95-891: Introduction to Artificial Intelligence

Session 9: Computer Vision

David Steier

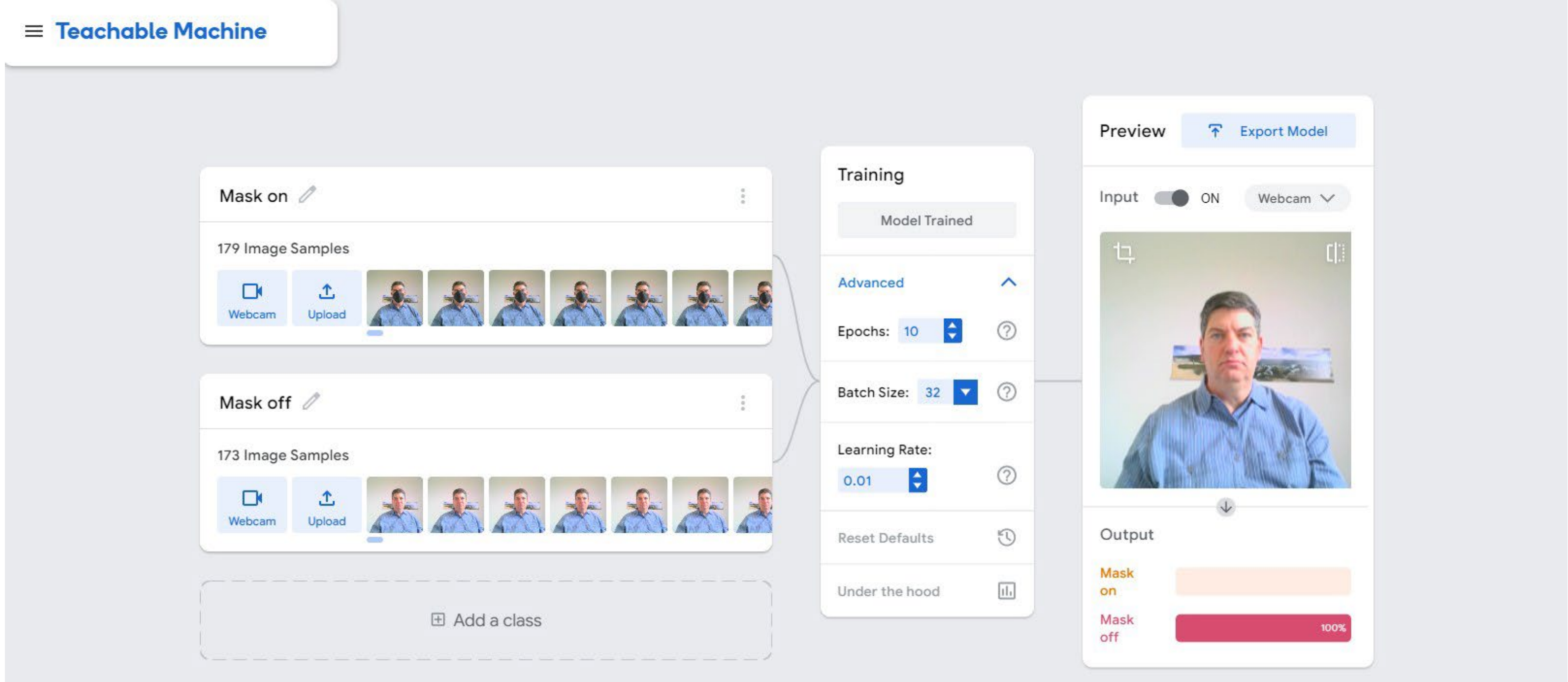
steier@andrew.cmu.edu

September 23, 2025

Agenda

- Exercise: Google’s Teachable Machine
- Introduction to computer vision
- Edge and motion detection
- Object detection
- Object classification
- Facial recognition
- Appendix: Image segmentation

