CS 6476-A: Computer Vision

Instructor: James Hays

TAs: **Mengyu Yang, Sriman Goel** (head TAs), Borun Song, Eric Zhang, Esther Shen, Kristen Pereira, Nathaniel Koehler, Qihang Hu, Sirish Gambira, Yiming Chen, Zhenyu Wu

Image by kirkh.deviantart.com

Today’s Class

- Who am I?
- What is Computer Vision?
- Specifics of this course
- Geometry of Image Formation
- Questions

A bit about me

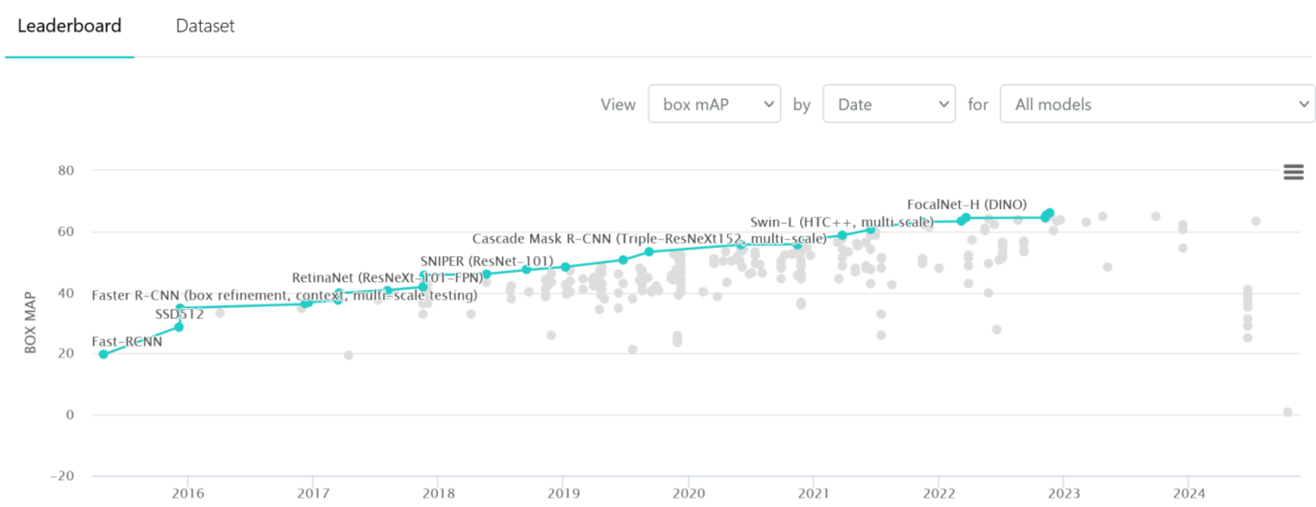
What type of stuff do I work on?

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin[1], Michael Maire[2], Serge Belongie[1], James Hays[3], Pietro Perona[2], Deva Ramanan[4], Piotr Dollár[5], and C. Lawrence Zitnick[5]

(a) Image classification (b) Object localization (c) Semantic segmentation (d) This work

Object Detection on COCO test-dev



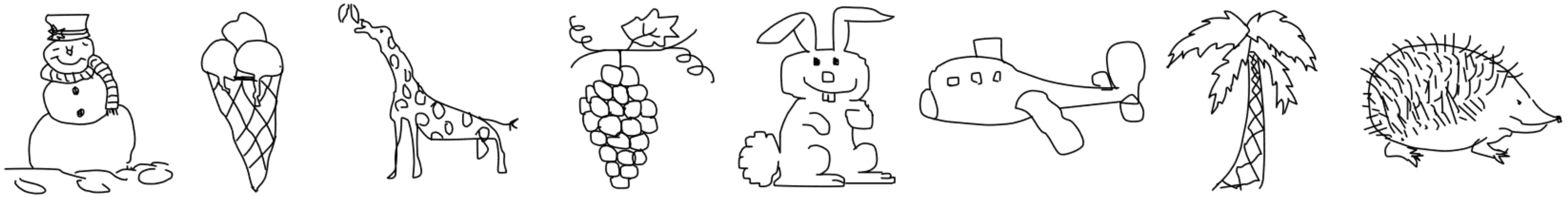
Winner of ECCV 2024 Koenderink Prize

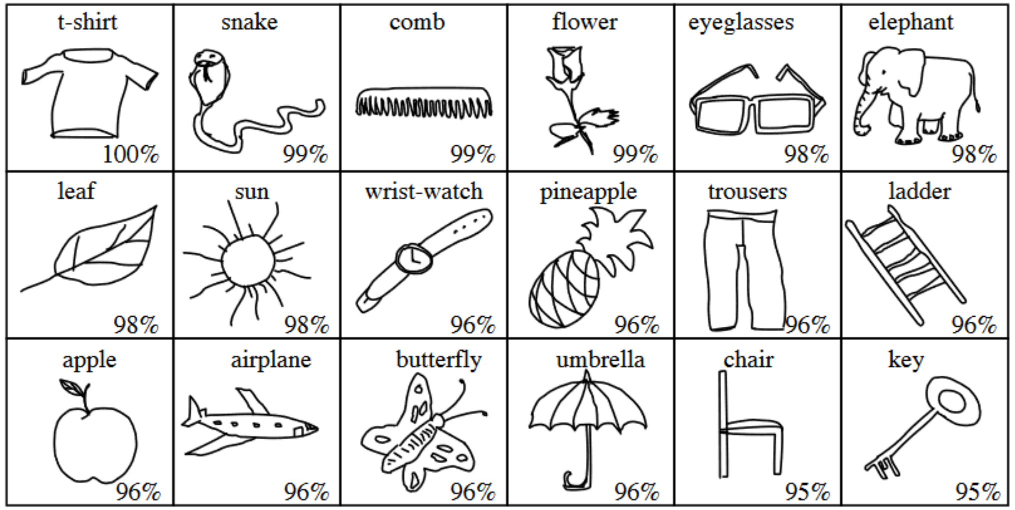
How Do Humans Sketch Objects?

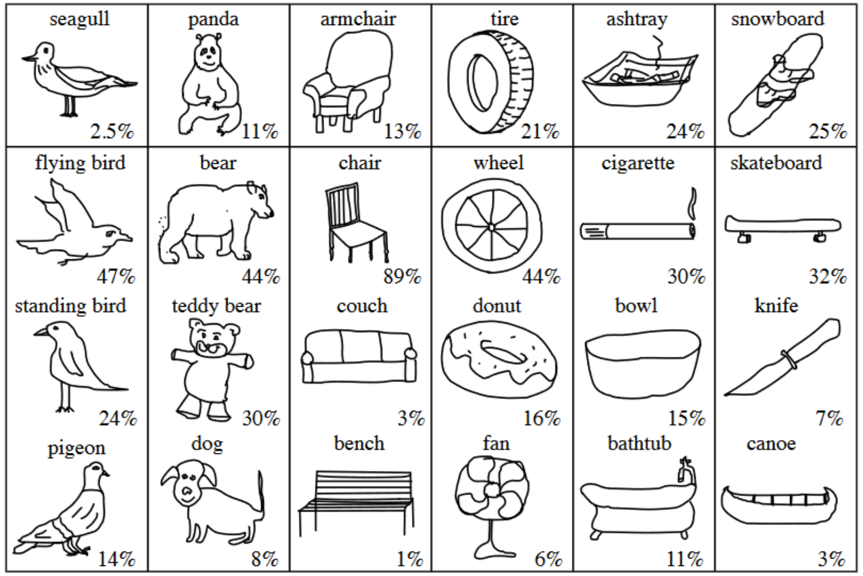
Mathias Eitz*
TU Berlin

James Hays†
Brown University

Marc Alexa ‡
TU Berlin





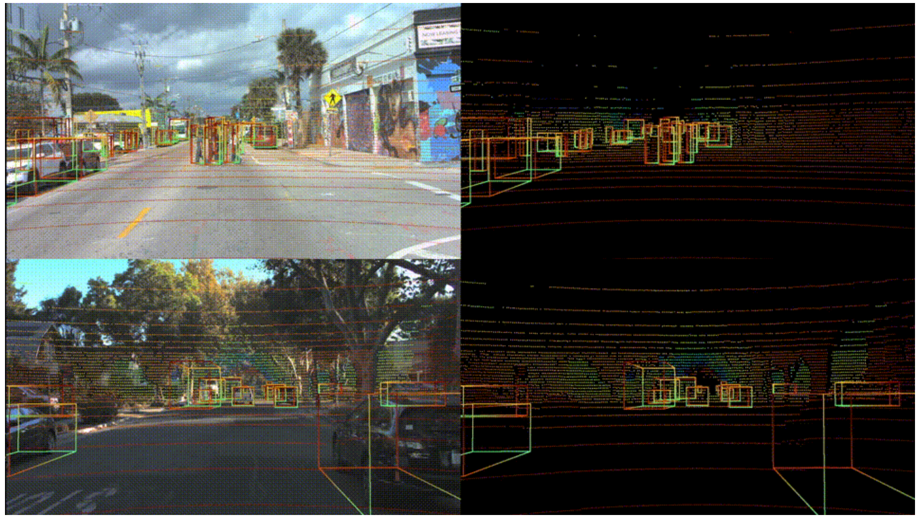


Siggraph 2012. **Won Siggraph Test of Time Award 2024.**

Personalized Residuals for Concept-Driven Text-to-Image Generation. Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, Tobias Hinz. CVPR 2024

Argoverse 2 (AV2) : Four Datasets

Sensor



Lidar

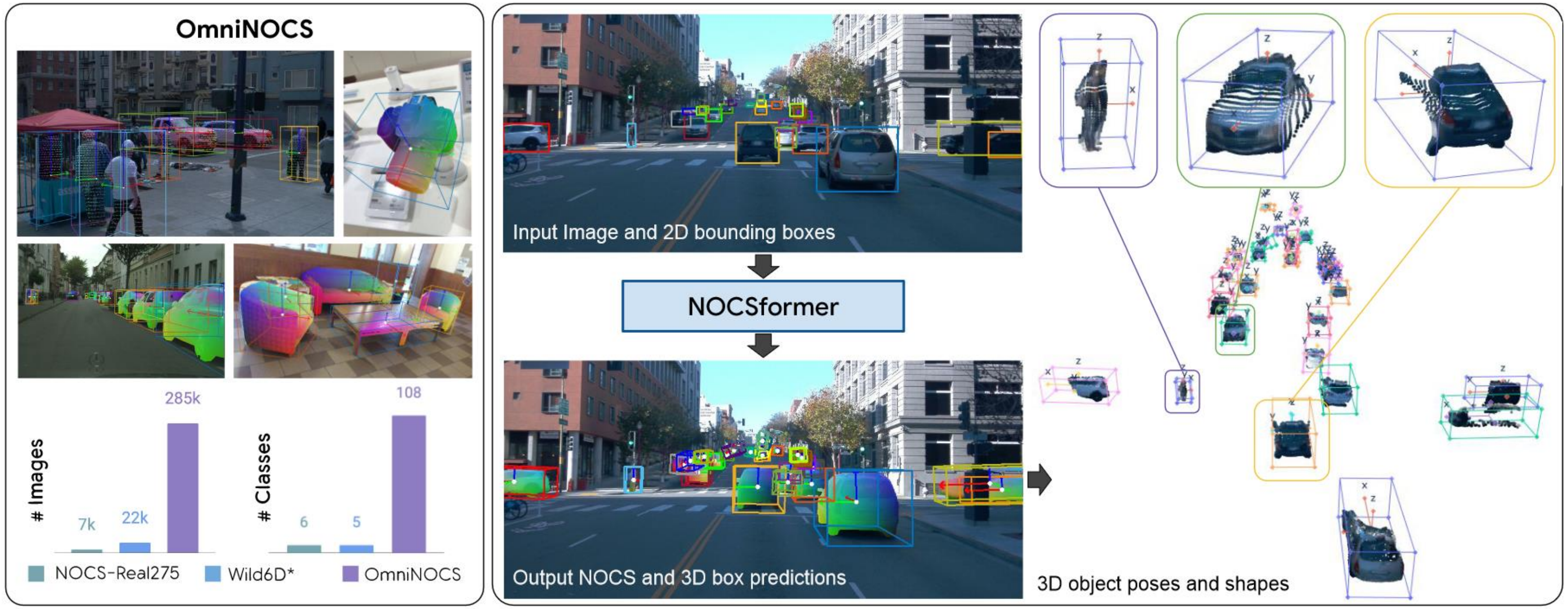
Motion Forecasting

Map Change

OmniNOCS: A unified NOCS dataset and model for 3D lifting of 2D objects

Akshay Krishnan[1,2], Abhijit Kundu[1], Kevis-Kokitsi Maninis[1], James Hays[2], and Matthew Brown[1]

[1] Google Research[†]
[2] Georgia Institute of Technology



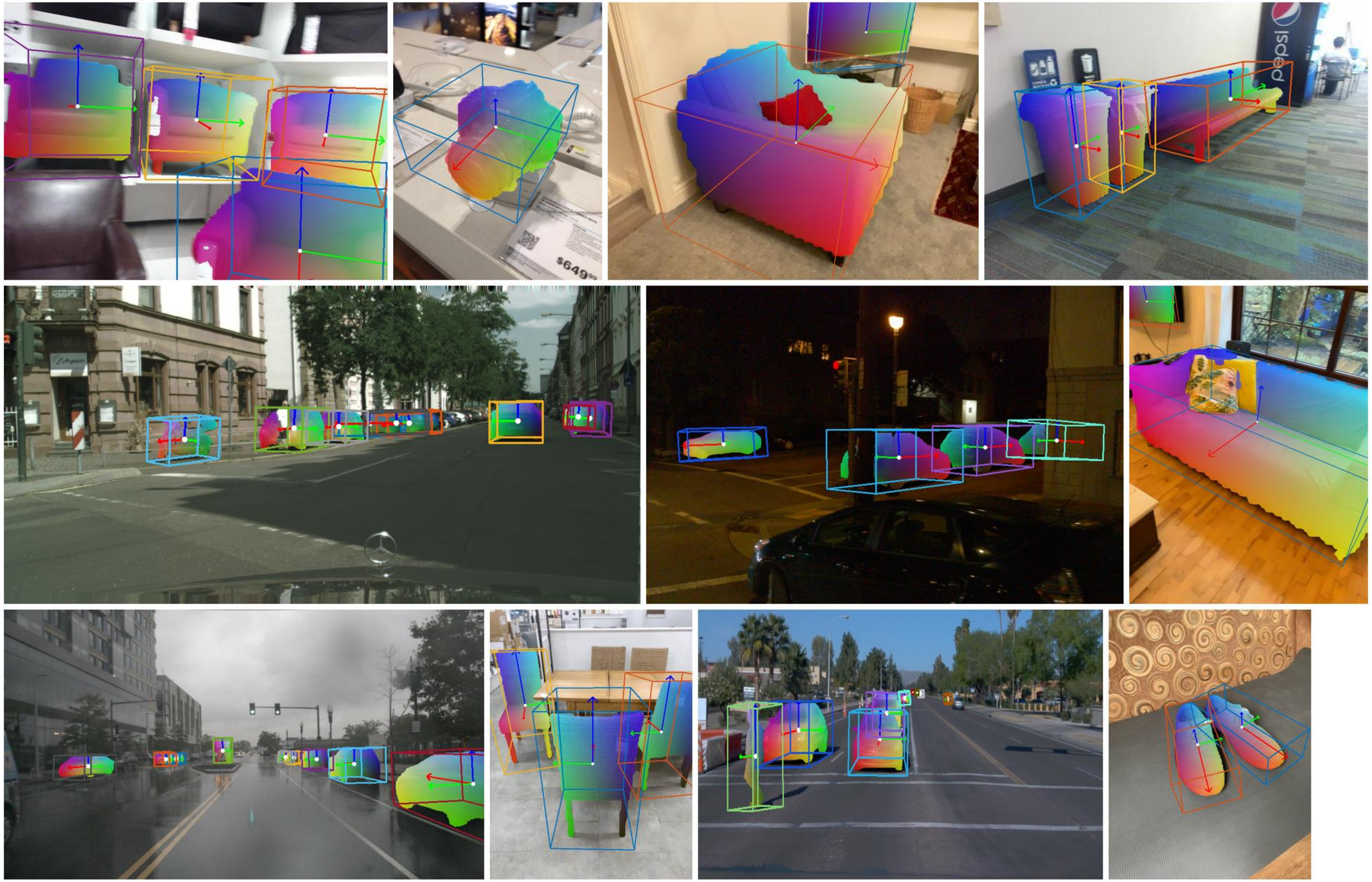
ECCV 2024 Oral. https://omninocs.github.io/

NOCS predictions on COCO objects

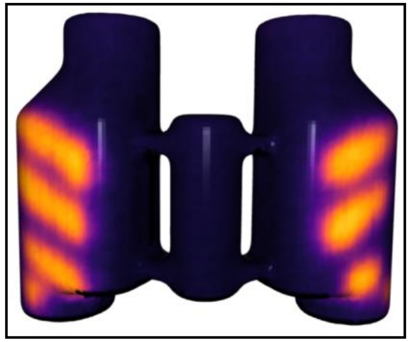
NOCSformer can generalize to in-the-wild objects in COCO images when trained on OmniNOCS.

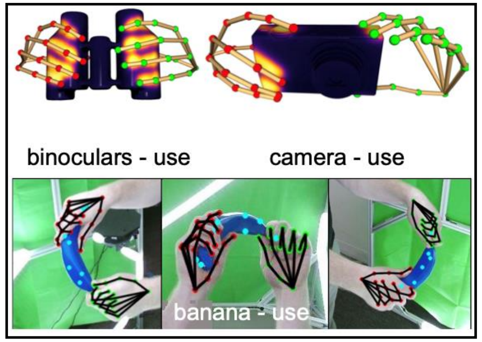
NOCS predictions across OmniNOCS

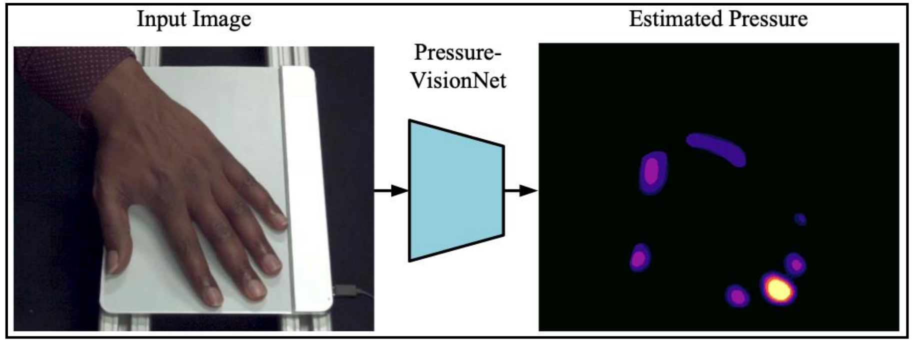
NOCSformer generalizes to the wide range of object classes and domains in OmniNOCS, including indoor and outdoor scenes, as well as object-centric images.



