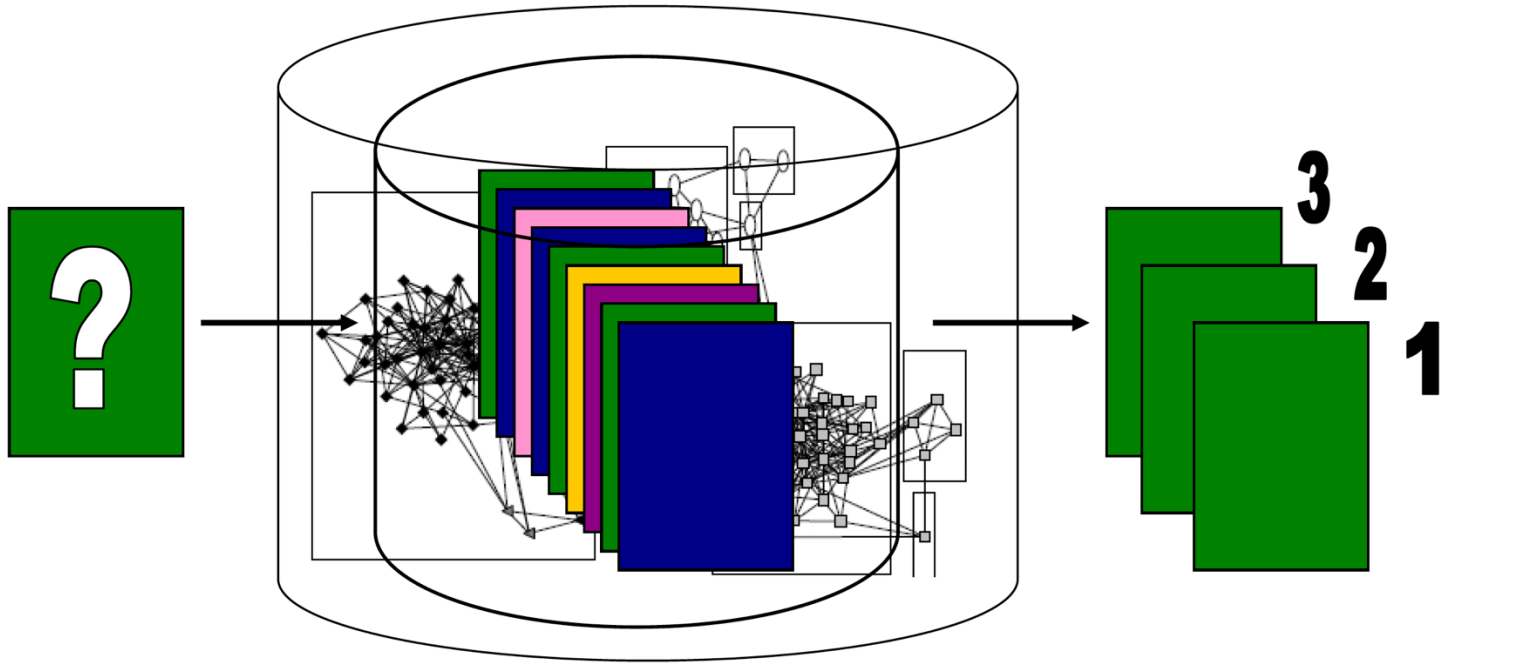
Snavely et al.

Tracking



Demirdjian et al.

Visual search organization



Goal of CV

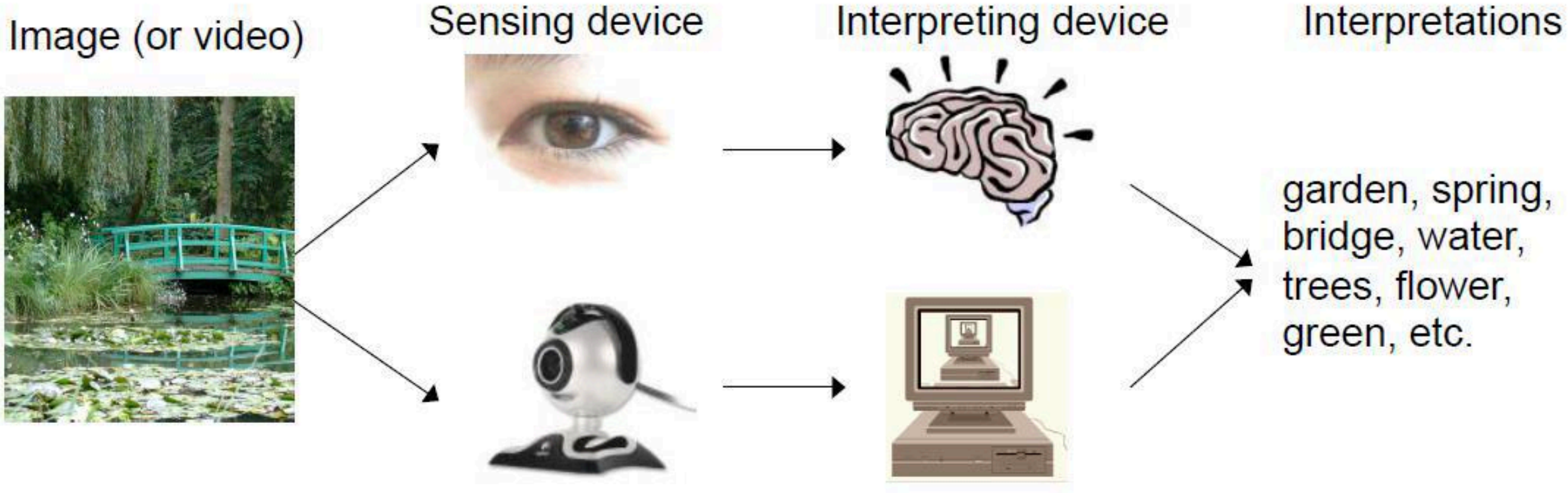
- To extract “meaning” from pixels

What we see

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

What a computer sees

Human vs Computer vision



Slide adapted from Fei-Fei Li

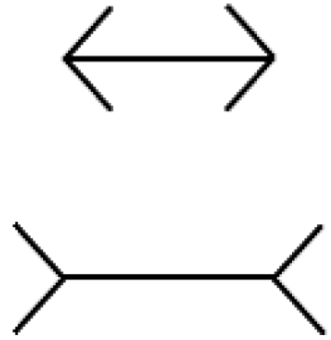
Can the computer match human perception?