Google’s Teachable Machine



Mask on

179 Image Samples

Mask off

173 Image Samples

Model Trained

Epochs: 10

Batch Size: 32

Learning Rate: 0.01

Input ON

Mask on

Mask off 100%

https://teachablemachine.withgoogle.com/

Training Teachable Machine

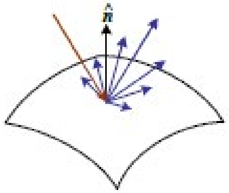
- Divide into groups of two. Make sure at least one of you has a laptop with a camera you can use for this exercise
- Open a web browser and browse to https://teachablemachine.withgoogle.com/
- Click on “Get Started”
- Click on “Image Project”
- Click on “Standard image model”
- Click on the pencil by “Class 1” and edit it so it shows your name
- Click on the pencil by “Class 2” and edit it so it shows your classmate’s name
- Click on the “Webcam” button in the top frame and focus the camera on the first team member. Press “Hold to Record” and record about 5 seconds of video of that person looking at the. camera from various angles. You should have captured about 100 frames from the video.
- Click on the “Webcam” button in the bottom frame and focus the camera on the second team member. Press “Hold to Record” and record about 5 seconds of video of that person looking at the camera from various angles. You should have captured about 100 frames from the video.
- Click “Training” and wait for the data to be prepared and the training to complete (may take about 10 seconds)
- The camera on the right should now be live with the trained model. The model will try to classify whether the image seen by the camera is you or your classmate.

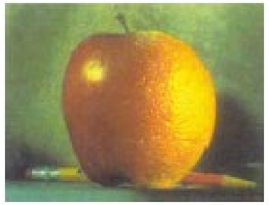
Experimenting With the Model

- How well can the model distinguish between the two members of your team? Try positioning your faces at different camera angles, or obscuring parts of your face to test the model.
- Experiment with the various options under the Advanced dropdown under training, such as Epochs, Batch size and Learning Rate, and retrain the model. Can you observe any effect on training time or accuracy of the model?
- If you have time, add a class representing a third team member, capture video of them, retrain the model and see how well it distinguishes among the three of you.
- Share your findings with the class.

What is Computer Vision?

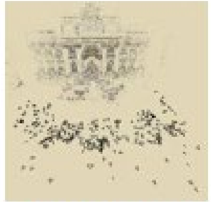
- Perception interprets the results of sensors to obtain information about the world
- Vision studies relationships among
  - Images: 2D
  - Geometry: 3D shape
  - Photometry: Object appearance
- Builds on digital image processing (image → image)

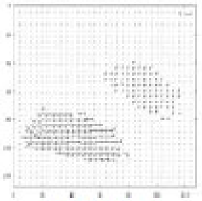

2. Image Formation

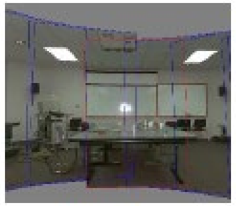

3. Image Processing

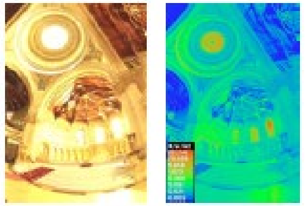
4. Features

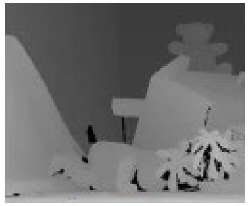
5. Segmentation


6-7. Structure from Motion

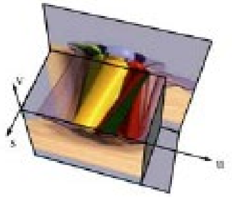

8. Motion


9. Stitching


10. Computational Photography


11. Stereo

12. 3D Shape

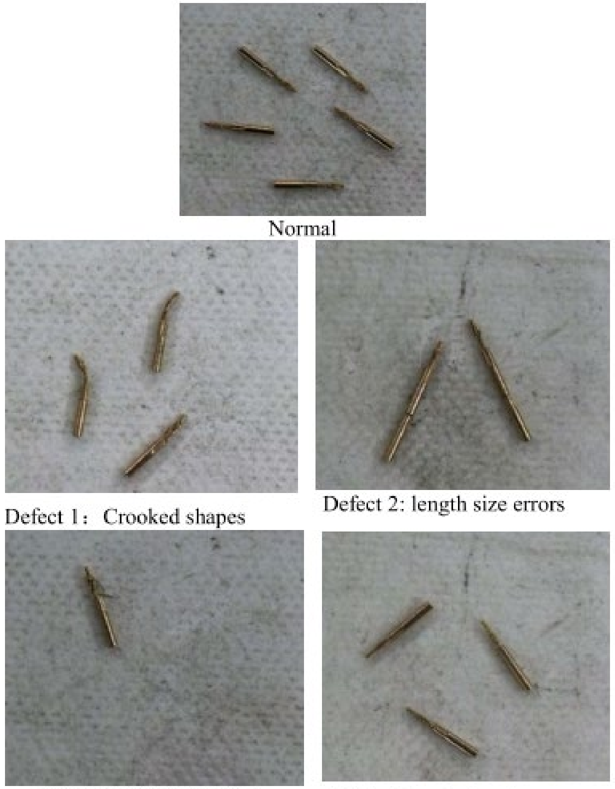

13. Image-based Rendering

14. Recognition

Szeliski, R, *Computer Vision: Algorithms and Applications*, 2010, https://szeliski.org/Book/