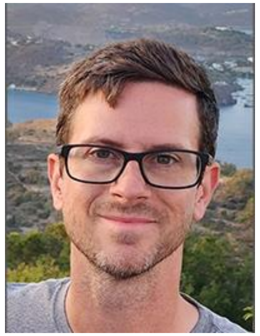
Creative Sensing for People and Robots







Presented by



James Hays

Georgia Institute of Technology

Overland AI

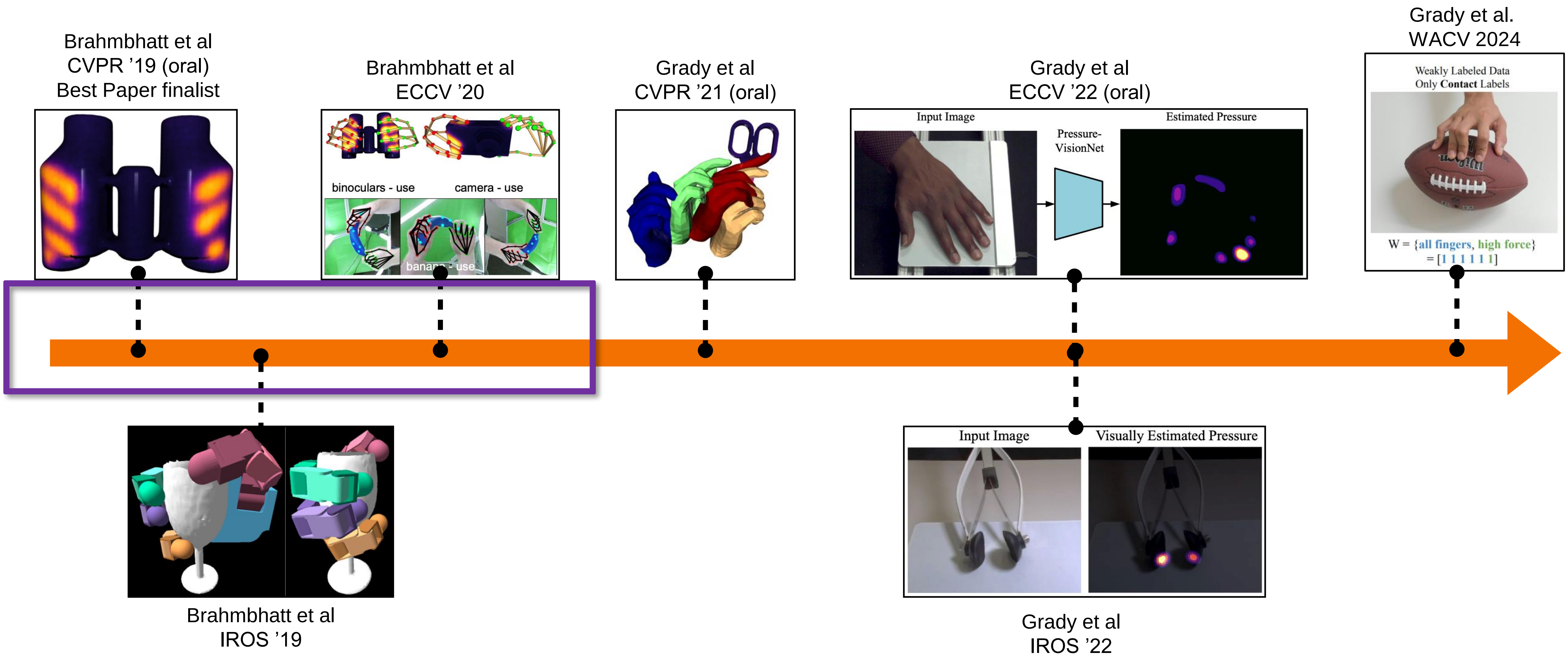
Samarth Brahmbhatt

Patrick Grady

Mengyu Yang

Charles C. Kemp

And collaborators Cusuh Ham, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Chengde Wan, Ankur Handa, Dieter Fox, Jeremy Collins

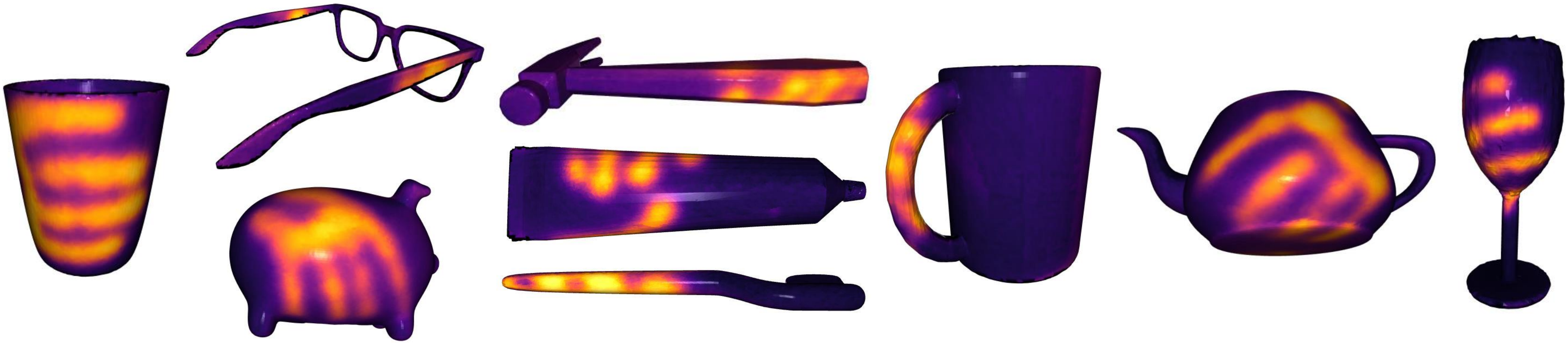


Brahmbhatt et al
CVPR ’19 (oral)
Best Paper finalist
Brahmbhatt et al
ECCV ’20
binoculars - use
camera - use
Grady et al
CVPR ’21 (oral)
Grady et al
ECCV ’22 (oral)
Input Image
Pressure-
VisionNet
Estimated Pressure
Grady et al.
WACV 2024
Weakly Labeled Data
Only Contact Labels
W = {all fingers, high force}
= [1 1 1 1 1 1]
Brahmbhatt et al
IROS ’19
Input Image
Visually Estimated Pressure
Grady et al
IROS ’22

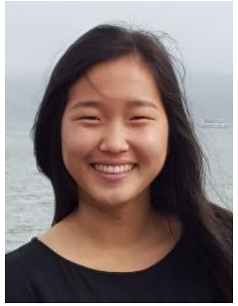
Why is observing contact difficult?