- Yes and no (mainly no)
  - computers can be better at “easy” things
  - humans are much better at “hard” things
- But huge progress has been made
  - Especially in the last 10 years
  - What is considered “hard” keeps changing
- With enough data, computers can surpass humans

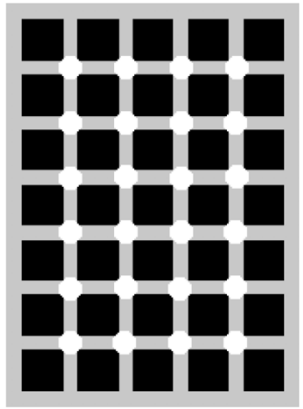
Is vision really hard ?

- Vision is an amazing feat of natural intelligence
  - Visual cortex occupies about 50% of Macaque brain. highly specialized to interpret visual stimuli: motion, colour, shape, depth, object recognition, etc.
  - One third of human brain is devoted to vision (more than anything else)

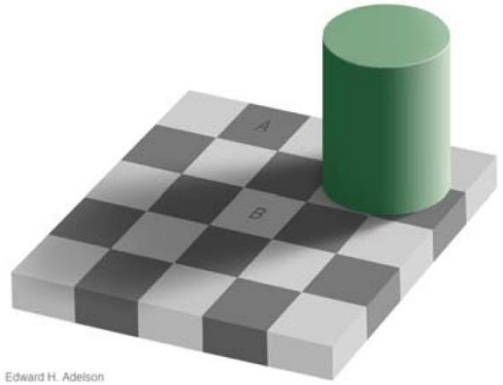
Why is it hard ?



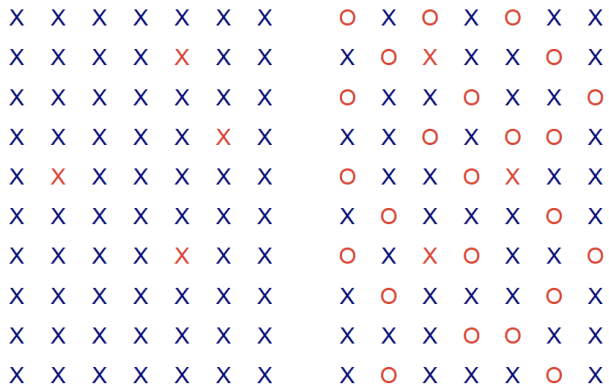
Müller-Lyer illusion: Which line is longer?



Variation of Hermann grid illusion: What do you see at the intersections?



Adelson's brightness constancy illusion: Which is brighter, A or B?

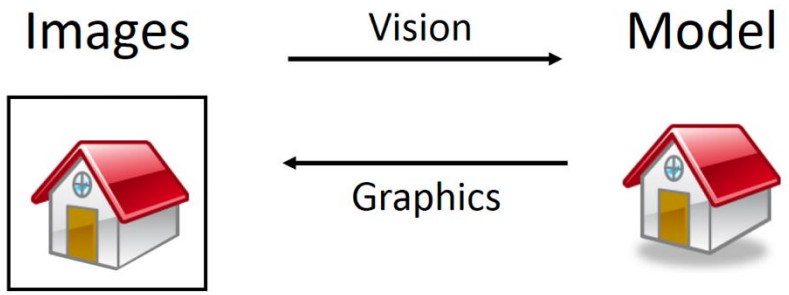


Count the red Xs in both figures, which is harder?

[1] Credit: Szeliski, Computer Vision: Algorithms and Applications, 2010

Why is it hard ?

- Many practical use cases are inverse model applications
  - No knowledge of how an image was taken or camera parameters - but need to model the real world in which picture/video was taken (shape, lighting, color, objects, interactions) => Need to model from incomplete/partial noisy information
  - Forward models are used in physics (radiometry, optics, and sensor design) and in computer graphics