Occlusion

ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging




Samarth Brahmbhatt


Cusuh Ham


Charles C. Kemp


James Hays

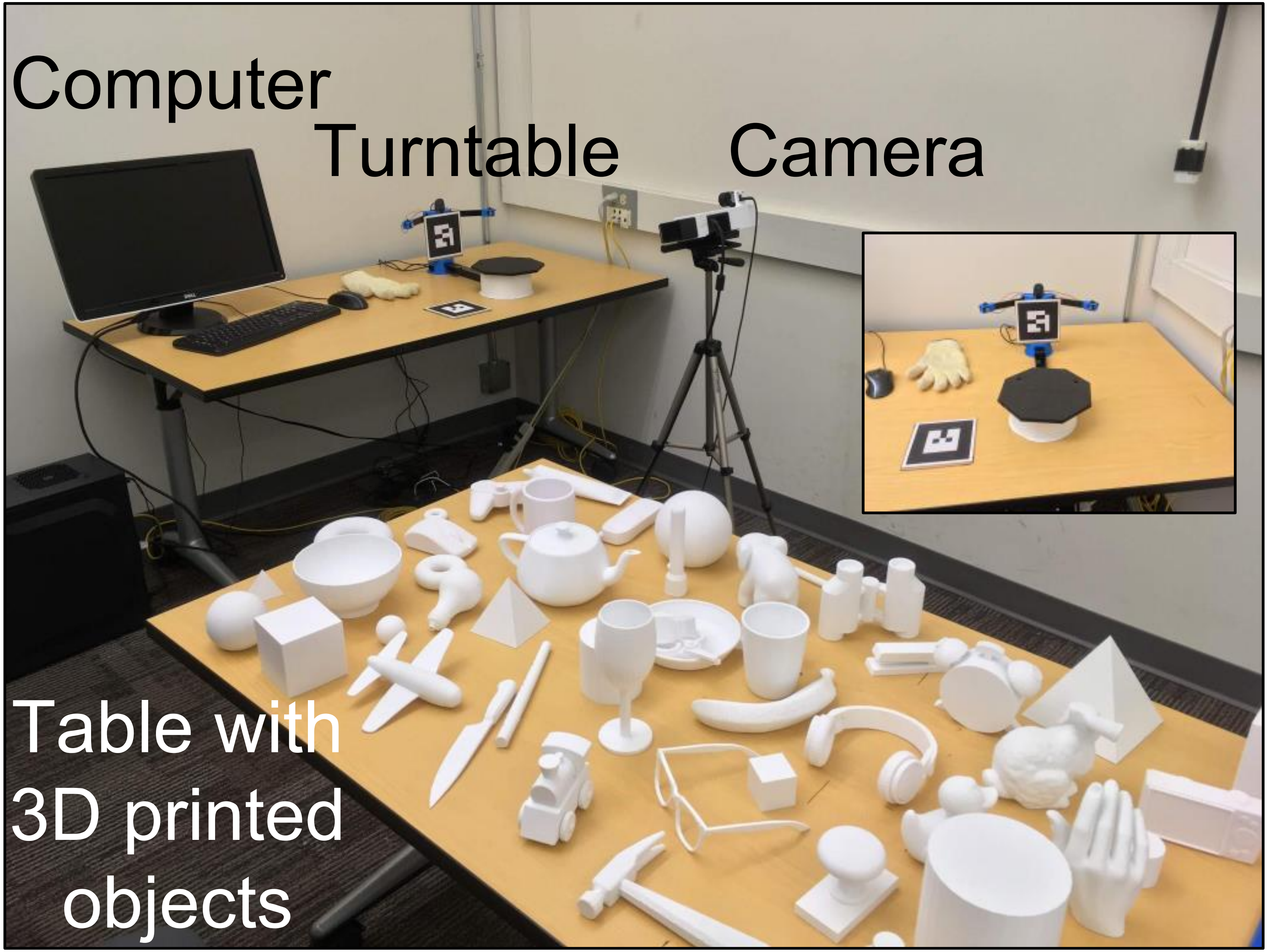
Georgia Institute of Technology

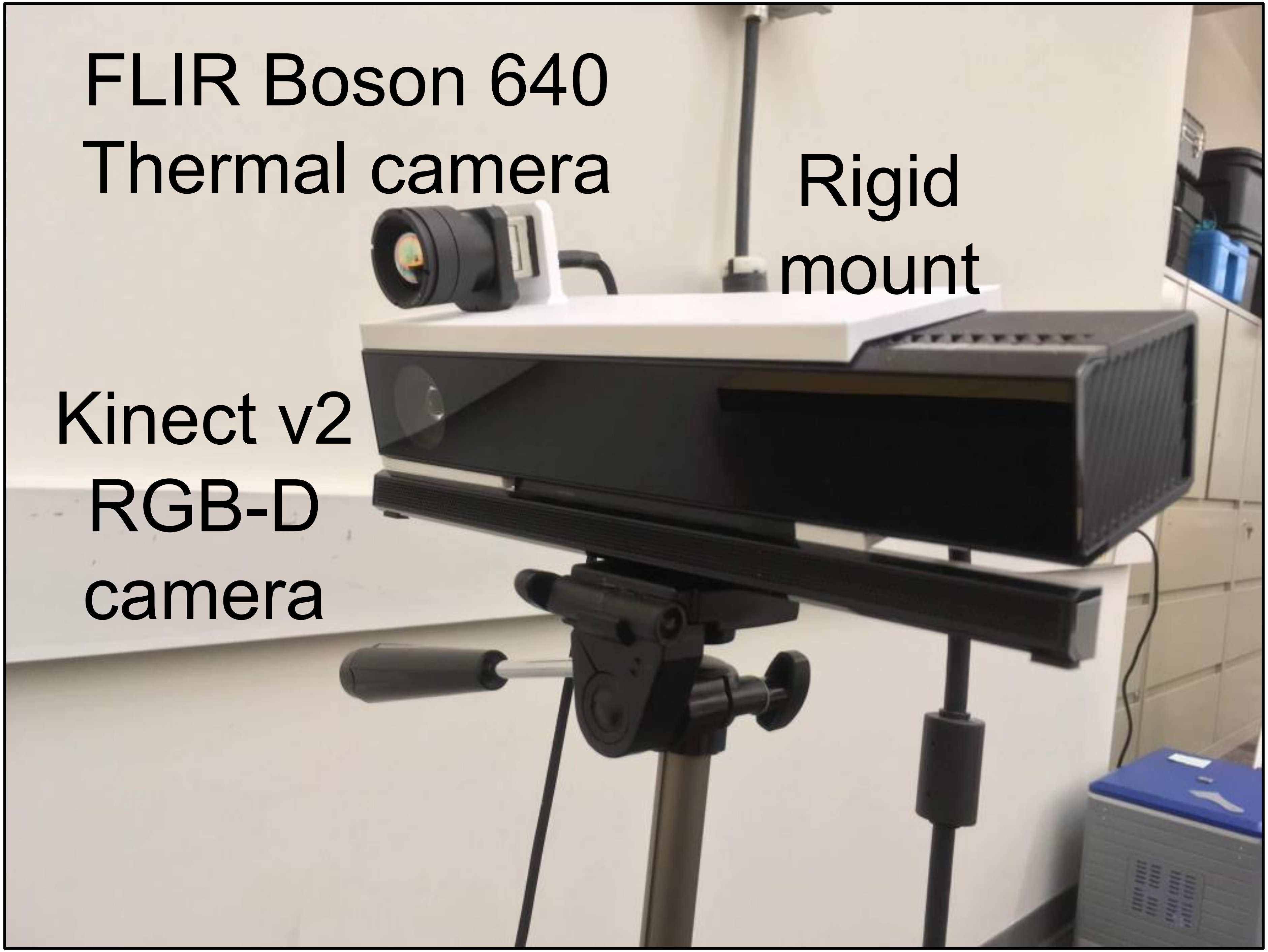
CVPR ’19 oral and best paper finalist

2 seconds

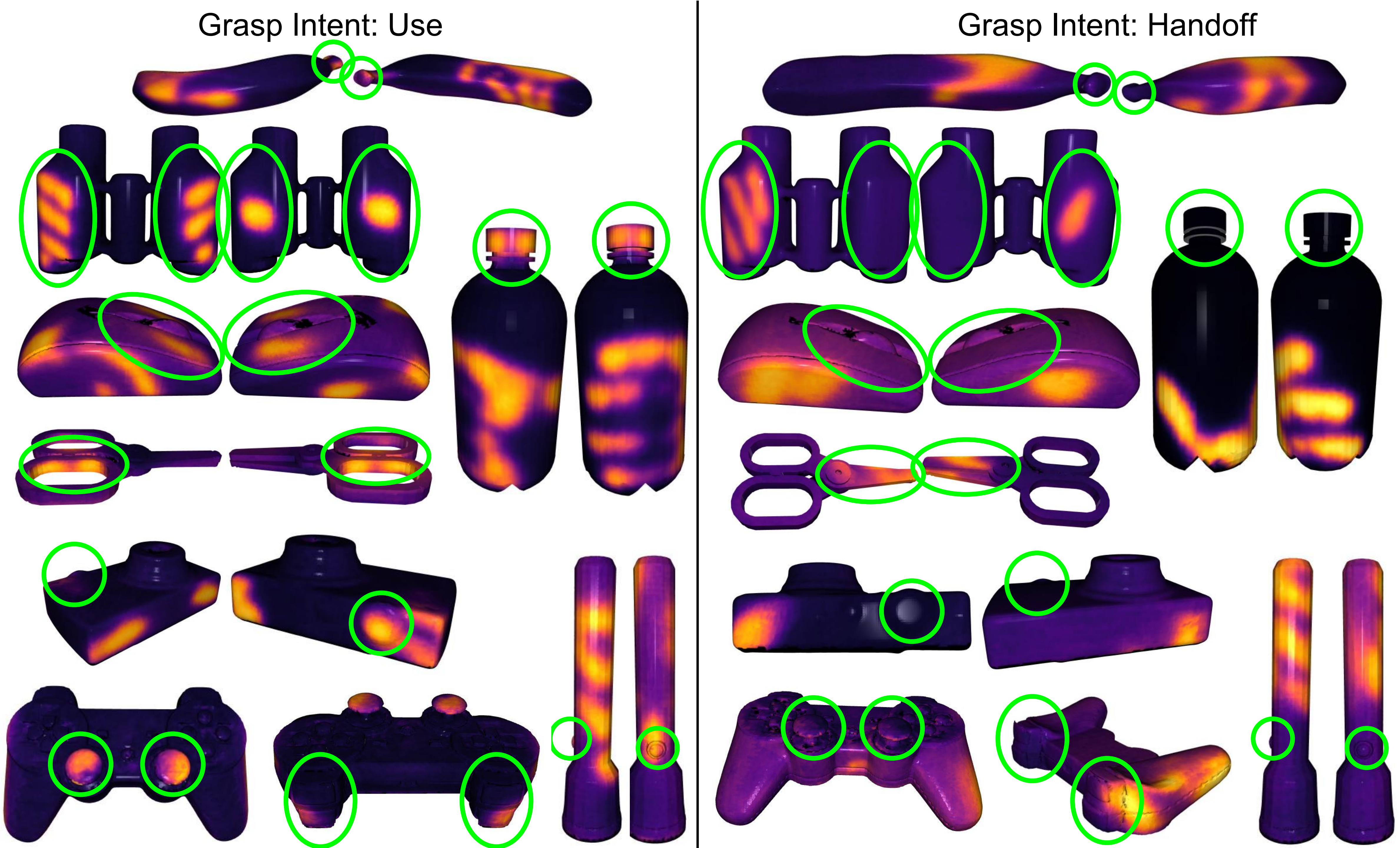
5 seconds

10 seconds

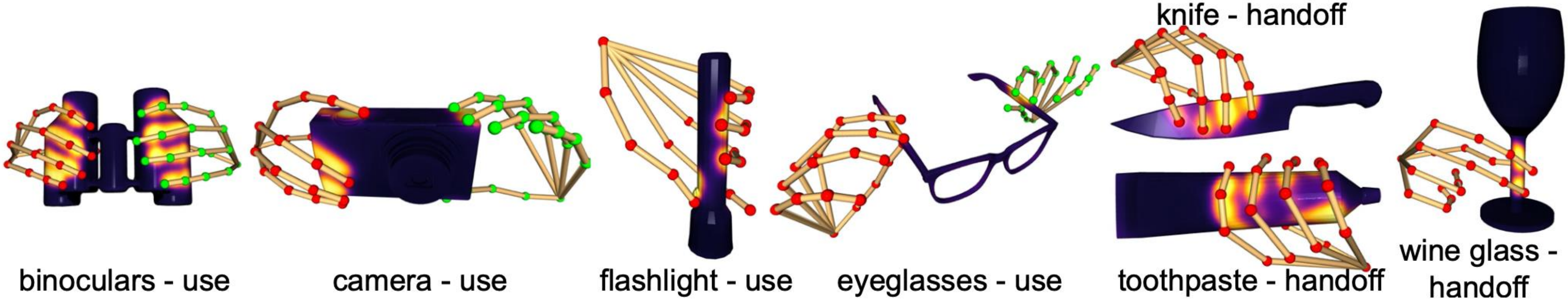

Computer
Turntable
Camera
Table with
3D printed
objects

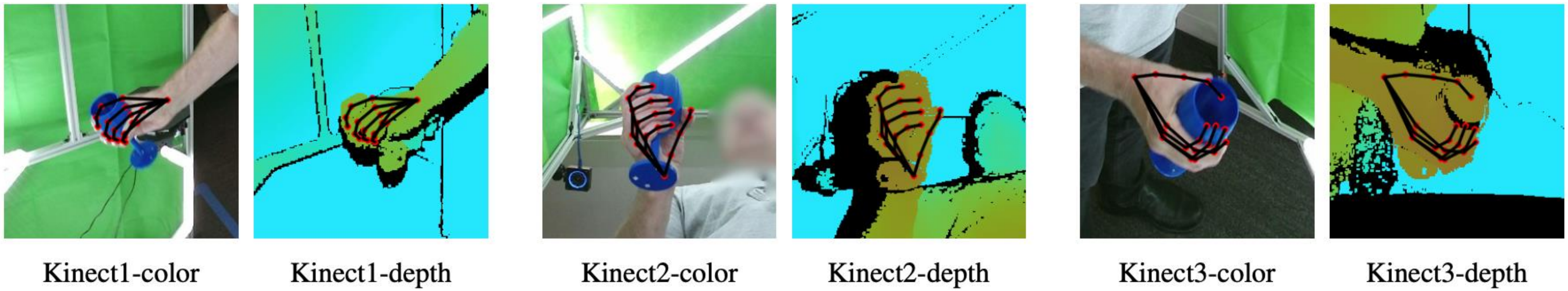

FLIR Boson 640
Thermal camera
Rigid
mount
Kinect v2
RGB-D
camera

Contact map


Grasp Intent: Use
Grasp Intent: Handoff

ContactPose: Capturing Contact + Hand Pose

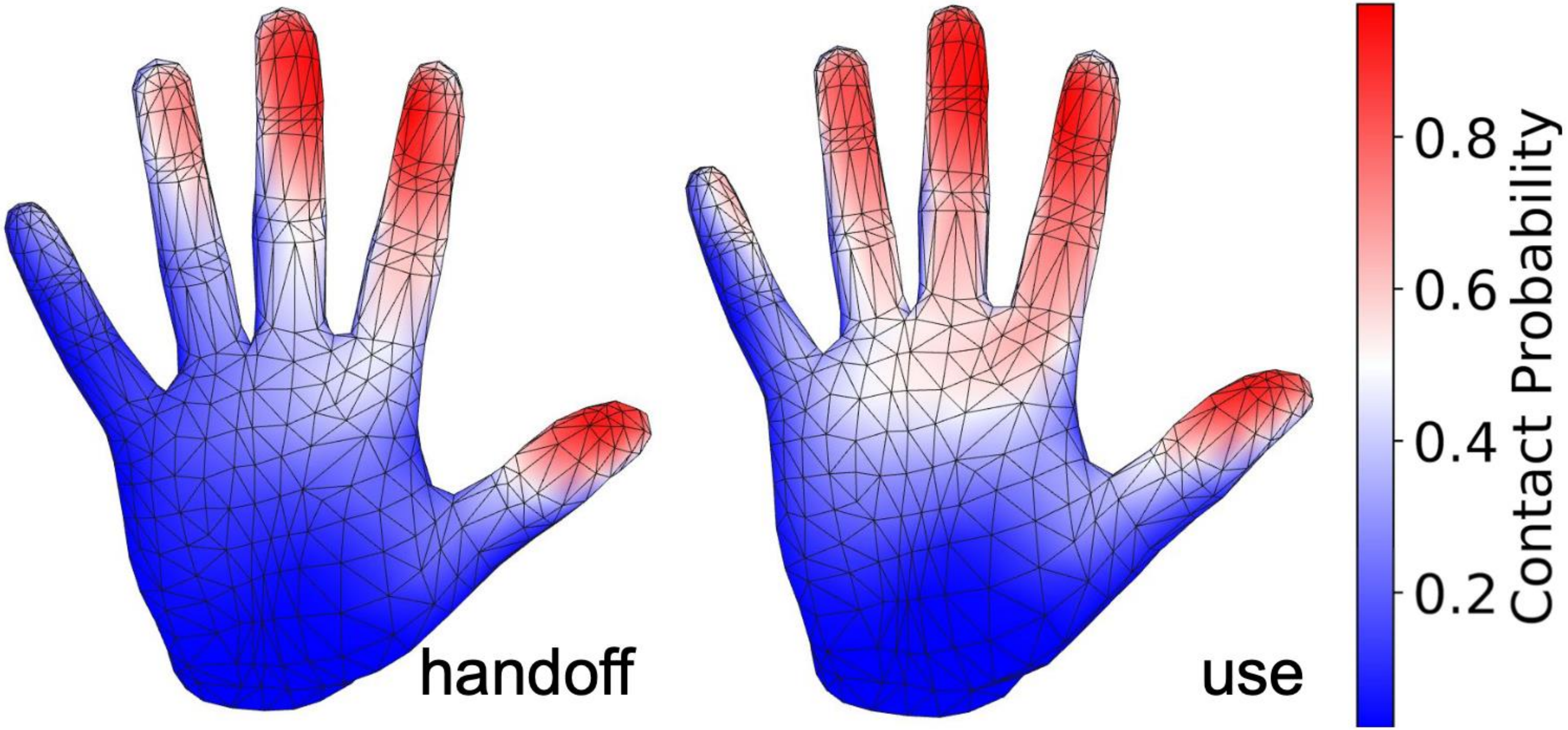


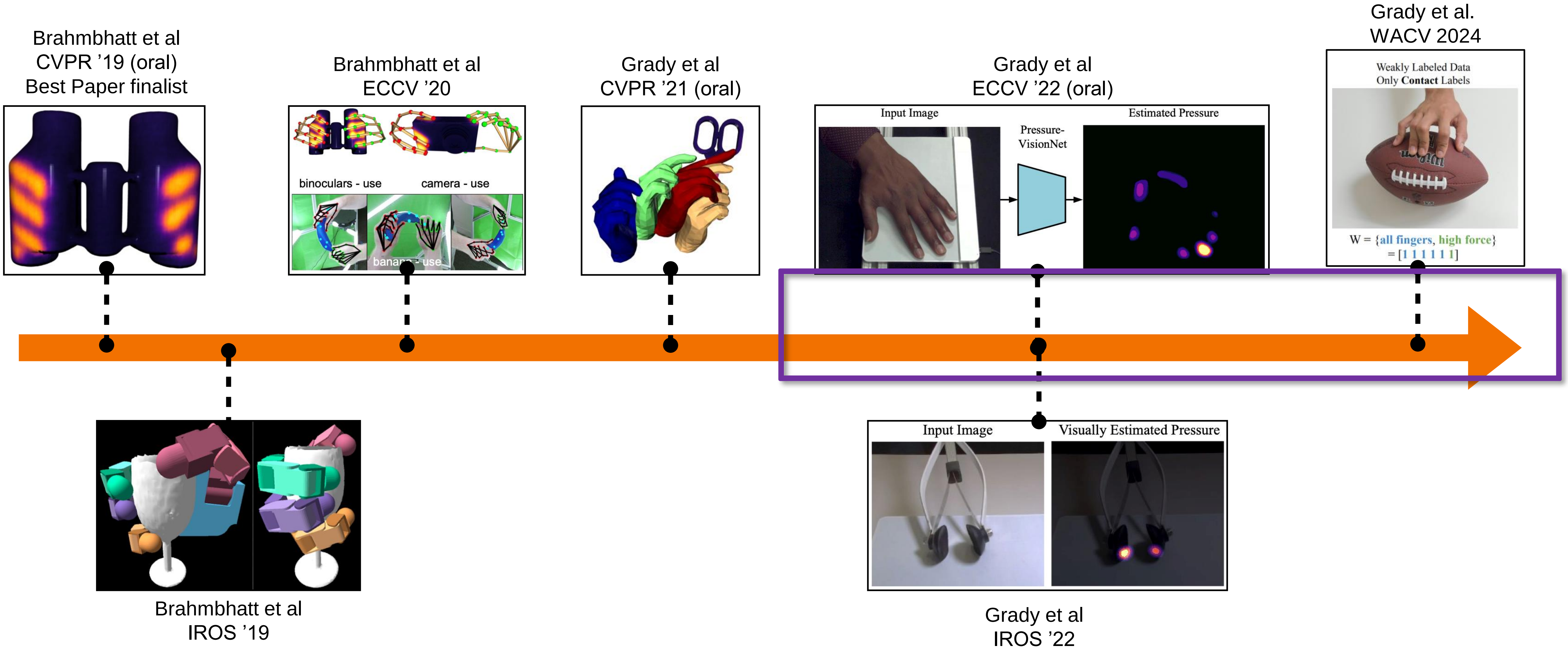


Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays

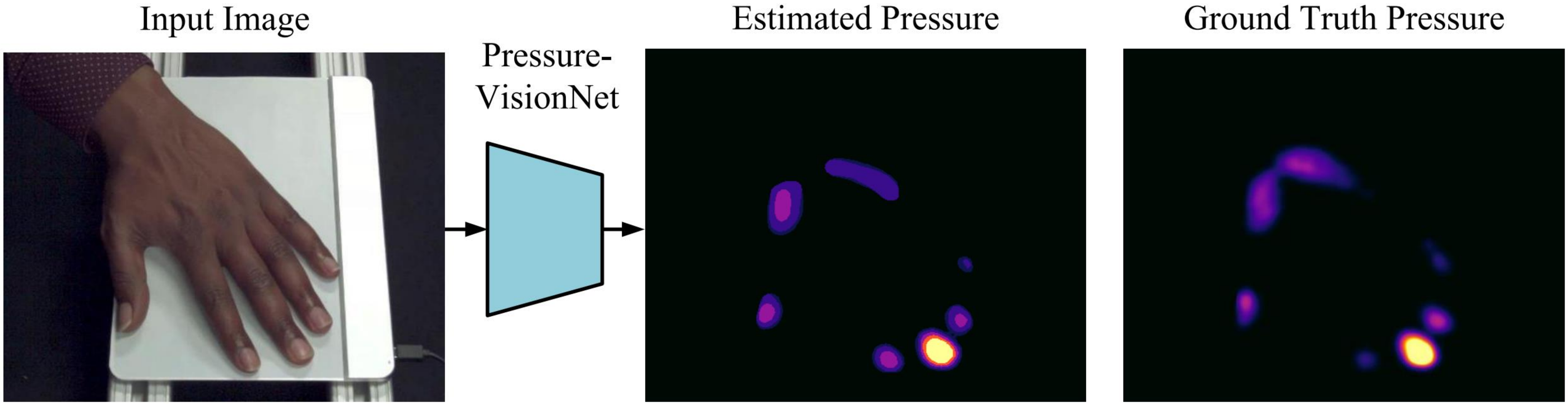
ECCV 2020

Hand Contact Probability




Brahmbhatt et al
CVPR ’19 (oral)
Best Paper finalist
Brahmbhatt et al
ECCV ’20
binoculars - use
camera - use
Grady et al
CVPR ’21 (oral)
Grady et al
ECCV ’22 (oral)
Input Image
Pressure-
VisionNet
Estimated Pressure
Grady et al.
WACV 2024
Weakly Labeled Data
Only Contact Labels
W = {all fingers, high force}
= [1 1 1 1 1 1]
Brahmbhatt et al
IROS ’19
Input Image
Visually Estimated Pressure
Grady et al
IROS ’22

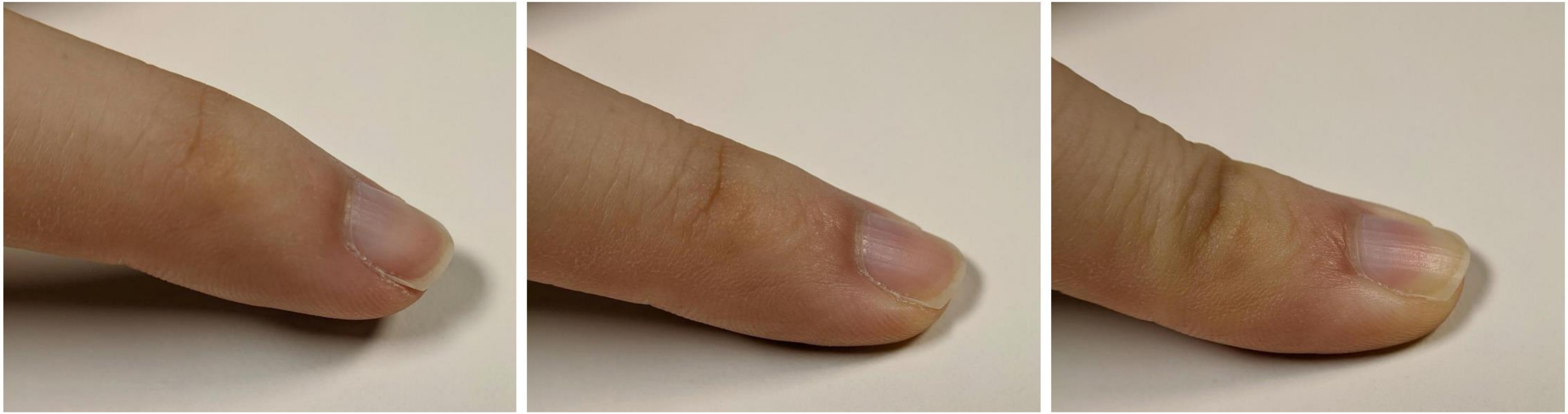
PressureVision: Estimating Hand Pressure from a Single RGB Image



Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D. Twigg, Chengde Wan, James Hays, and Charles C. Kemp

ECCV 2022 Oral

No Contact | Low Force | High Force

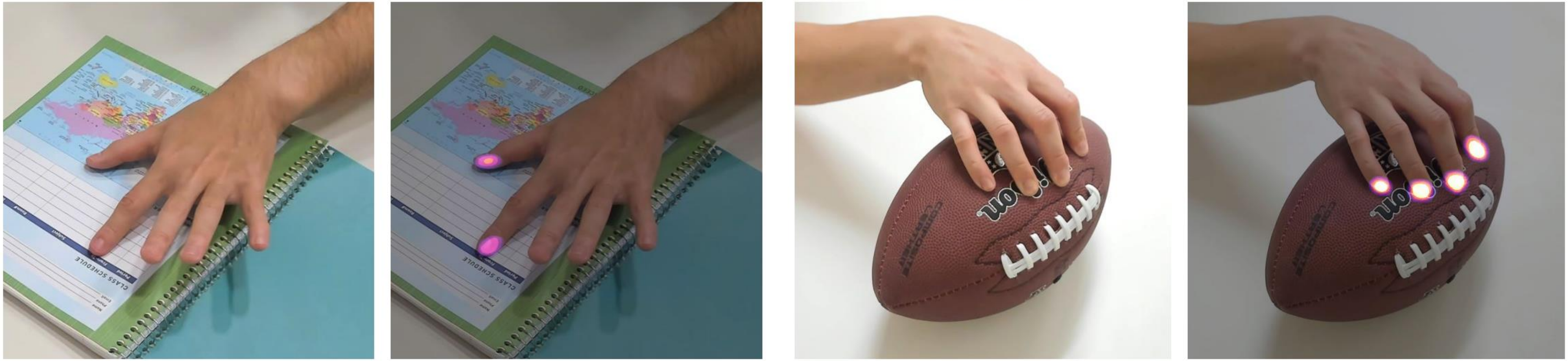


We train a deep network, PressureVisionNet,
to estimate pressure from a single RGB image.

The pressure for each frame is calculated independently.