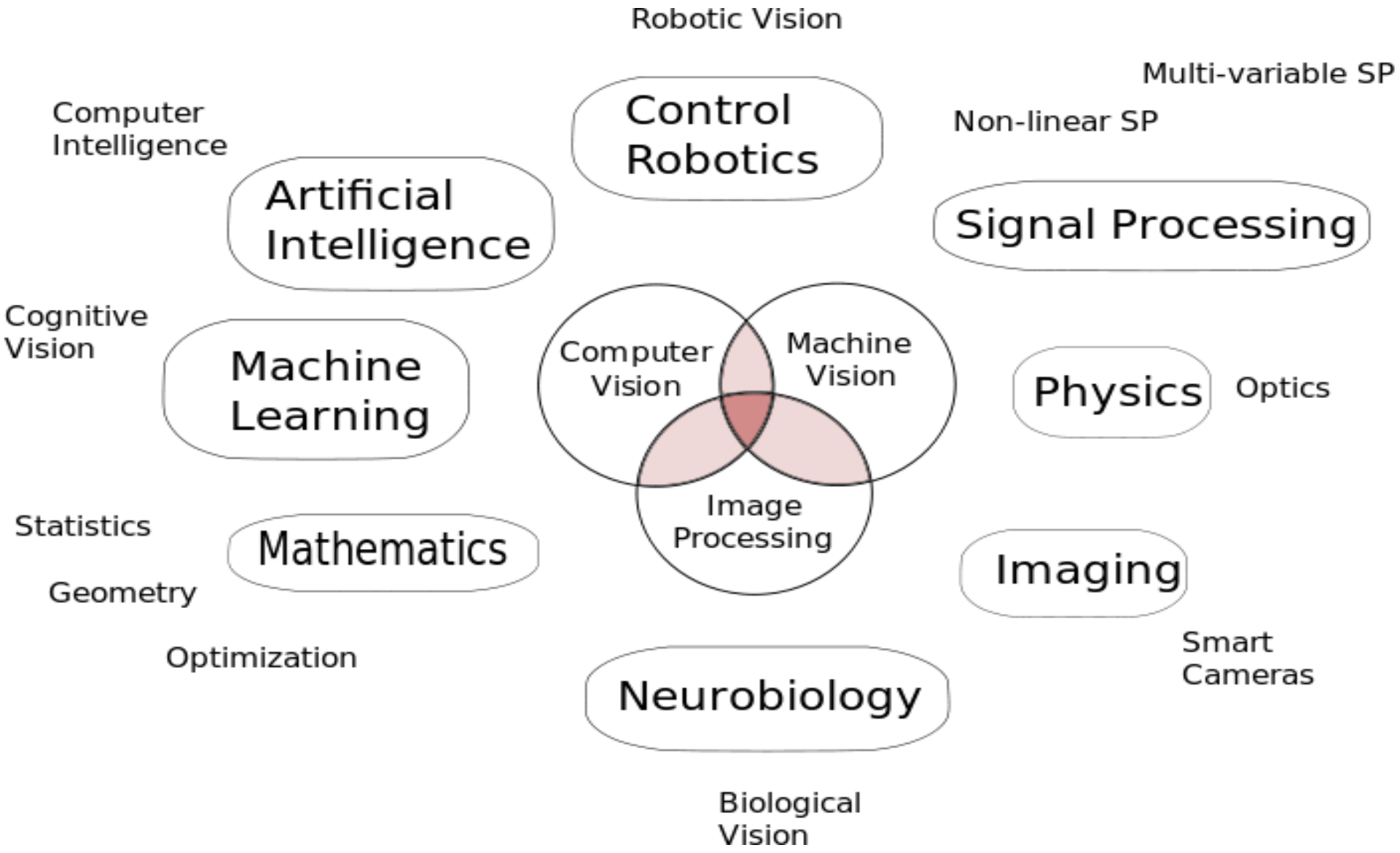


Inverse problems: analysis and synthesis.

Why is it hard ?

- High-dimensional data => heavy computational requirements
- No complete models of the human visual system exist
  - Existing models largely related to subsystems, not holistic
  - What is perceived, and what is cognized? When is an object important for a task, and when is the context important?

Perspectives of Study



Slide adapted from DL for Vision by Prof. Vineeth N (NPTEL)

Why to Study CV ?

- Millions of images being captured all the time

Google Image Search

Picasa™

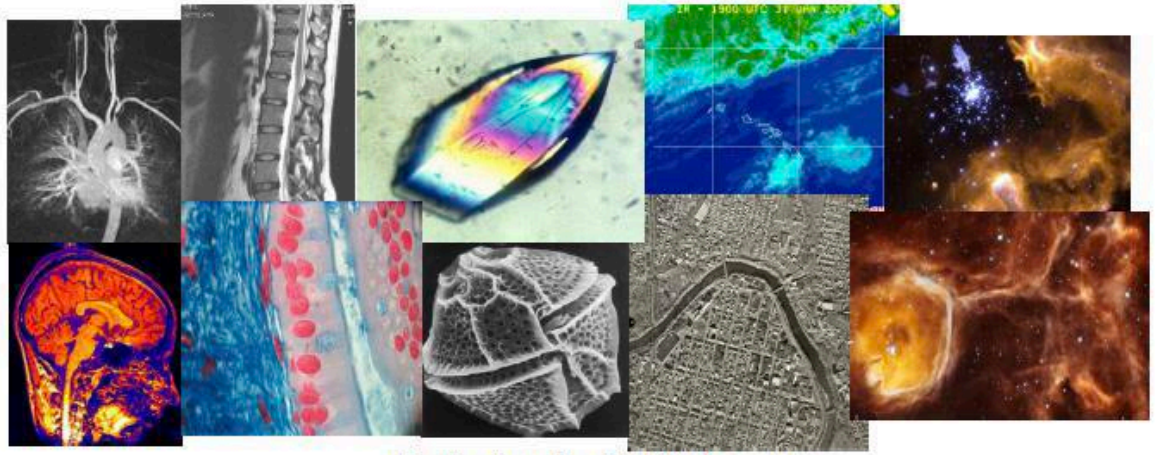
flickr™

webshots

picsearch™

YouTube Broadcast Yourself™

Surveillance and security



Medical and scientific images

Visual Data on Internet

- Flickr
  - –10+ billion photographs
  - –60 million images uploaded a month
- Facebook
  - –250 billion+
  - –300 million a day
- Instagram: 55 million a day
- YouTube: 100 hours uploaded every minute

Vision in Space

NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

Vision systems (JPL) used for several tasks

- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read "Computer Vision on Mars" by Matthies et al.

Applications of CV

Autonomous Vehicles

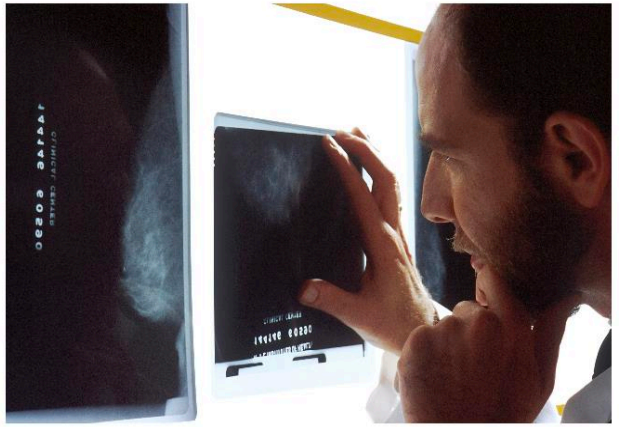
Credit: smoothgrover22, Flickr CC License

Surveillance

Credit: Yeong Nam, Flickr CC License

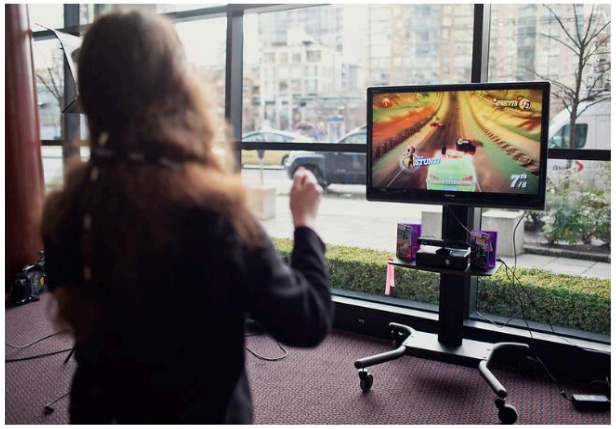
Factory Automation

Credit: KUKA Roboter GmbH, Bachmann



Medical Imaging

Credit: National Cancer Institute



Human-Computer Interaction

Credit: Vancouver Film School

Visual Effects

Credit: AntMan3001, Flickr CC License

Applications of CV

- **Smartphones:** QR codes, computational photography (Android Lens Blur, iPhone Portrait Mode), panorama construction (Google Photo Spheres), Night Sight (Pixel), iPhone Pro 3D scanning (LiDAR), body workout form detection, face filters, FaceID (iPhone)
- **Smartwatches:** Heart rate detection, proximity detection
- **Security:** Fingerprint/iris/face scanning (offices, airports), CCTV monitoring
- **Laptops/Desktops:** Biometrics auto-login (face recognition, 3D)
- **Web:** Image search, Google photos (face recognition, object recognition, scene recognition, geolocalization from vision), Facebook (image captioning), Google maps aerial imaging (image stitching), YouTube (content categorization), Photoshop, PowerPoint (captioning, design suggestions)
- **Virtual Worlds:** VR/AR head tracking (Meta Quest, Apple Vision Pro), simultaneous localization and mapping, person tracking (Kinect), gesture recognition, virtual try-on, digital humans
- **Telepresence:** Virtual backgrounds (Zoom, Google Meet), webcam person/face following
- **Media:** Visual effects for film/TV, virtual sports replay, automatic camera control, semantics-based auto edits
- **Transportation:** Assisted driving (cruise control, self-driving), face tracking/iris dilation for safety
- **Supermarkets:** Cashier-less checkout, theft detection (Walmart), fruits/vegetables sorting, packaging + manufacture
- **Medical imaging:** CAT / MRI reconstruction, assisted diagnosis, surgery planning, automatic pathology, connectomics
- **Space Exploration:** Mar rovers, space telescopes (Hubble, James Webb)
- **Industry:** Vision-based robotics (human+robot spaces in Amazon warehouses), online shopping (Amazon, Walmart), machine-assisted tools (routers, jigs), OCR (USPS), ANPR (number plates for tolls), drones
- **Creative:** Photoshop, vision-language models for image and video generation (Dall-E, SORA), automatic video editors

Applications of CV

- **Retail and Retail Security** (Amazon Go, Virtual Try-on, StopLift)
- **Healthcare** (Blood Loss Detector, DermLens)
- **Agriculture** (SlantRange, Cainthus - Livestock facial recognition)
- **Banking and Finance** (Mobile Deposit, Insurance Risk Profiling)
- **Remote Sensing** (Land Use Understanding, Forestry Modeling)
- **Structural Health Monitoring** (Oilwell Inspection, Drone-based Bridge Inspection and 3D Reconstruction)
- **Document Understanding** (Optical Character Recognition, Robotic Process Automation)
- **Tele- and Social Media** (Image Understanding, Brand Exposure Analytics)
- **Augmented Reality** (TechSee Visual Support, Warehouse and Enterprise Management)

Reconstruction: 3D from photo collections

Colosseum, Rome, Italy

San Marco Square, Venice, Italy

Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. Seitz, The Visual Turing Test for Scene Reconstruction, 3DV 2013

YouTube Video

Slide adapted from **SVETLANA LAZEBNIK**

Reconstruction: 4D from depth cameras

Figure 1: Real-time reconstructions of a moving scene with DynamicFusion; both the person and the camera are moving. The initially noisy and incomplete model is progressively denoised and completed over time (left to right).