PressureVision++: Estimating Fingertip Pressure From Diverse RGB Images

Input Image | Estimated Pressure | Input Image | Estimated Pressure



Prompt: Press thumb and index, low force | Prompt: Press all fingers, high force

Patrick Grady, Jeremy Collins, Chengcheng Tang,
Christopher D. Twigg, James Hays, and Charles C. Kemp
WACV 2024


Surface/Prompt
Image
ContactLabelNet
Wall
Press index
Low force
Foam mat
Press all fingers
High force
Foam mat
No contact
Mirror
Press middle
Low force
Mirror
Press all fingers
High force
Surface/Prompt
Image
ContactLabelNet
Football
Press pinky
High force
Notebook
Press all fingers
Low force
Notebook
Press index, thumb
Low force
Glass
Press index
High force
Glass
Press ring
Low force

The Un-Kidnappable Robot: Acoustic Localization of Sneaking People

Mengyu Yang, Patrick Grady, Samarth Brahmbhatt,
Arun Balajee Vasudevan, Charles C. Kemp, James Hays

ICRA2024 YOKOHAMA | JAPAN

Georgia Tech
intel labs
Carnegie Mellon University
hello robot

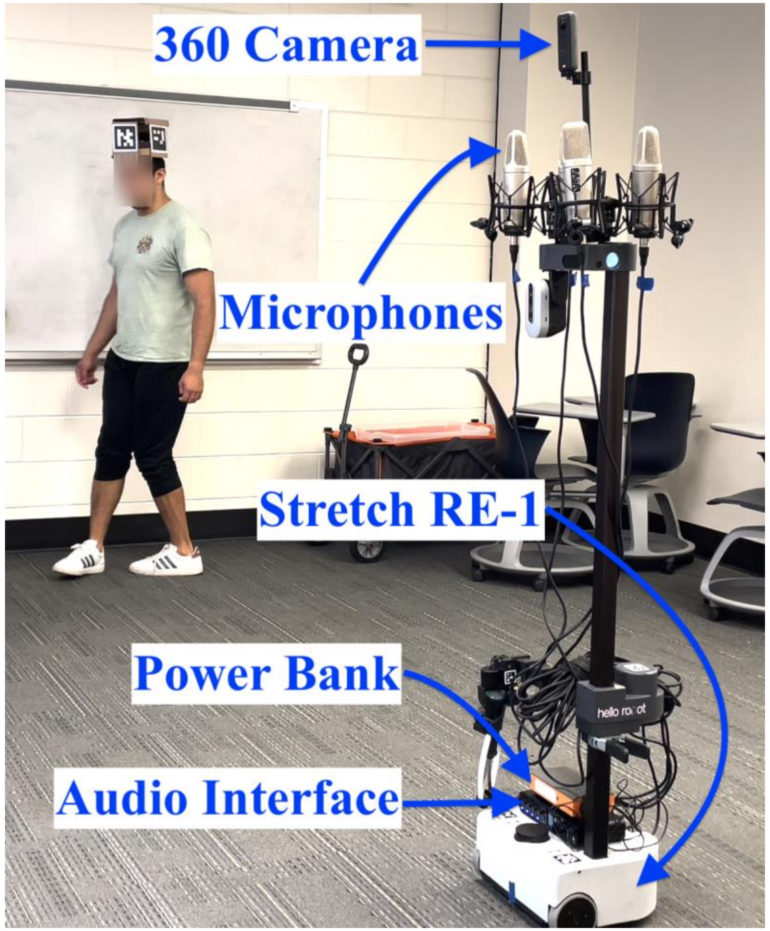
ICRA2024
YOKOHAMA | JAPAN

How easy is it to sneak up on a robot?

We train robots to detect people using *only* the subtle and incidental sounds they produce as they move around, even when they try to be quiet.

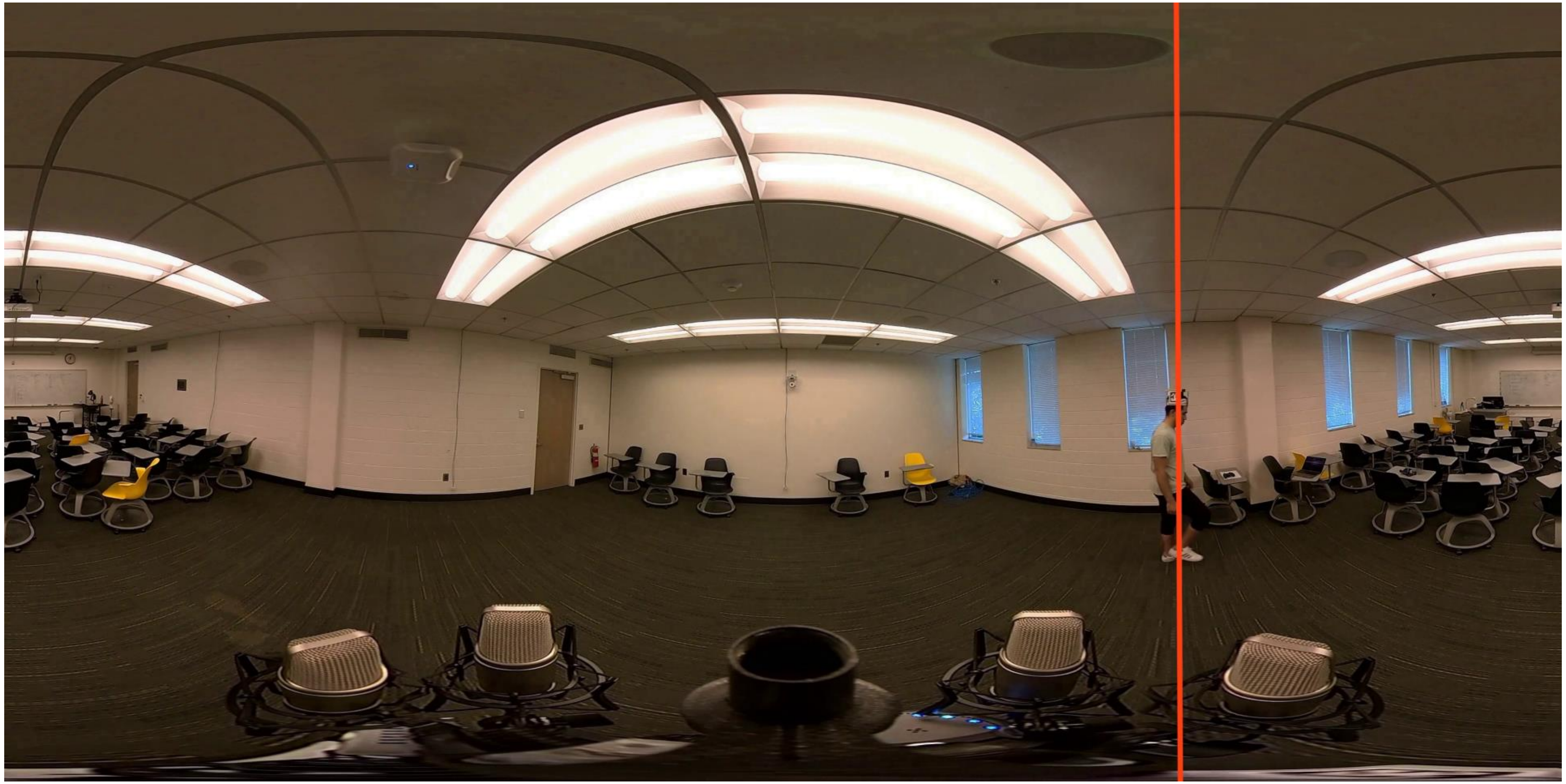
Real Time

The Robot Kidnapper Dataset



- 4-channel audio
- 360 degree egocentric video
- 12 participants in 8 indoor settings:
  - Standing still
  - Walking quietly
  - Walking normally
  - Walking loudly

Data Annotation

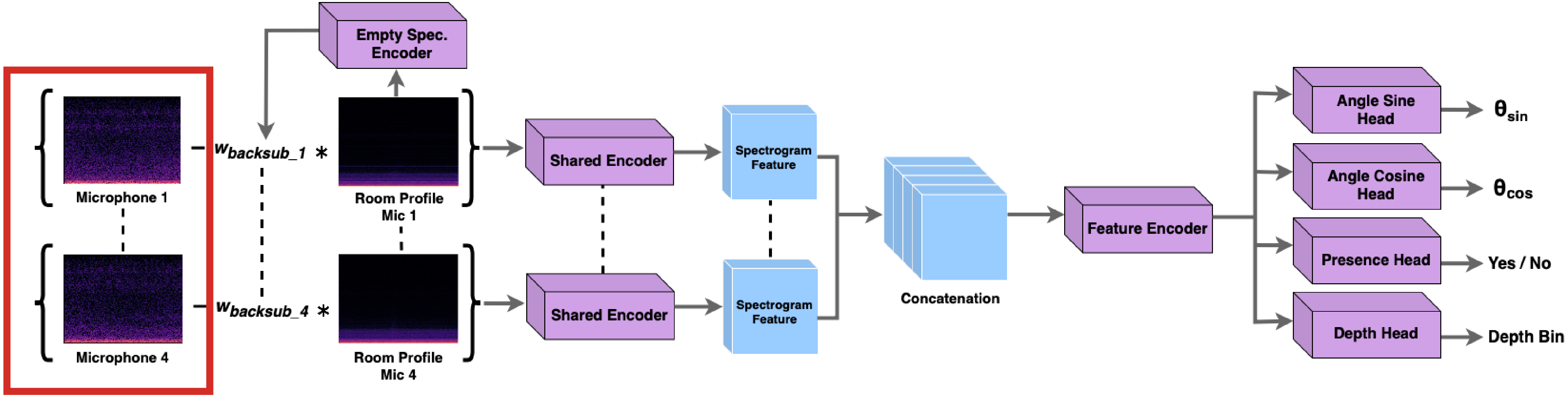


Azimuthal angle

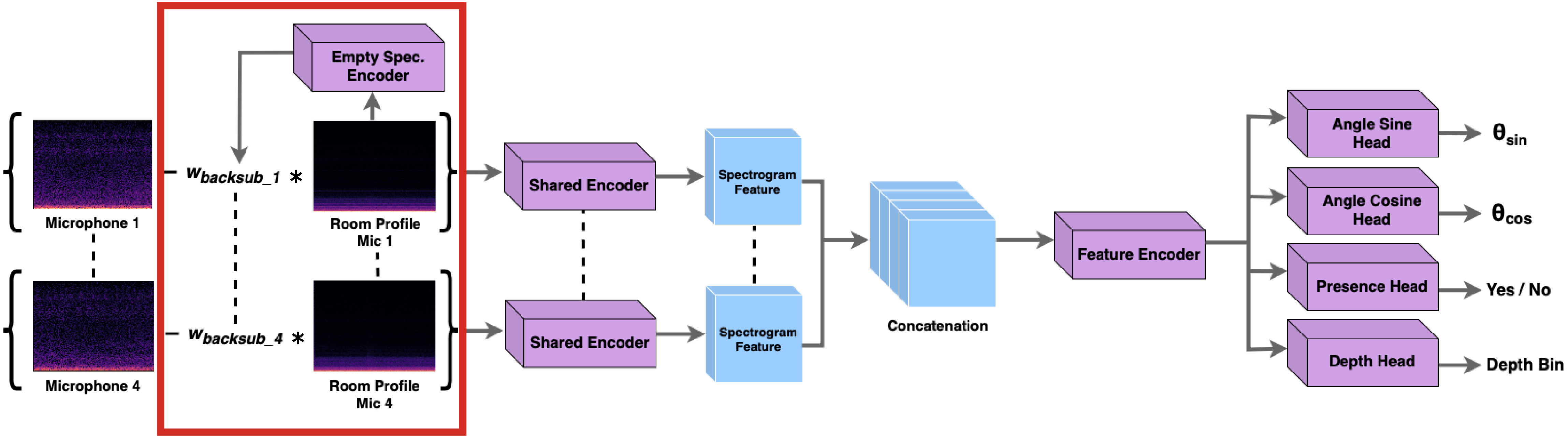
Data Annotation

Radial distance from detecting ArUco markers

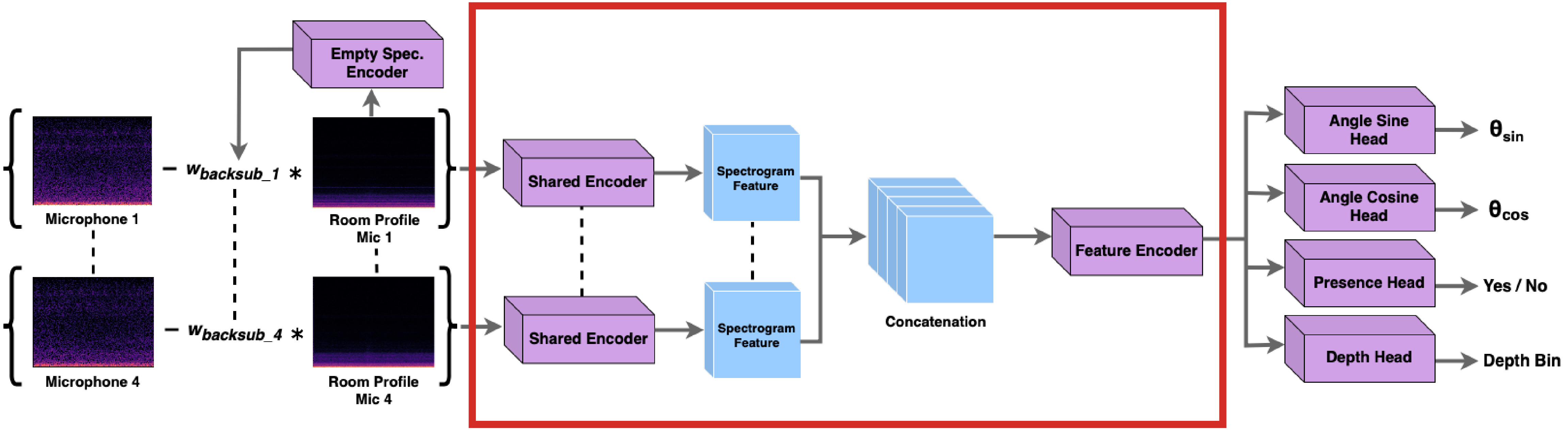
Architecture



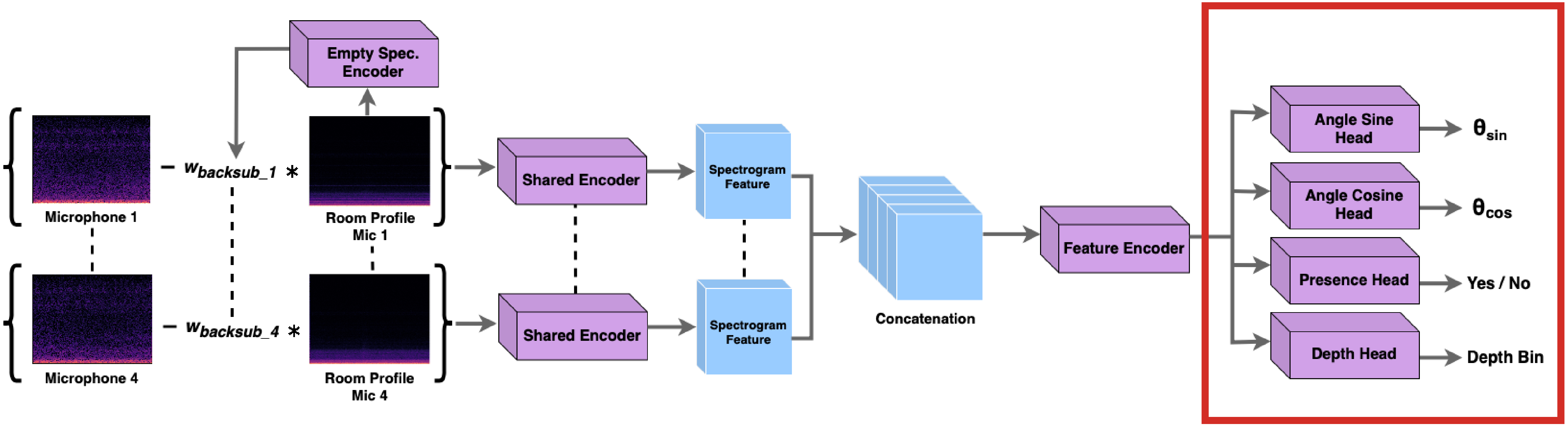
Architecture



Architecture



Architecture



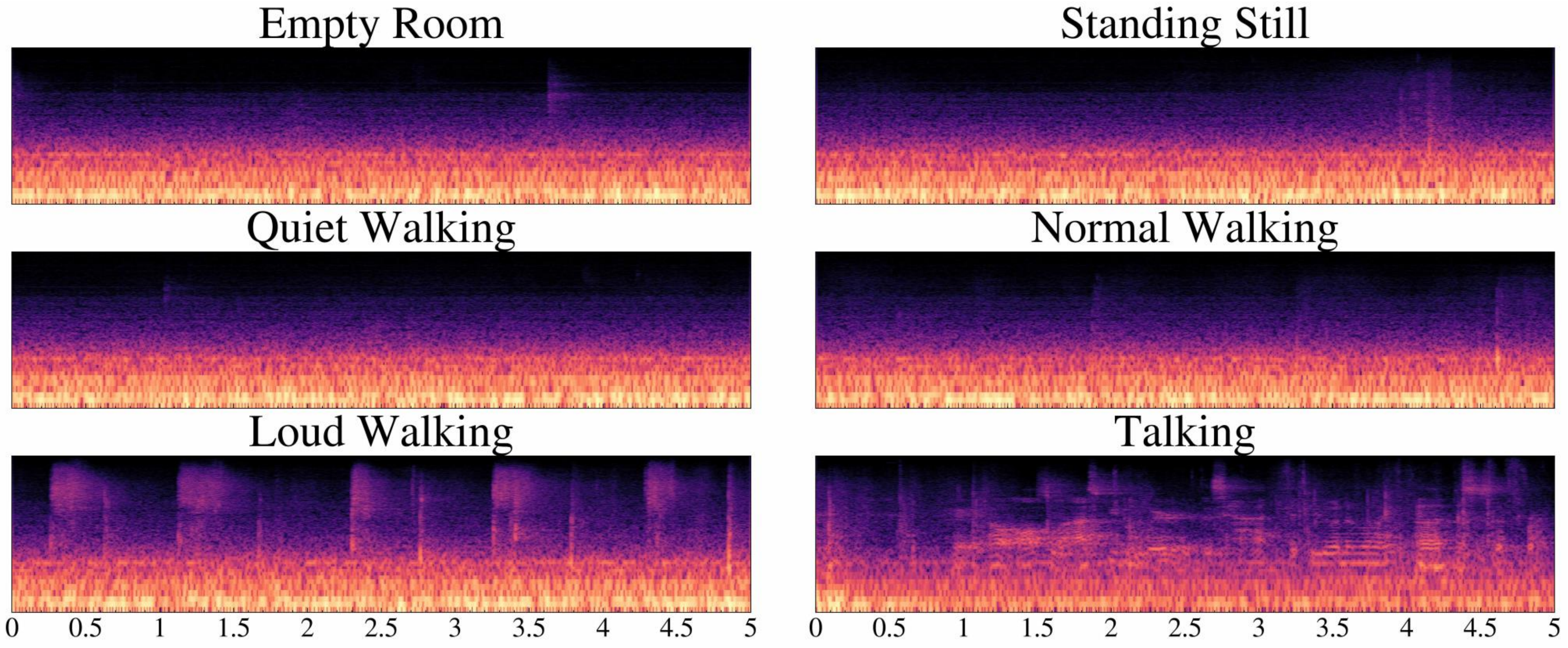
Azimuthal Angle Prediction

ICRA2024
YOKOHAMA | JAPAN

CATEGORY	MODEL	*Quiet* Sta.	*Quiet* Dyn.	*Normal* Sta.	*Normal* Dyn.	*Loud* Sta.	*Loud* Dyn.
Random	Uniform 360°	90	90	90	90	90	90
Oracle Mic Pair	Constant Front	50	43	50	46	50	43
	GCC-PHAT [25]	44	47	45	43	46	47
	StereoCRW [26]	52	46	51	48	37	34
Ours	1 Mic	67	75	64	71	64	74
	2 Mics	37	54	37	48	36	47
	Base 4 Mics	47	55	50	48	49	47
	4 Mics	**21**	**26**	**22**	**24**	**19**	**22**

Mean absolute error (MAE) in degrees

Qualitative Comparisons



Conclusion

- Human detection with only subtle incidental sounds of them moving
- Robot Kidnapper dataset collected on robot in real-world indoor environments
- Our model outperforms previous sound localization methods
- Real-time detection on robot

Today's Class

- ~~Who am I?~~
- What is Computer Vision?
- Specifics of this course
- Geometry of Image Formation
- Questions

What is Computer Vision?