R. Newcombe, D. Fox, and S. Seitz, DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time, CVPR 2015

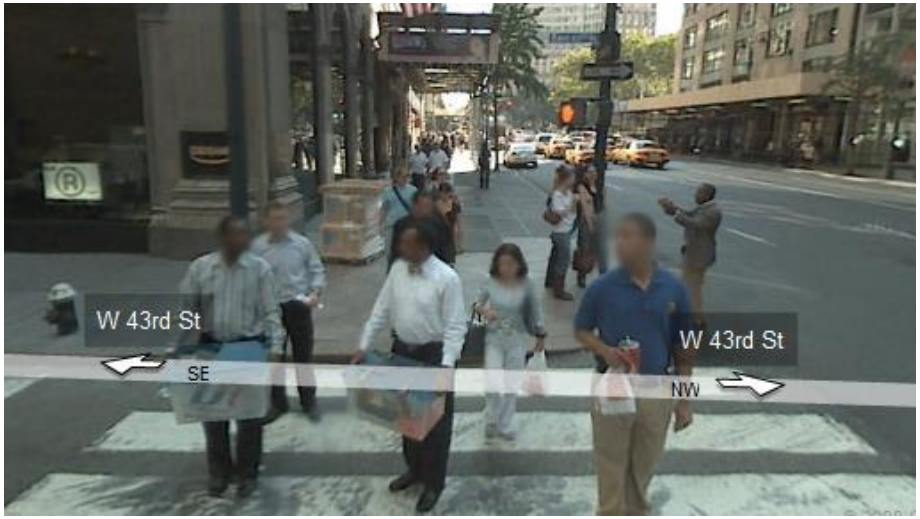
YouTube Video

Slide adapted from **SVETLANA LAZEBNIK**

Recognition: “Simple” patterns

Slide adapted from **SVETLANA LAZEBNIK**

Recognition: Faces



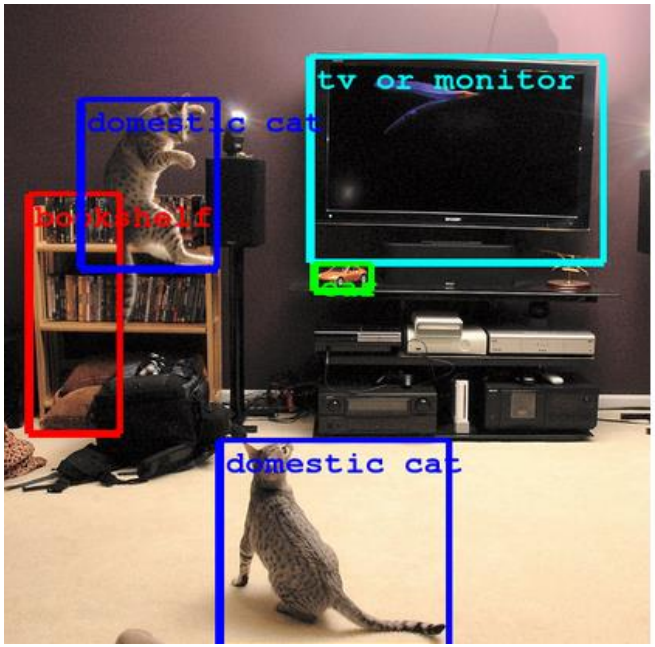
Slide adapted from **SVETLANA LAZEBNIK**

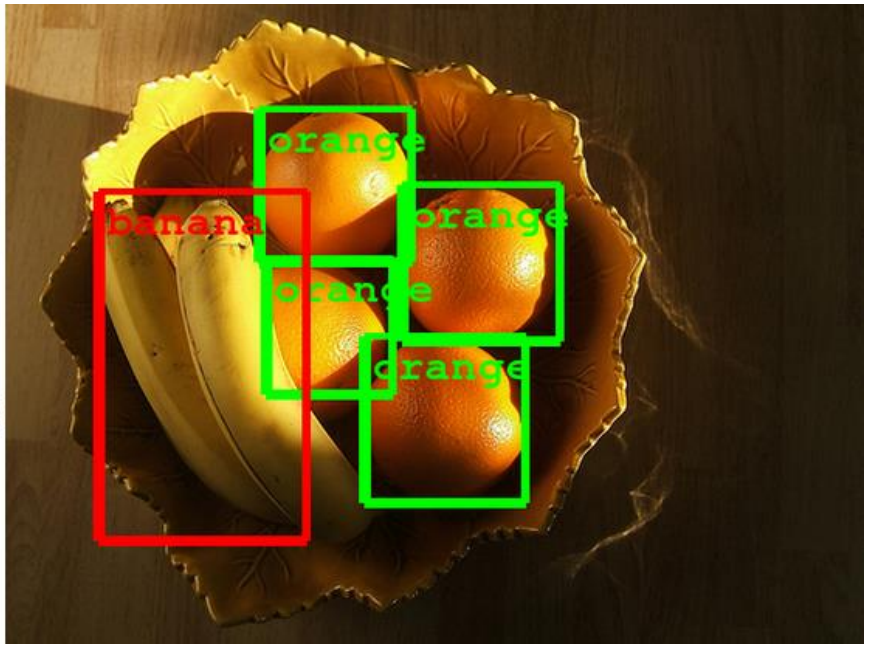
Slide adapted from **SVETLANA LAZEBNIK**

Concerns about face recognition

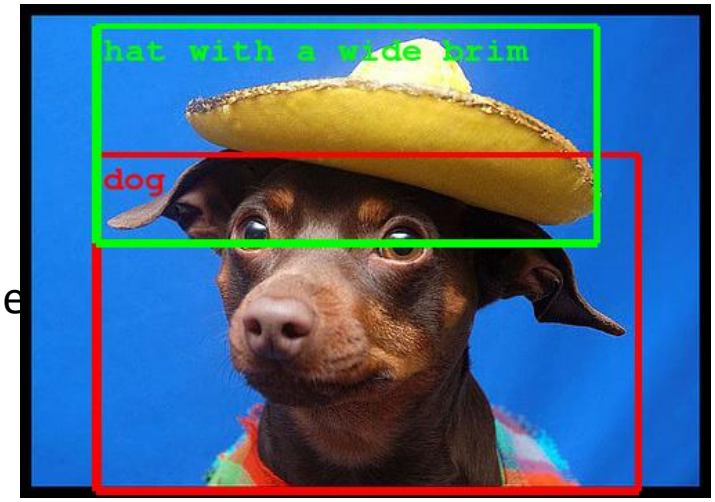
[Beijing bets on facial recognition in a big drive for total surveillance](#) – Washington Post, 1/8/2018

Recognition: General categories





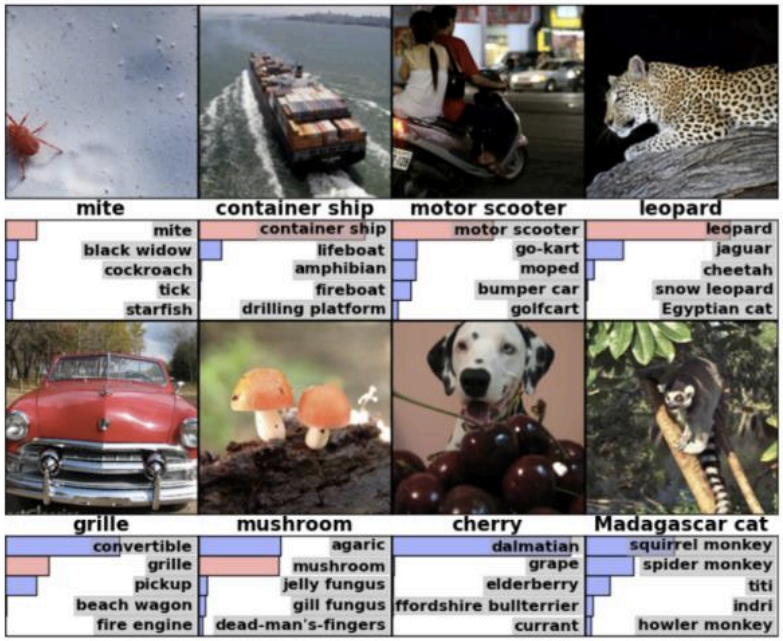
- Computer Eyesight Gets a Lot More Accurate, NY Times Bits blog, August 18, 2014
- Building A Deeper Understanding of Images, Google Research Blog, September 5, 2014

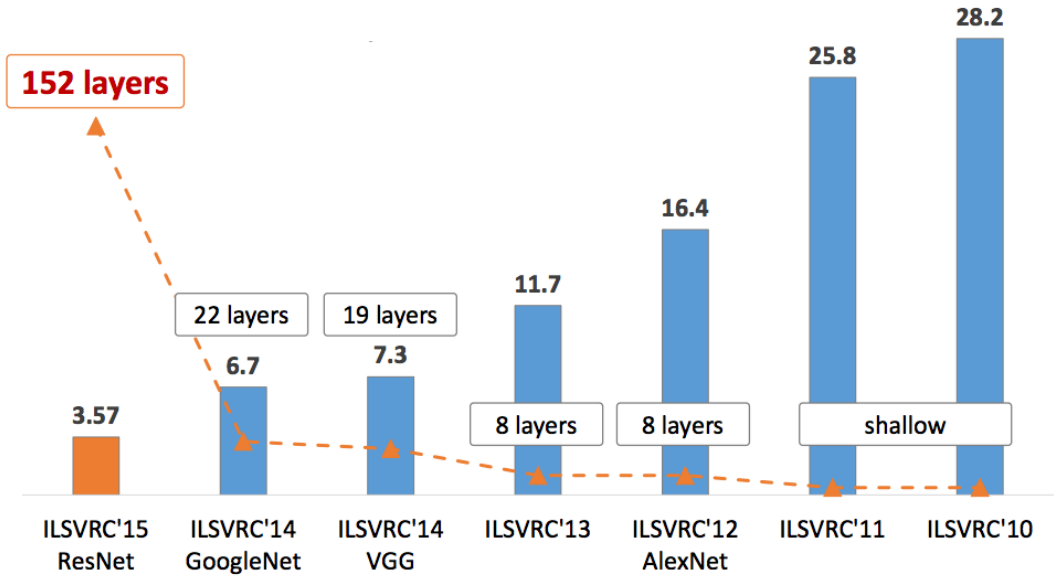


Slide adapted from **SVETLANA LAZEBNIK**

Recognition: General categories

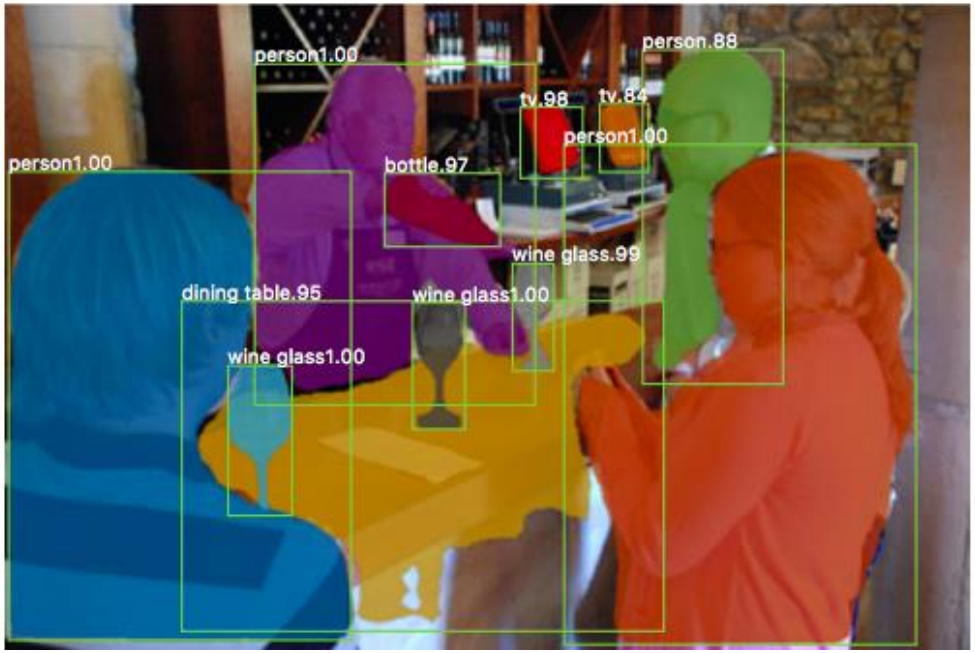
- ImageNet challenge





Slide adapted from **SVETLANA LAZEBNIK**

Object detection, instance segmentation



K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask R-CNN,
ICCV 2017 (Best Paper Award)