Derogatory summary of computer vision:
Machine learning applied to visual data

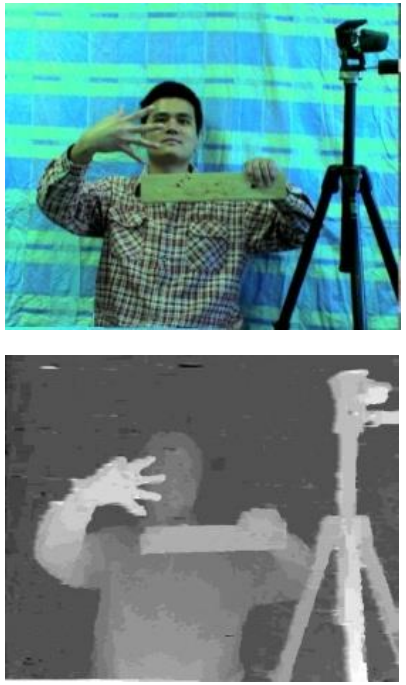
Computer Vision

- Automatic understanding of images and video
    1. Computing properties of the 3D world from visual data *(measurement)*

Slide credit: Kristen Grauman

1. Vision for measurement

Real-time stereo



Wang et al.

Structure from motion

Snavely et al.

Tracking

Demirdjian et al.

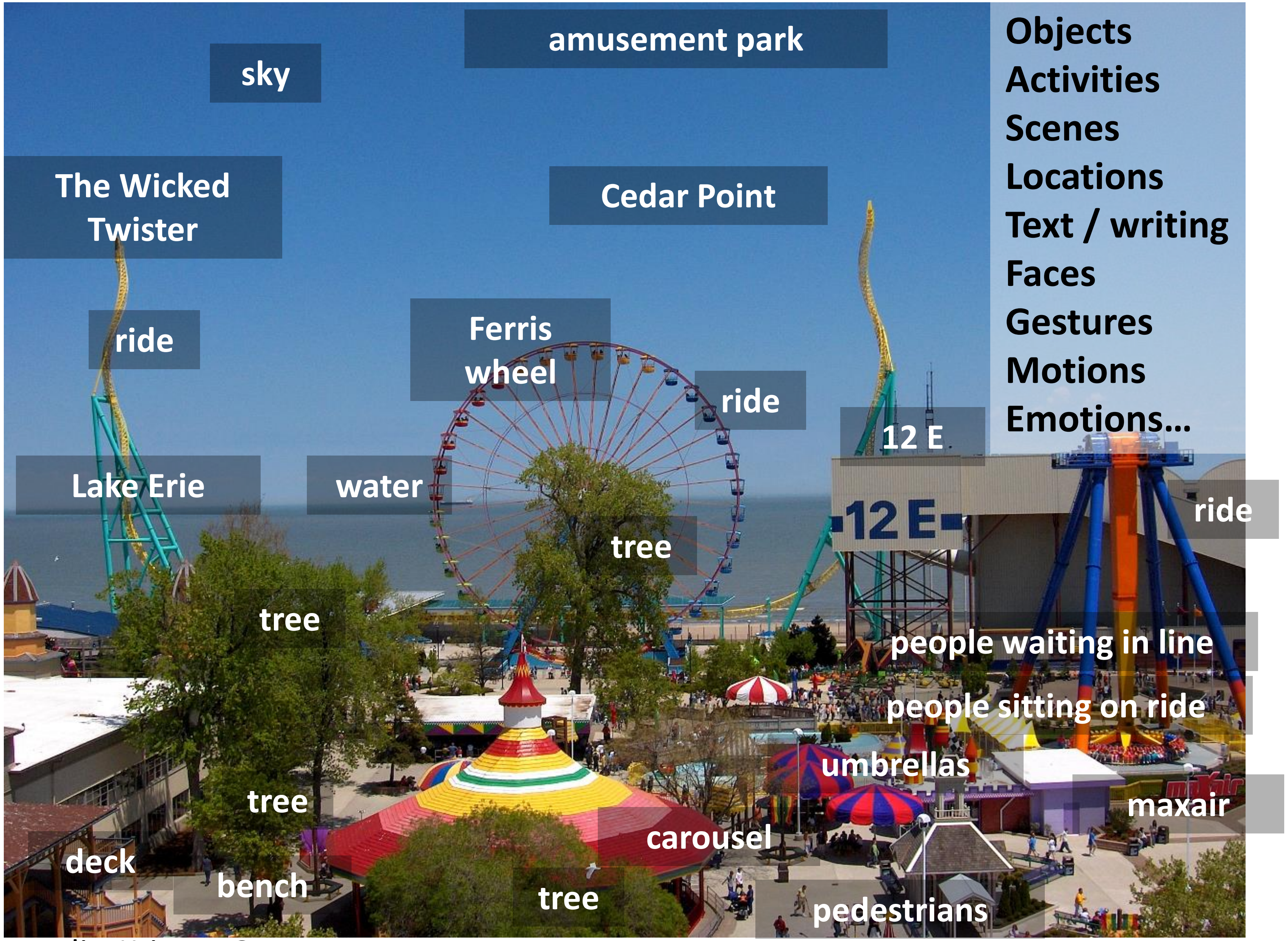
Slide credit: Kristen Grauman

Computer Vision

- Automatic understanding of images and video
  1. Computing properties of the 3D world from visual data *(measurement)*
  2. Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. *(perception and interpretation)*

Slide credit: Kristen Grauman

2. Vision for perception, interpretation



Objects
Activities
Scenes
Locations
Text / writing
Faces
Gestures
Motions
Emotions…

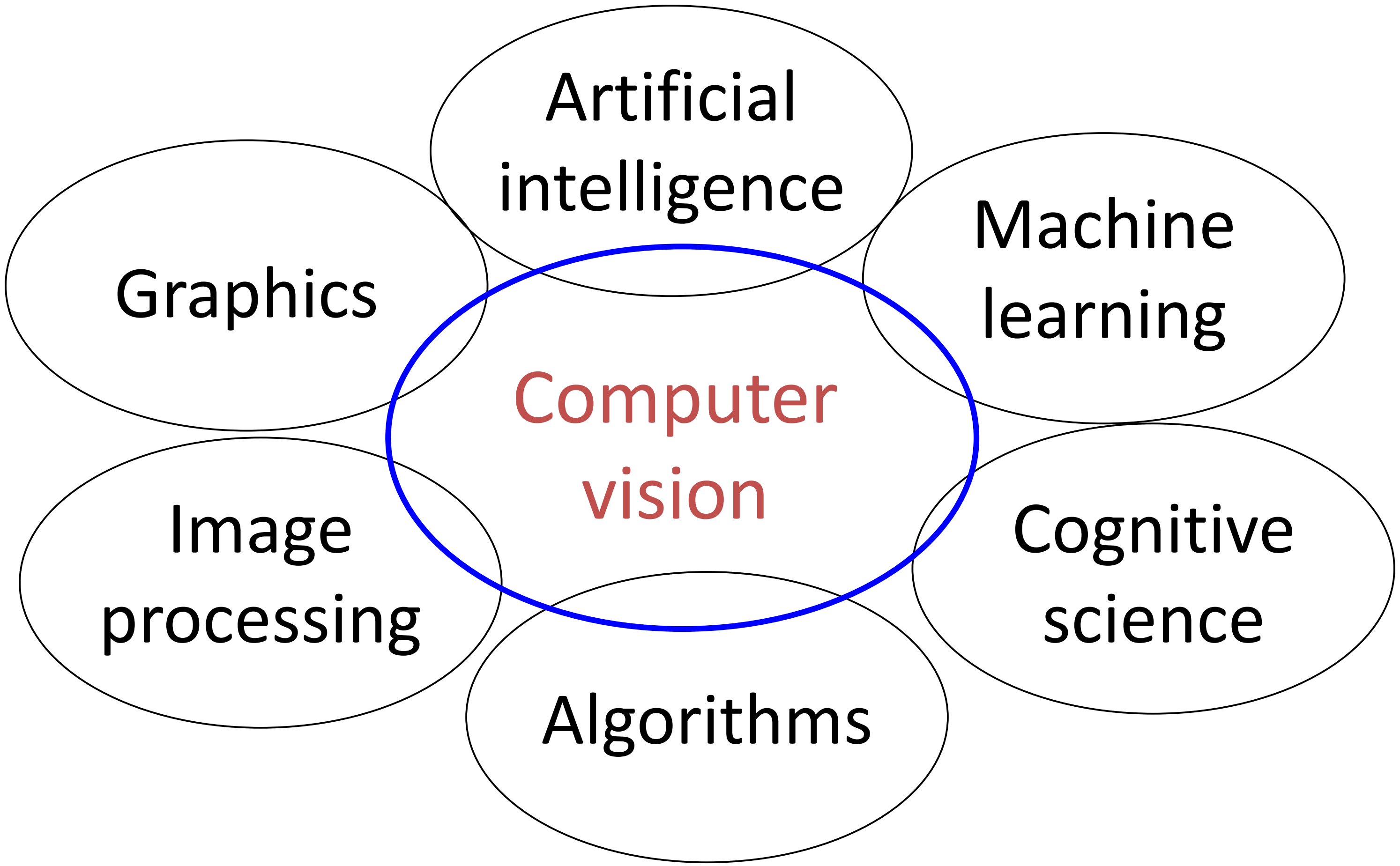
Slide credit: Kristen Grauman

Computer Vision

- Automatic understanding of images and video
    1. Computing properties of the 3D world from visual data *(measurement)*
    2. Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. *(perception and interpretation)*
    3. Algorithms to mine, search, and interact with visual data (*interaction)*

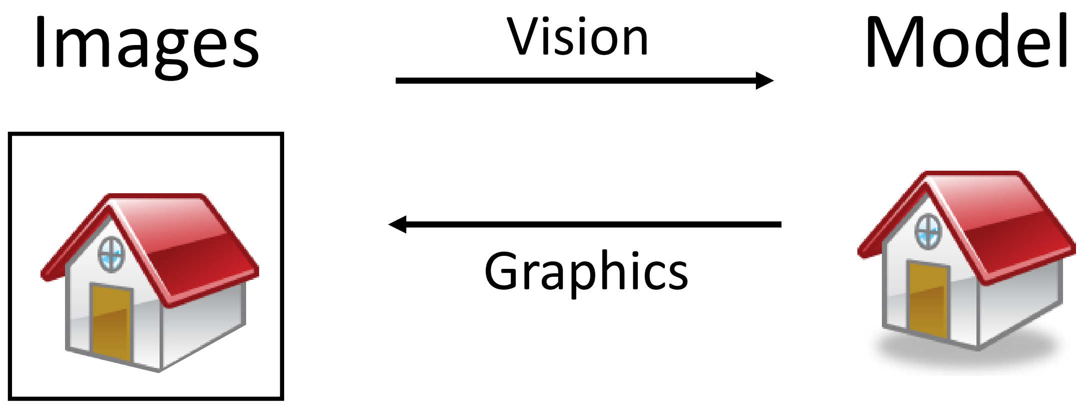
3. Interaction

Related disciplines



Slide credit: Kristen Grauman

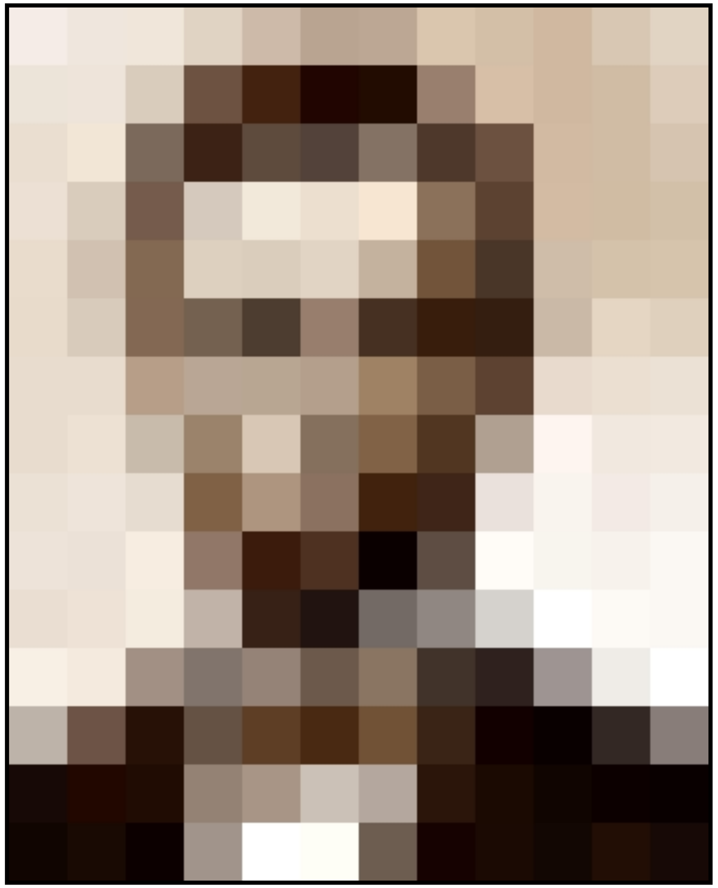
Vision and graphics



Inverse problems: analysis and synthesis.

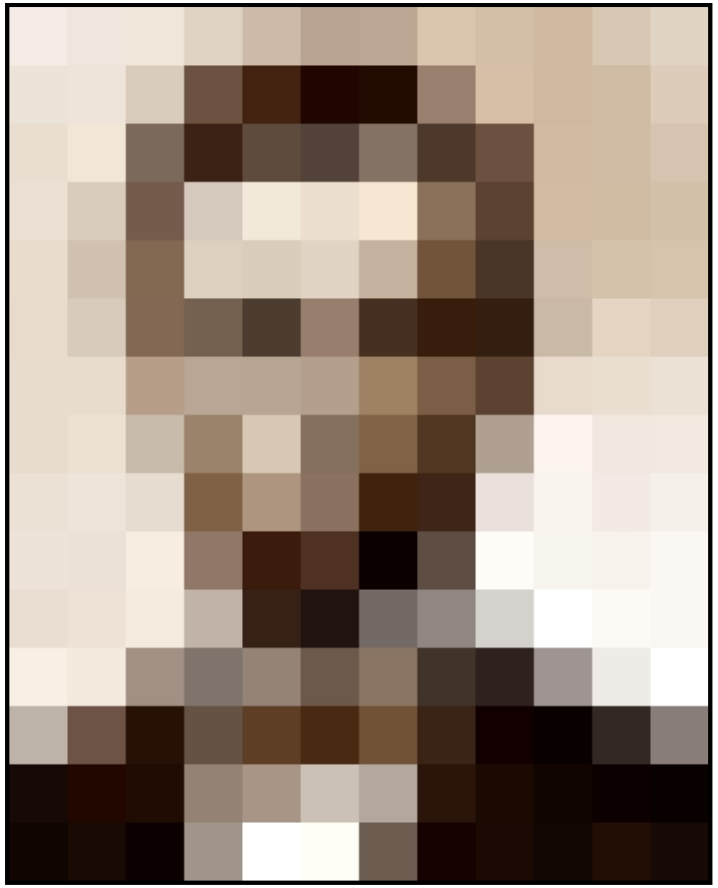
Slide credit: Kristen Grauman

What humans see



Slide credit: Larry Zitnick

What computers see



Slide credit: Larry Zitnick

What do humans see?

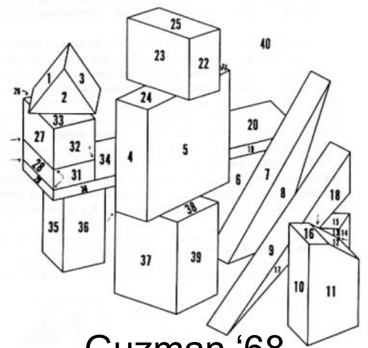
Slide credit: Larry Zitnick

Vision is really hard

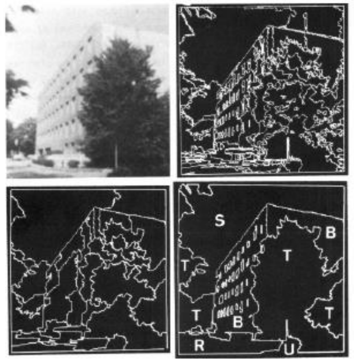
- Vision is an amazing feat of natural intelligence
  - Visual cortex occupies about 50% of Macaque brain
  - One third of human brain devoted to vision (more than anything else)

Ridiculously brief history of computer vision

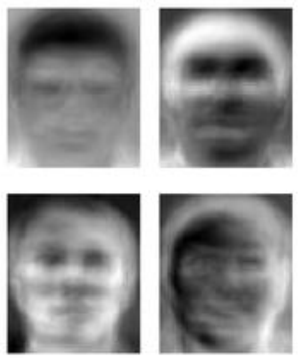
- 1966: Minsky assigns computer vision as an undergrad summer project
- 1960’s: interpretation of synthetic worlds
- 1970’s: some progress on interpreting selected images
- 1980’s: ANNs come and go; shift toward geometry and increased mathematical rigor
- 1990’s: face recognition; statistical analysis in vogue
- 2000’s: broader recognition; large annotated datasets available; video processing starts
- 2010’s: Deep learning with ConvNets
- 2020’s: Widespread autonomous vehicles?
- 2030’s: robot uprising?