Slide adapted from **SVETLANA LAZEBNIK**

Image generation

- Faces: 1024x1024 resolution, CelebA-HQ dataset

T. Karras, T. Aila, S. Laine, and J. Lehtinen, Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018

Follow-up work

Slide adapted from **SVETLANA LAZEBNIK**

Image generation

- BigGAN: 512 x 512 resolution, ImageNet

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, arXiv 2018

Slide adapted from **SVETLANA LAZEBNIK**

Image generation

- BigGAN: 512 x 512 resolution, ImageNet

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, arXiv 2018

Slide adapted from **SVETLANA LAZEBNIK**

Image generation

- BigGAN: 512 x 512 resolution, ImageNet

Easy classes

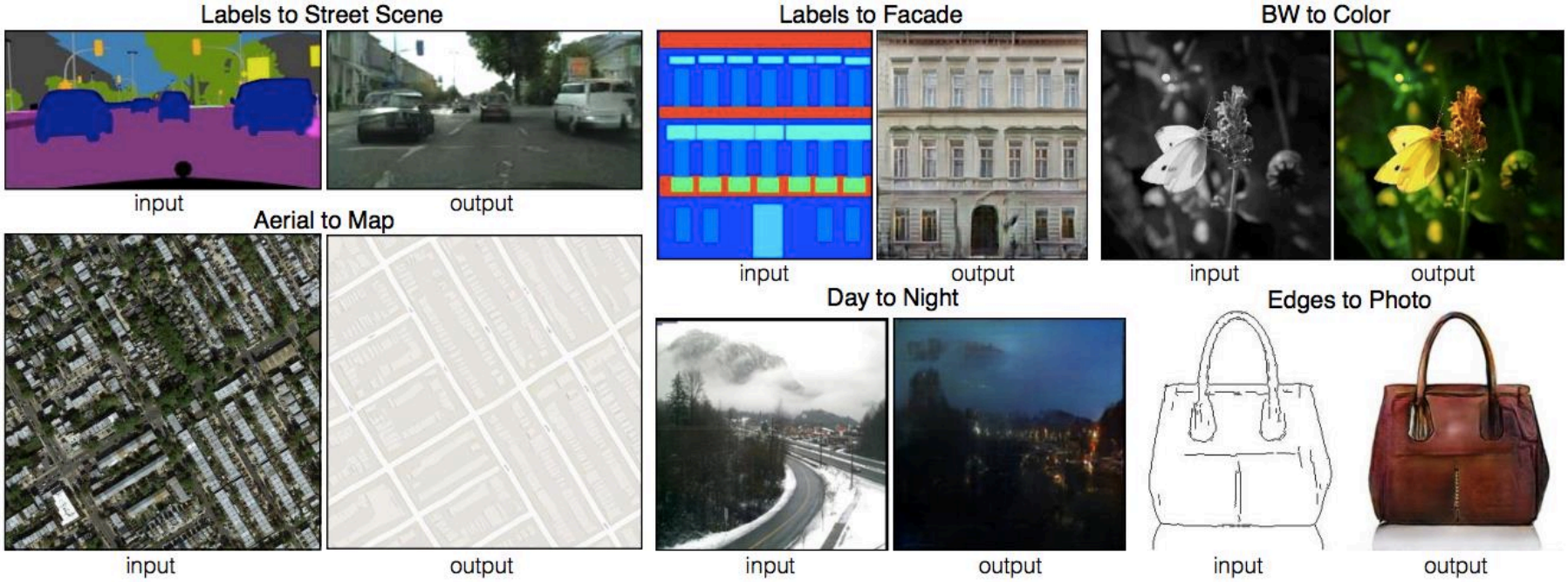
Difficult classes

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, arXiv 2018

Slide adapted from **SVETLANA LAZEBNIK**

Image generation

- Image-to-image translation

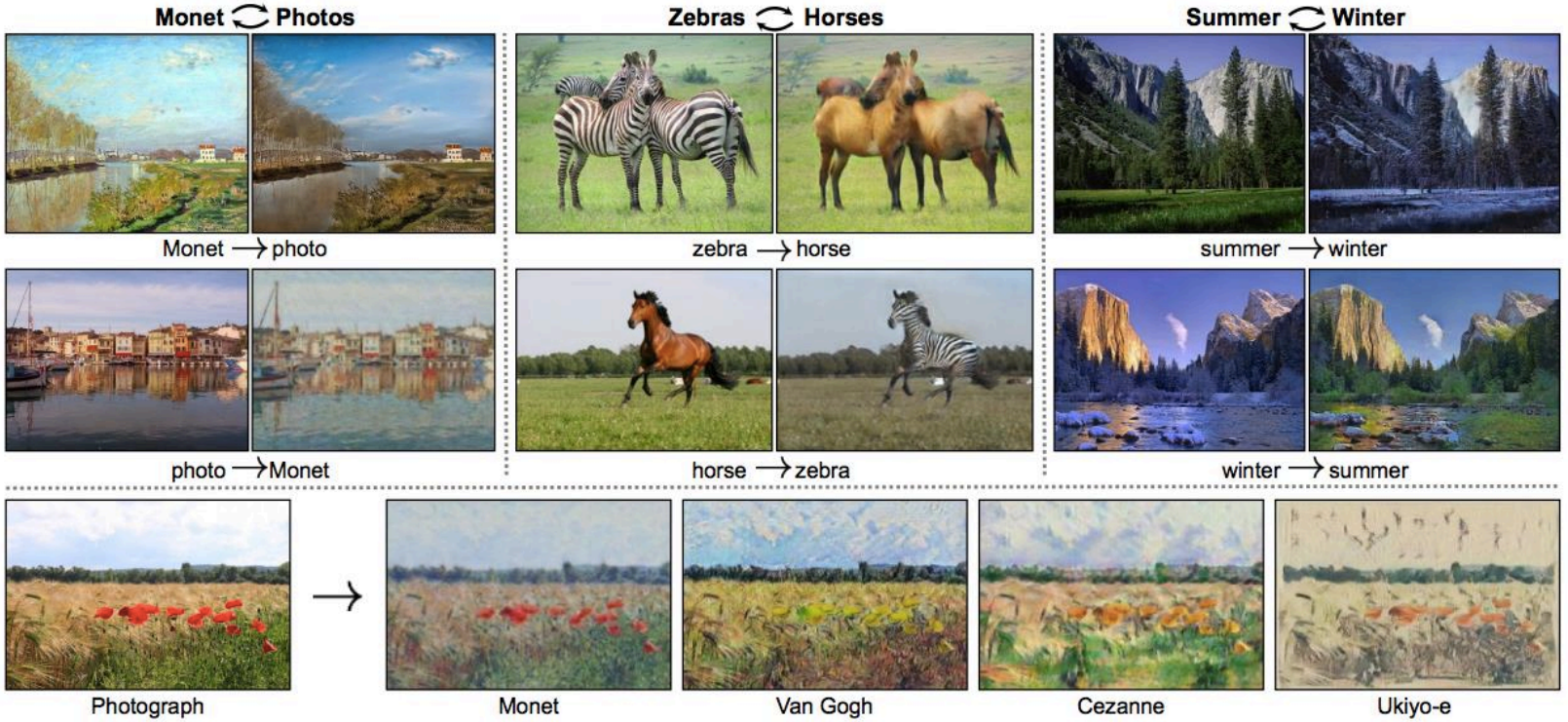


P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017

Slide adapted from **SVETLANA LAZEBNIK**

Image generation

- Unpaired image-to-image translation



J.-Y. Zhu, T. Park, P. Isola, A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, ICCV 2017

Slide adapted from **SVETLANA LAZEBNIK**

Unsupervised image-to-image translation

Figure 4: Dog breed translation results.