Guzman ‘68



Ohta Kanade ‘78



Turk and Pentland ‘91

How vision is used now

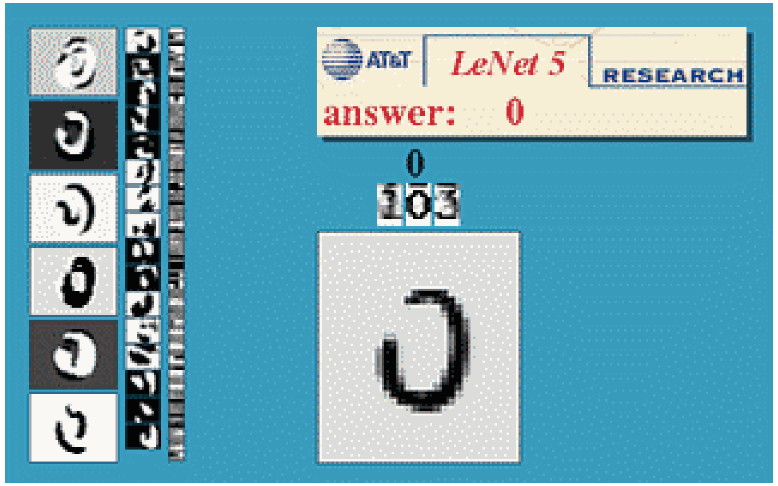
- Examples of real-world applications

Some of the following slides by Steve Seitz

Optical character recognition (OCR)

Technology to convert scanned docs to text

- If you have a scanner, it probably came with OCR software



Digit recognition, AT&T labs
http://www.research.att.com/~yann/

License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

Optical character recognition (OCR)

- Most US postal service mail is automatically read.
- In 1997, there were 55 offices reviewing images of 19 billion pieces of mail that OCR failed on.
- Today, there is 1 office, and they only looked at 1.2 billion pieces of mail this year.

https://www.youtube.com/watch?v=XxCha4Kez9c

Face detection

- Digital cameras detect faces

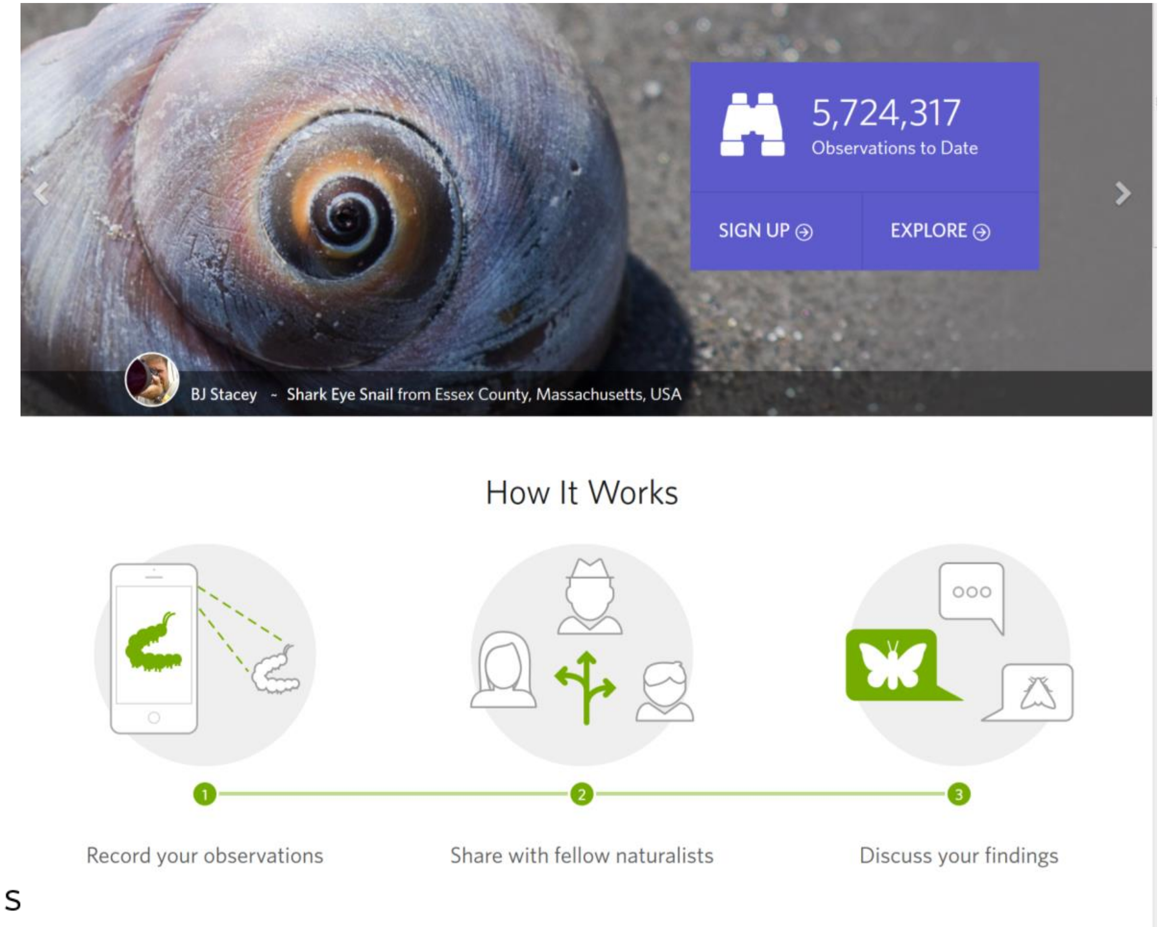
Vision in space

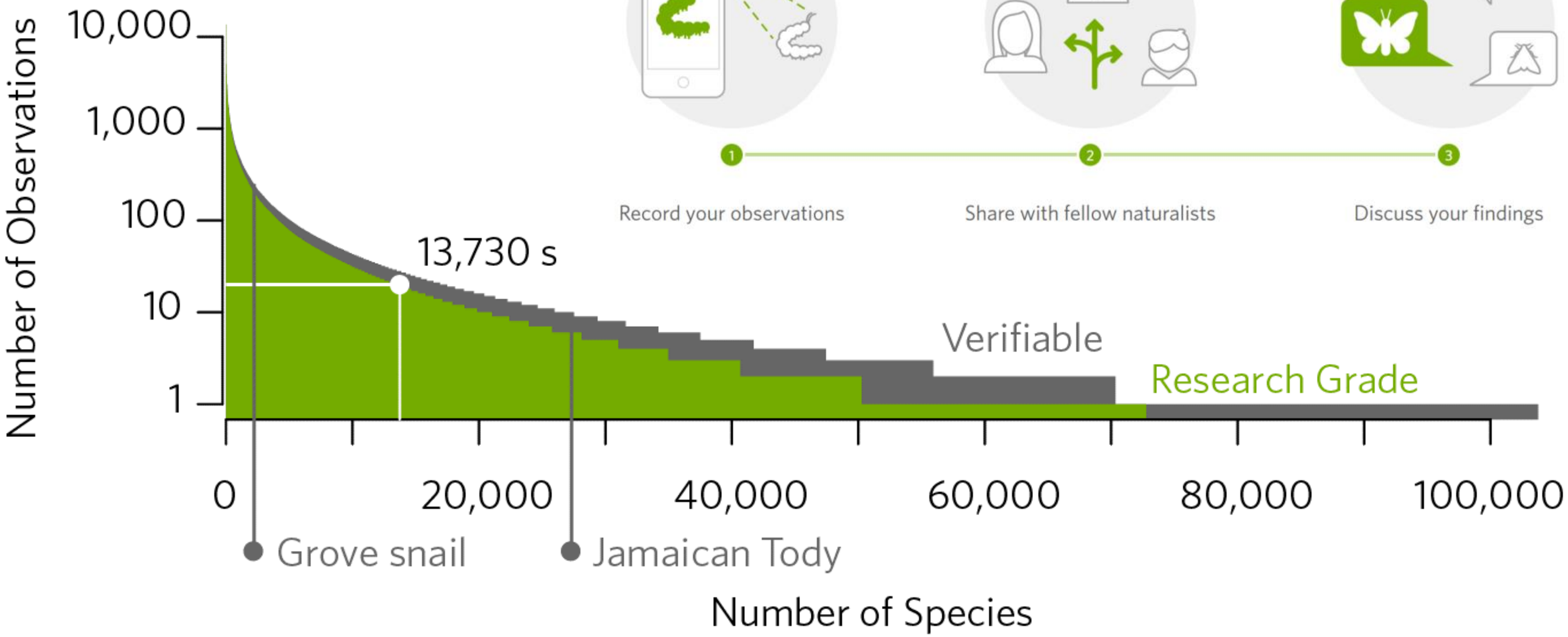
NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

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.

iNaturalist





https://www.inaturalist.org/pages/computer_vision_demo

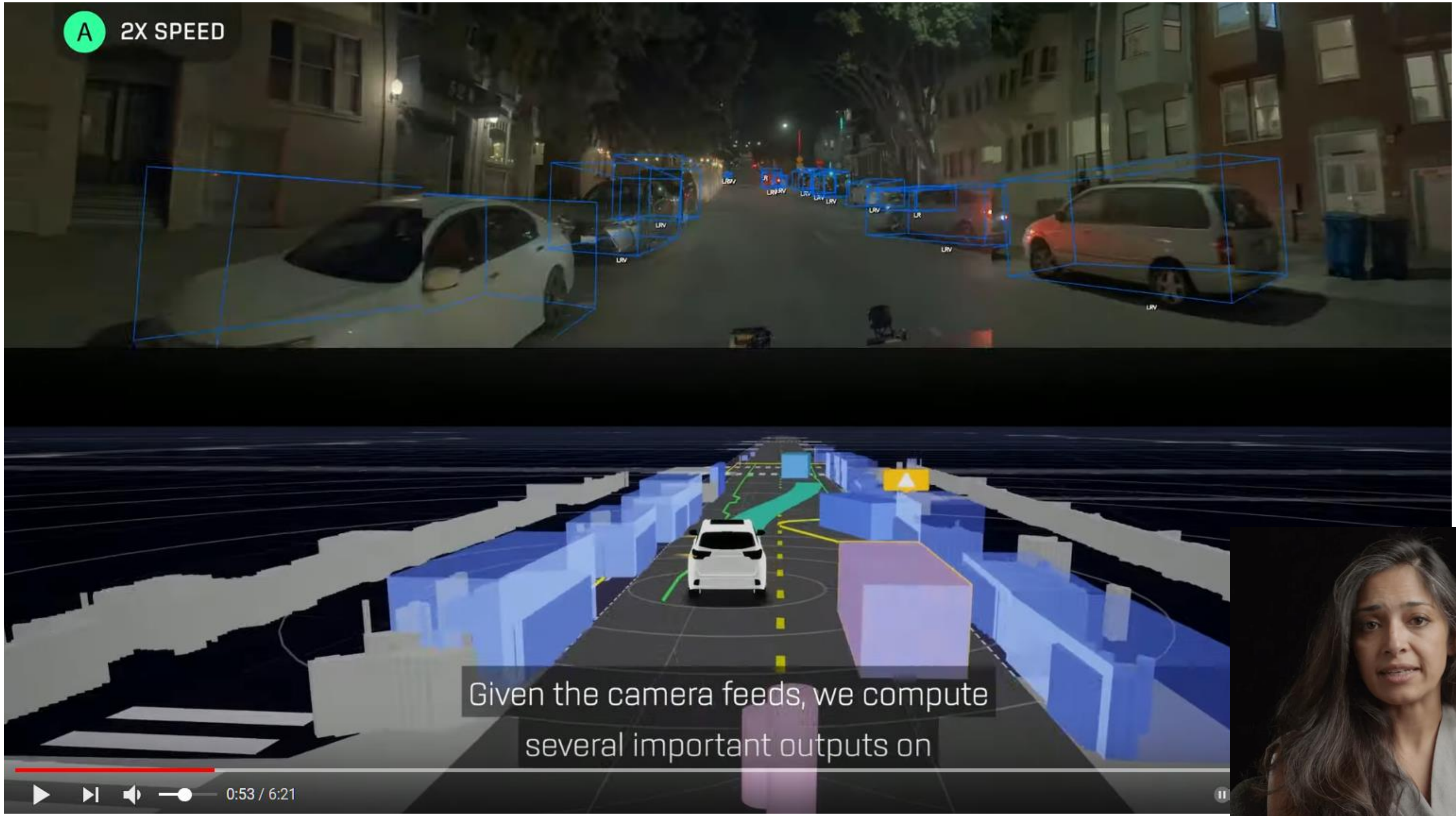
Amazon Prime Air

https://www.amazon.com/b?node=8037720011

Skydio

https://www.skydio.com/

Zoox Computer Vision Demo

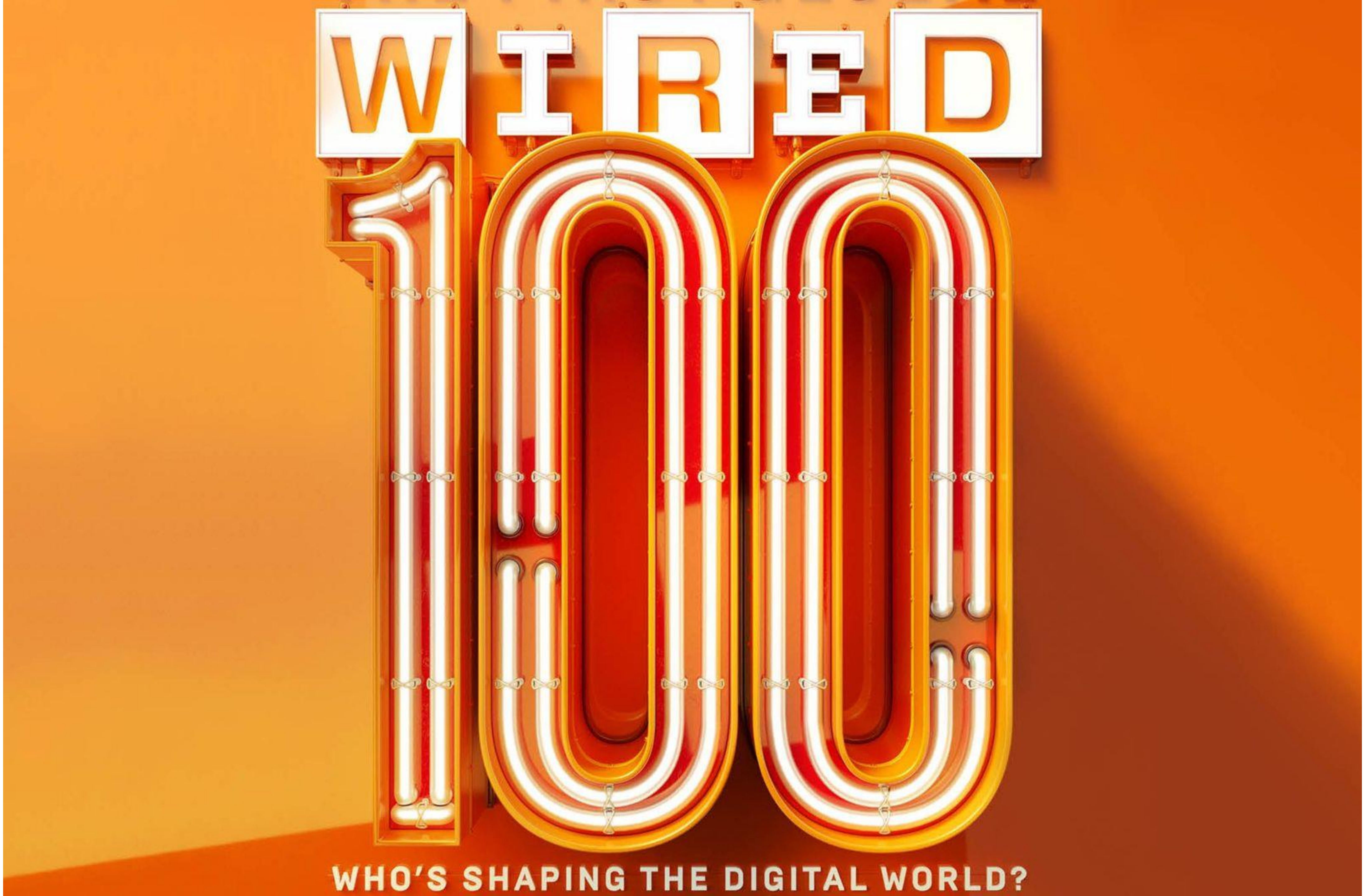


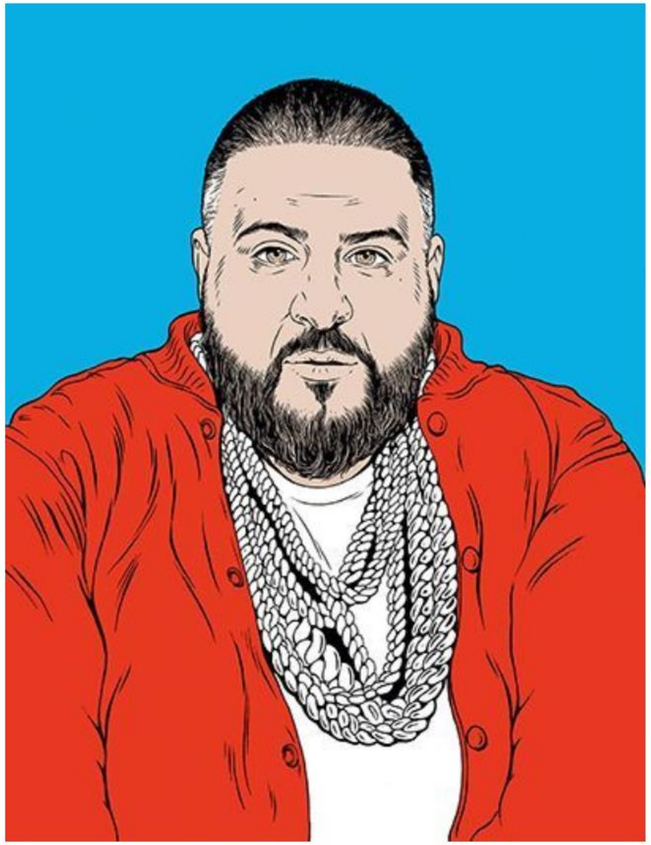
https://www.youtube.com/watch?v=BVRMh9NO9Cs

State of the art today?