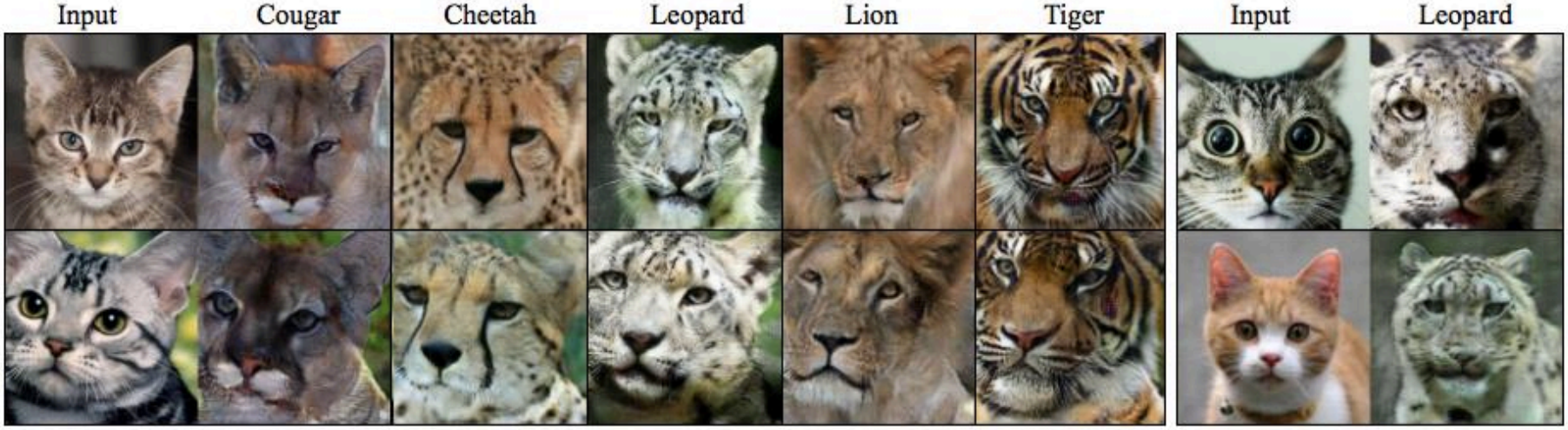
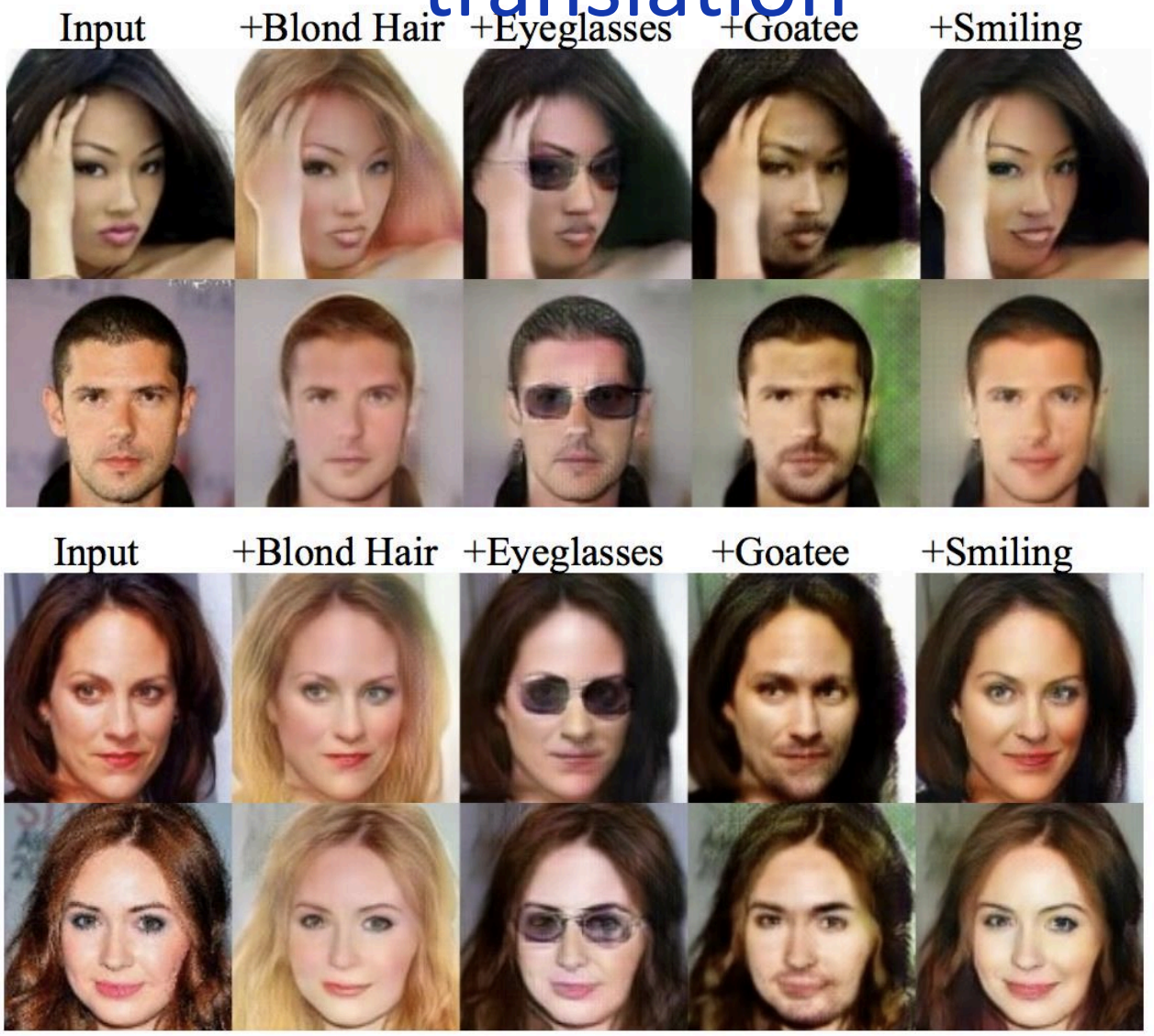


Figure 5: Cat species translation results.

M.-Y. Liu, T. Breuel, and J. Kautz, Unsupervised Image-to-Image Translation Networks, NIPS 2017

Unsupervised image-to-image translation



M.-Y. Liu, T. Breuel, and J. Kautz, Unsupervised Image-to-Image Translation Networks, NIPS 2017