With enough training data, computer vision ~~nearly~~ matches human vision at most recognition tasks

Deep learning has been an enormous disruption to the field. More and more techniques are being “deepified”.


WIRED
100
WHO'S SHAPING THE DIGITAL WORLD?



DJ Khaled

Credit **Louise Zergaeng Pomeroy**

73. DJ Khaled

Snapchat icon; DJ and producer

Louisiana-born Khaled Mohamed Khaled, aka DJ Khaled, cut his musical chops in the early 00s as a host for Miami urban music radio WEDR. He proceeded to build a solid if not dazzling career as a mixtape DJ and music producer (he founded his label We The Best Music Group in 2008, and was appointed president of Def Jam South in 2009).

69. Geoffrey Hinton

Psychologist, computer scientist; researcher, Google Toronto

British-born Hinton has been dubbed the "godfather of deep learning". The Cambridge-educated cognitive psychologist and computer scientist started being an ardent believer in the potential of neural networks and deep learning in the 80s, when those technologies enjoyed little support in the wider AI community.

But he soldiered on: in 2004, with support from the Canadian Institute for Advanced Research, he launched a University of Toronto programme in neural computation and adaptive perception, where, with a group of researchers, he carried on investigating how to create computers that could behave like brains.

Hinton's work – in particular his algorithms that train multilayered neural networks – caught the attention of tech giants in Silicon Valley, which realised how deep learning could be applied to voice recognition, predictive search and machine vision.

The spike in interest prompted him to launch a free course on neural networks on e-learning platform Coursera in 2012. Today, 68-year-old Hinton is chair of machine learning at the University of Toronto and moonlights at Google, where he has been using deep learning to help build internet tools since 2013.

63. Yann Lecun

Director of AI research, Facebook, Menlo Park

LeCun is a leading expert in deep learning and heads up what, for Facebook, could be a hugely significant source of revenue: understanding its user's intentions.

62. Richard Branson

Founder, Virgin Group, London

Branson saw his personal fortune grow £550 million when Alaska Air bought Virgin America for $2.6 billion in April. He is pressing on with civilian space travel with Virgin Galactic.

61. Taylor Swift

Entertainer, Los Angeles

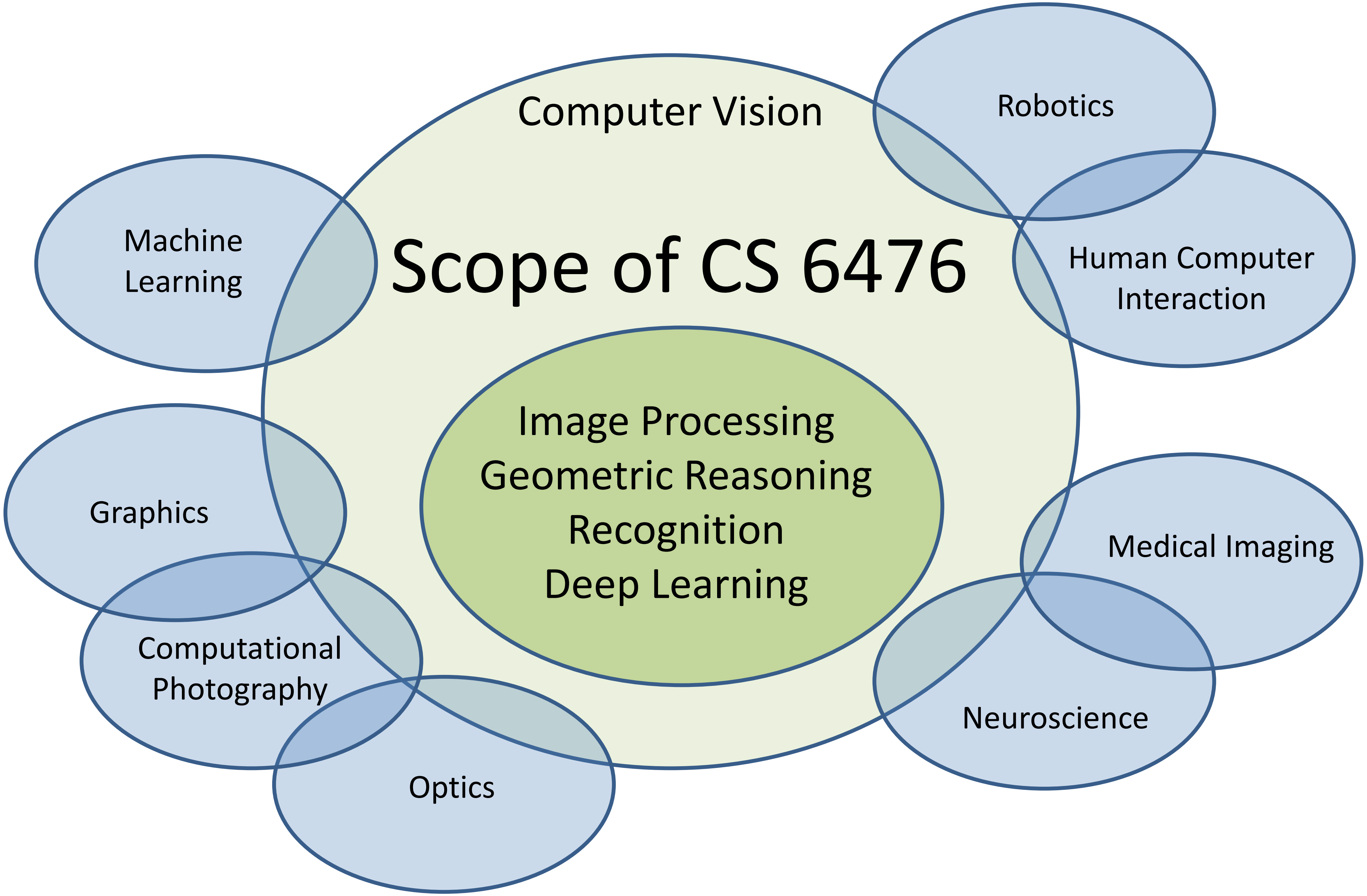
Today’s Class

- ~~Who am I?~~
- ~~What is Computer Vision?~~
- Specifics of this course
- Geometry of Image Formation
- Questions

Grading

- 75% programming projects (5 total + 1 extra credit, maybe)
- 25% 2 quizzes in class

- We will have no final exam. The last project might extend into the final exam period.


Computer Vision
Scope of CS 6476
Image Processing
Geometric Reasoning
Recognition
Deep Learning
Machine Learning
Graphics
Computational Photography
Optics
Robotics
Human Computer Interaction
Medical Imaging
Neuroscience

Textbook

Computer Vision: Algorithms and Applications, 2nd ed.

© 2020 Richard Szeliski, Facebook

http://szeliski.org/Book/

Prerequisites

- **Linear algebra**, basic calculus, and probability
- Experience with image processing will help but is not necessary
- Experience with Python or Python-like languages will help

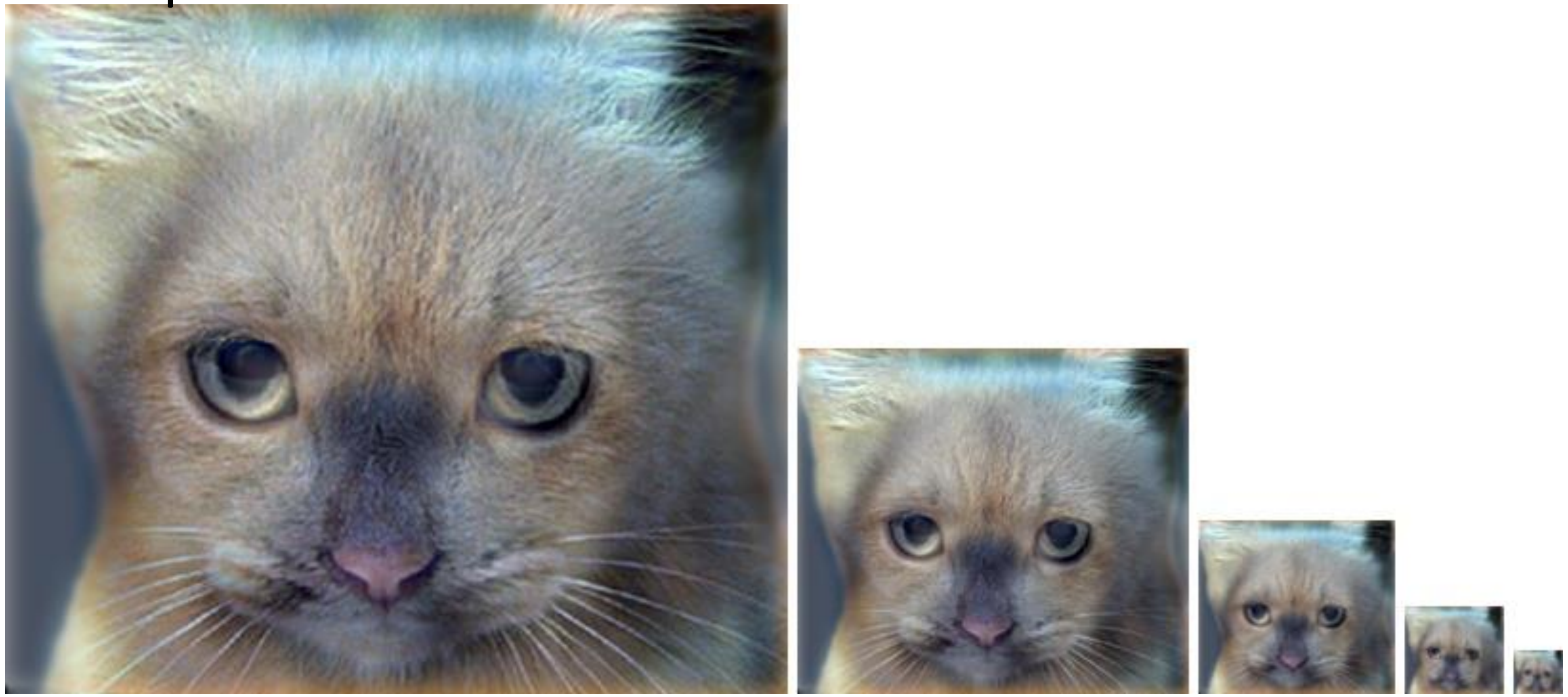
Projects

- (project 0 to test environment setup and handin)
- Image Filtering and Hybrid Images
- Local Feature Matching and Ransac
- Image Classification with Deep Learning
- Semantic Segmentation with Deep Learning
- Point cloud classification with PointNet
- Extra credit project: Neural Radiance Fields (NeRF)

You may want to buy a month or two of Google Colab Pro
near the end of the semester

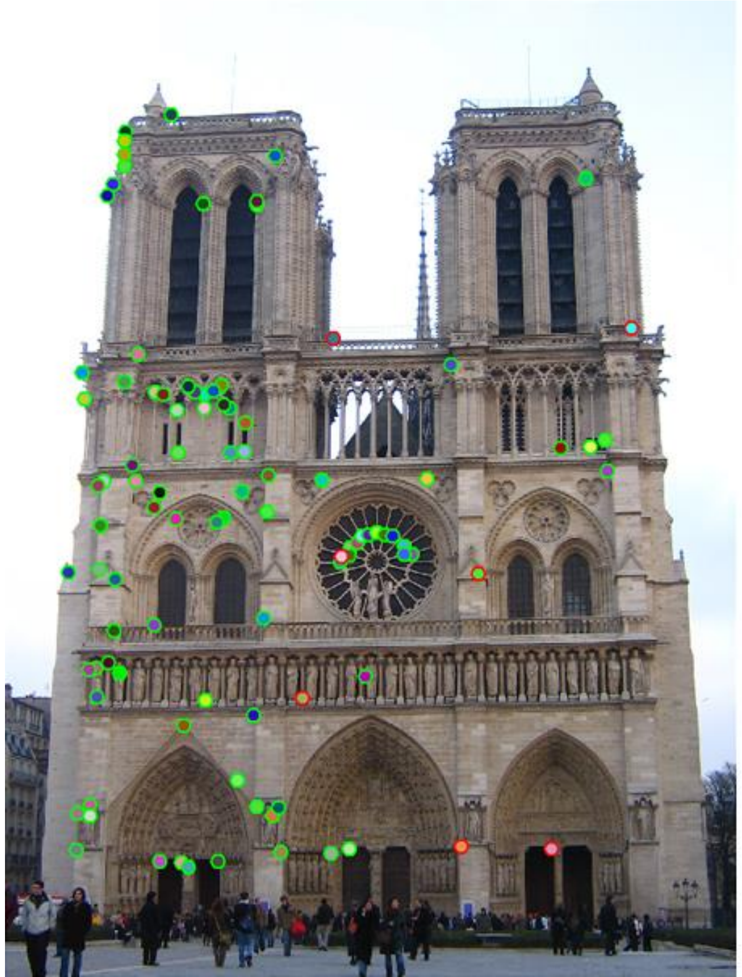
Proj1: Image Filtering and Hybrid Images

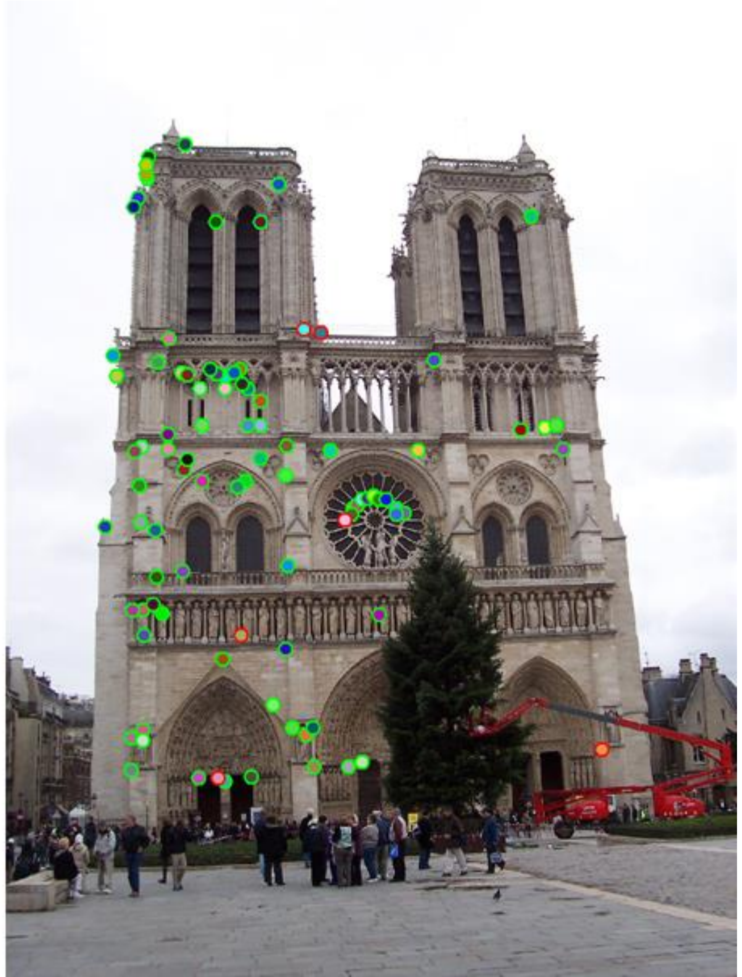
- Implement image filtering to separate high and low frequencies
- Combine high frequencies and low frequencies from different images to create an image with scale-dependent interpretation



Proj2: Local Feature Matching

- Implement interest point detector, SIFT-like local feature descriptor, and simple matching algorithm.





Course Syllabus (tentative)

https://faculty.cc.gatech.edu/~hays/compvision/

Code of Conduct

Your work must be your own. We’ll look for cheating. Don’t talk at the level of code with other students.

Today's Class

- ~~Who am I?~~
- ~~What is Computer Vision?~~
- ~~Specifics of this course~~
- Geometry of Image Formation
- Questions


Cuisinart
WÓDKA
[VODKA]
VISIONS of JAPAN
TOMATSU Shomei
BENTLEY AND HUMPHREYS
SNOW CRYSTALS
DOVER
The Unexplained
MYSTERIES OF MIND SPACE & TIME
FOTOGRAFIE UND DAS UNSICHTBARE
STOWASSER
LATEINISCH-DEUTSCHES
SCHULWÖRTERBUCH
MATHEMATICS
MAN AND SPACE
DRUGS
STACK-ON


WÓDKA
Cuisinart
SNOW CRYSTALS
MATHEMATICS
MAN AND SPACE
DRUGS
STACK-ON

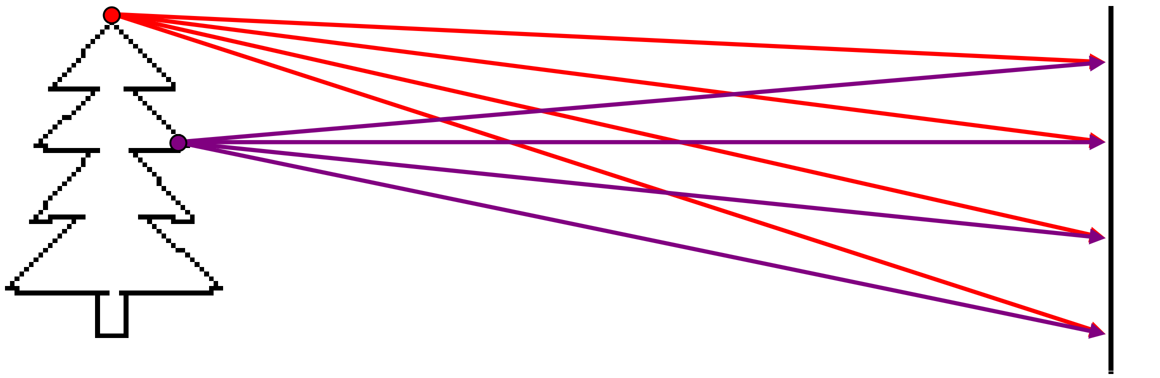
The Geometry of Image Formation

Mapping between image and world coordinates

– Pinhole camera model
– Projective geometry
    - Vanishing points and lines
– Projection matrix

What do you need to make a camera from scratch?

Image formation

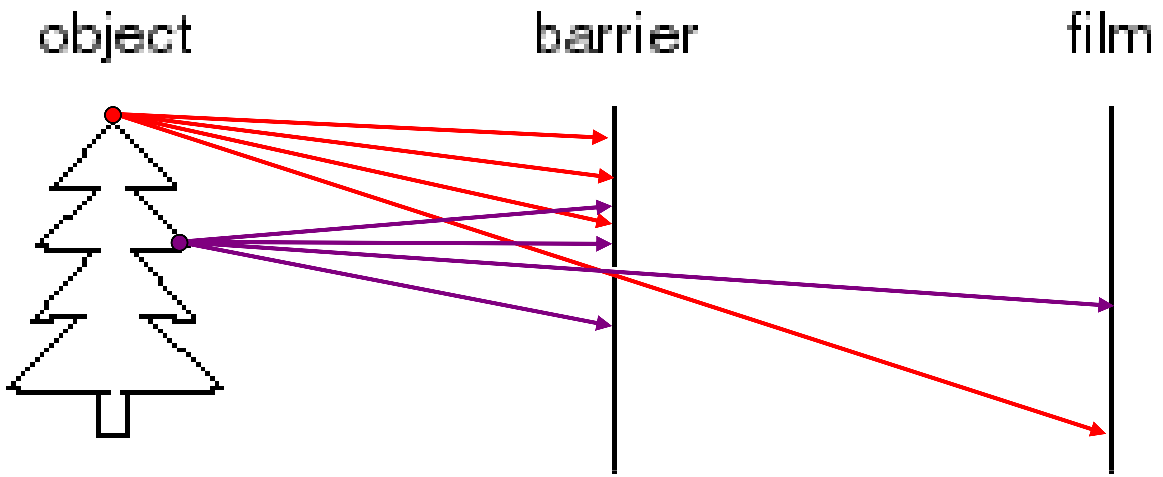


Let’s design a camera

– Idea 1: put a piece of film in front of an object
– Do we get a reasonable image?

Slide source: Seitz

Pinhole camera

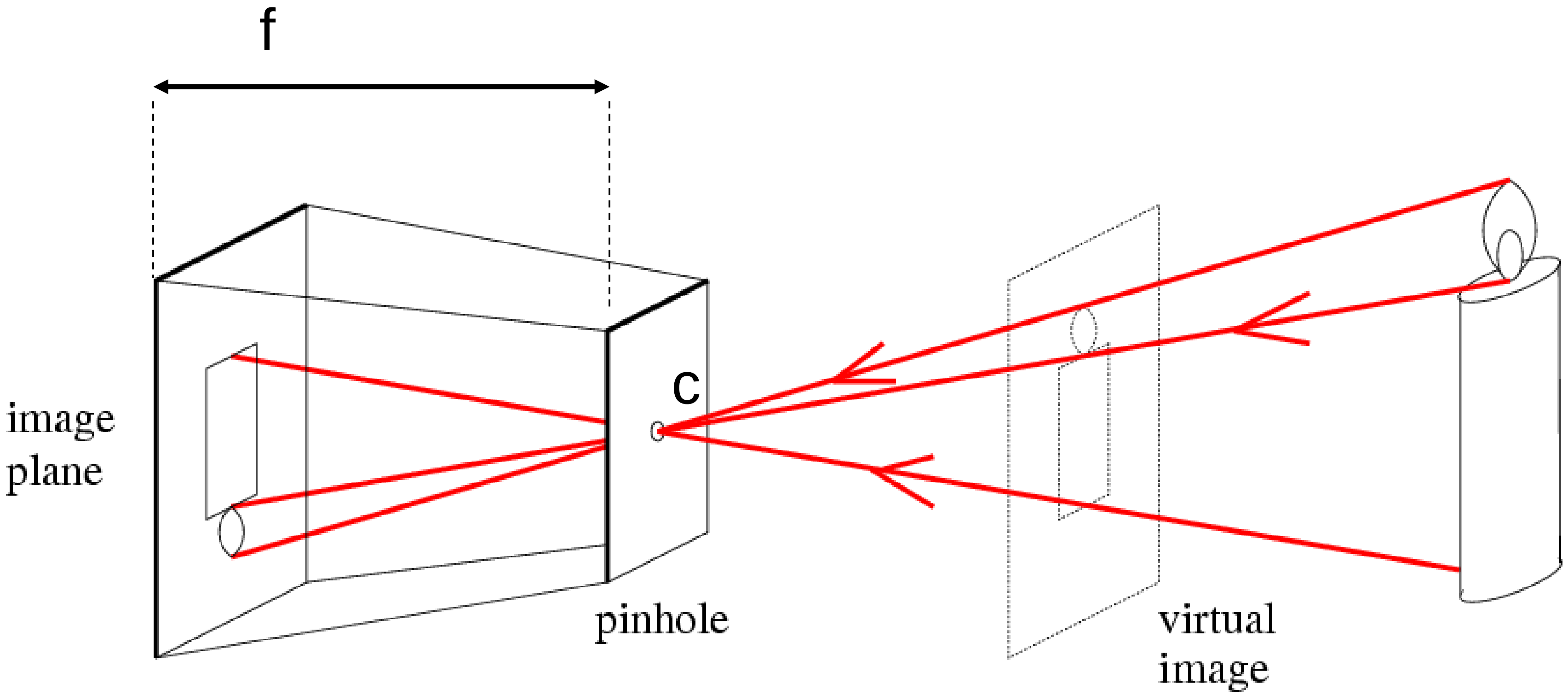


Idea 2: add a barrier to block off most of the rays

- This reduces blurring
- The opening known as the **aperture**

Slide source: Seitz

Pinhole camera



f = focal length
c = center of the camera

Figure from Forsyth

Camera obscura: the pre-camera

- Known during classical period in China and Greece (e.g. Mo-Ti, China, 470BC to 390BC)

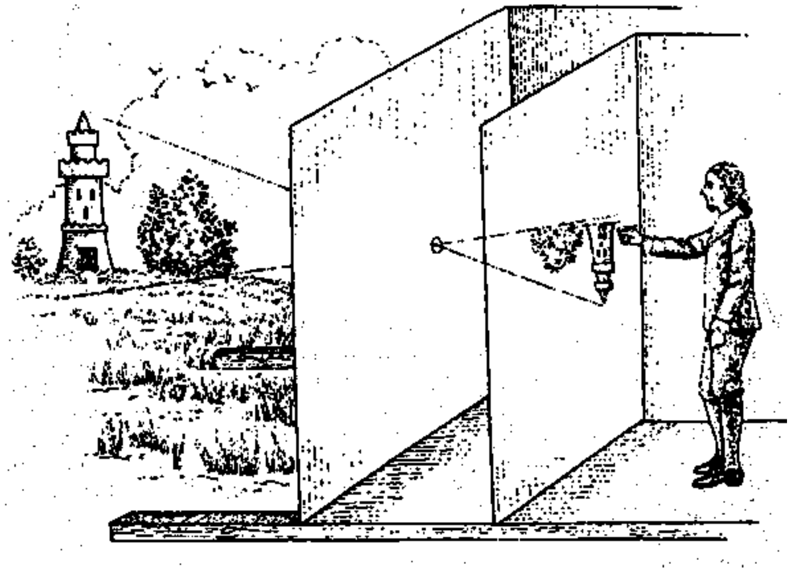
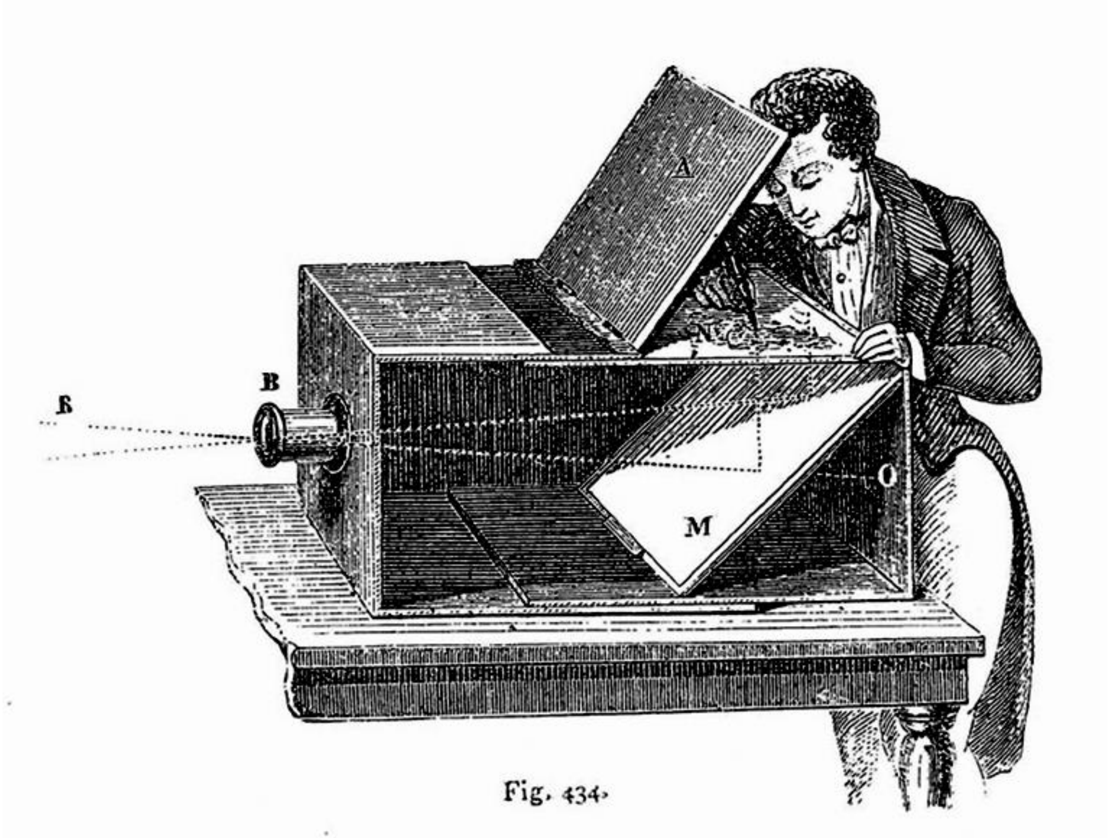


Illustration of Camera Obscura

Freestanding camera obscura at UNC Chapel Hill

Photo by Seth Ilys

Camera Obscura used for Tracing

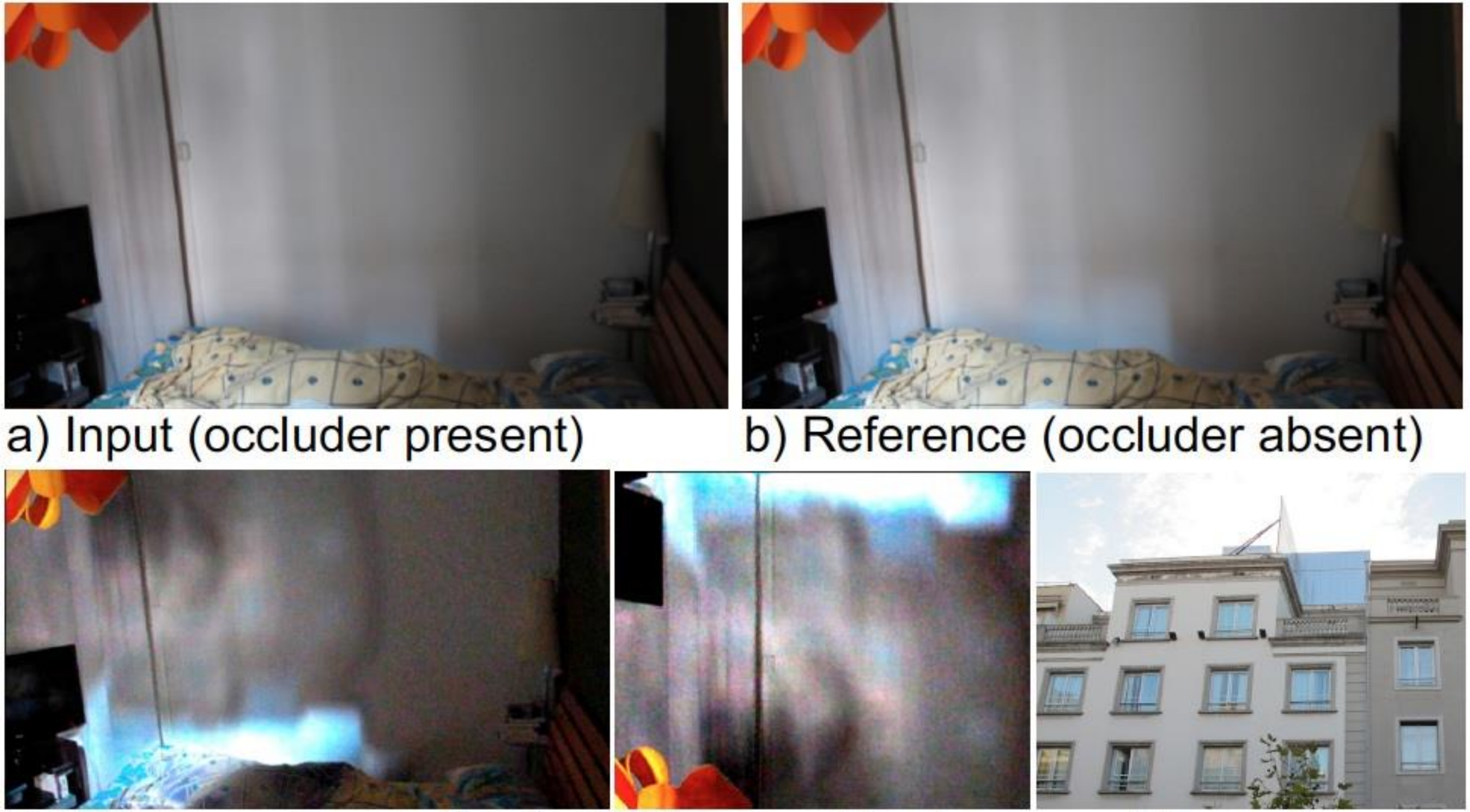


Lens Based Camera Obscura, 1568

Accidental Cameras

Accidental Pinhole and Pinspeck Cameras
Revealing the scene outside the picture.
Antonio Torralba, William T. Freeman

Accidental Cameras



a) Input (occluder present) b) Reference (occluder absent)

c) Difference image (b-a) d) Crop upside down e) True view

First Photograph

Oldest surviving photograph

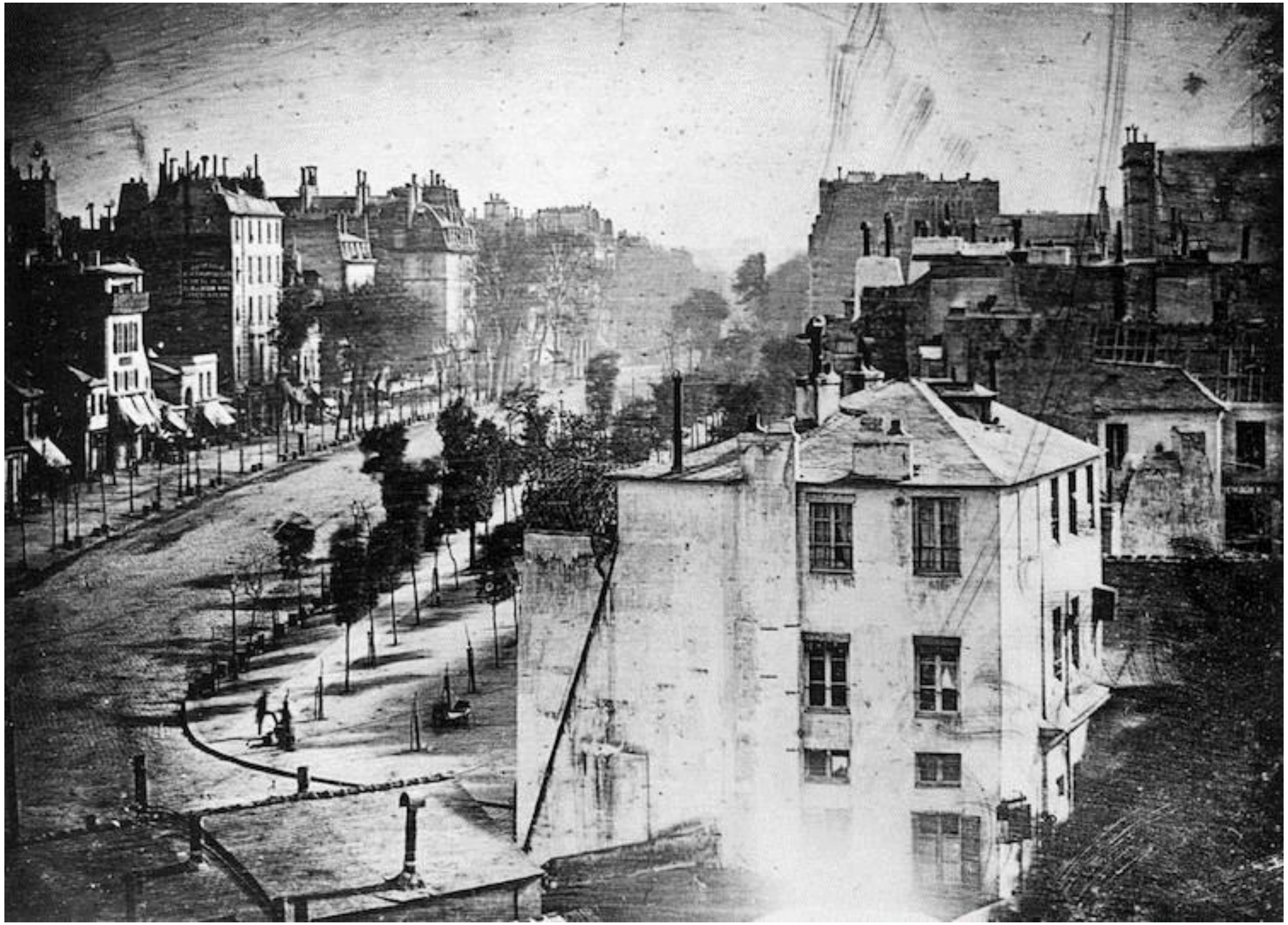
- Took 8 hours on pewter plate

Joseph Niepce, 1826

Photograph of the first photograph

Stored at UT Austin

Niepce later teamed up with Daguerre, who eventually created Daguerrotypes



“Louis Daguerre—the inventor of daguerreotype—shot what is not only the world's oldest photograph of Paris, but also the first photo with humans. The 10-minute long exposure was taken in 1839 in Place de la République and it's just possible to make out two blurry figures in the left-hand corner.”

[Source]

Great history lesson on the chemistry and engineering challenges of early photography from the "Technology Connections" YouTube channel.

https://www.youtube.com/watch?v=wbbH77rYaa8&list=PLv0jwu7G_DFV6yW240e6CbiwCLaZ0Z6PV