Applications of Computer Vision

- Understanding what people are doing
- Linking pictures and words
- Reconstruction from many views
- Geometry from a single view
- Facial recognition
- Inspection of parts for defects



Normal

Defect 1： Crooked shapes

Defect 2: length size errors

Defect 3: Wringing errors

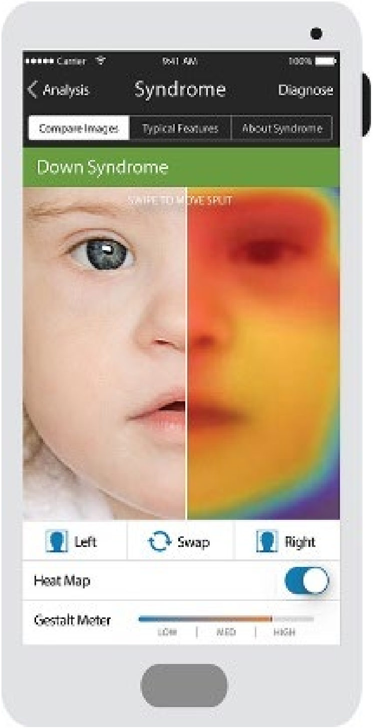
Defect 4: endpoint size errors

Adapted from Russell & Norvig, *AI: A Modern Approach*, 2020, Chapter 25.7, p. 908

Jing, Yang & Li, Shaobo & Wang, Zheng & Guanci, Yang. (2019). Real-Time Tiny Part Defect Detection System in Manufacturing Using Deep Learning. IEEE Access. 7. 89278 - 89291. 10.1109/ACCESS.2019.2925561. , https://www.researchgate.net/publication/334104857_Real-Time_Tiny_Part_Defect_Detection_System_in_Manufacturing_Using_Deep_Learning

Diagnosing Genetic Conditions with Computer Vision

- Down’s Syndrome is one of 10,000 genetic conditions that can be diagnosed from facial features
- Describing such features is very difficult
  - “When you just mention three or four traits, like upslanting eyes, depressed nasal bridge, uplifted nasal tip, or downturned corner of the mouth, that’s not really descriptive. That can be common to a lot of different diseases. How do you start describing patterns? There’s really no way you can verbalize that, you have to be able to classify them as a type of appearance. That’s what your brain is doing.” (Dekel Gelbman, CEO of FDNA, http://www.fdna.com/)
- FDNA’s **Face2Gene** uses computer vision to aid in diagnoses and is available only to clinical geneticists, **70%** of whom now use the app



N. Hurst, January 24, 2017, https://www.smithsonianmag.com/innovation/app-uses-facial-recognition-software-help-identify-genetic-conditions-180961897/#1dYXmJpp66DXvtLb.99

Photometric Image Formation

- Lighting
- Reflectance and shading
- Optics

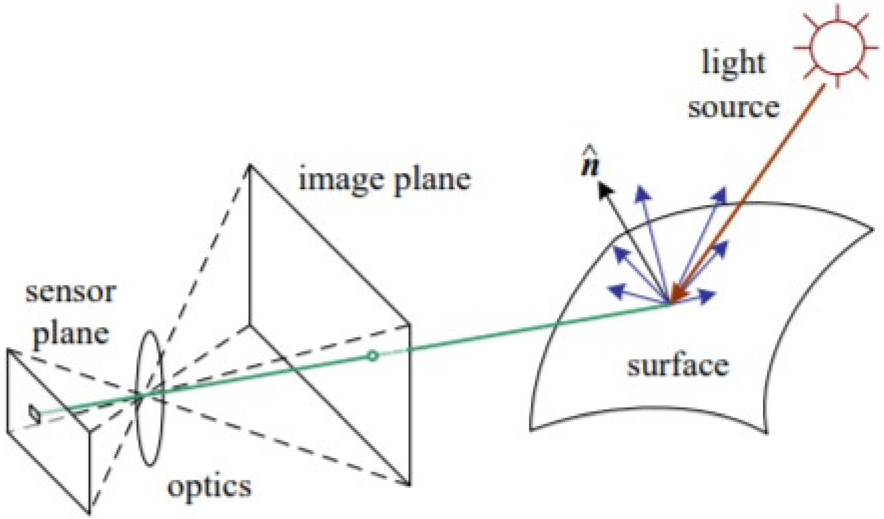
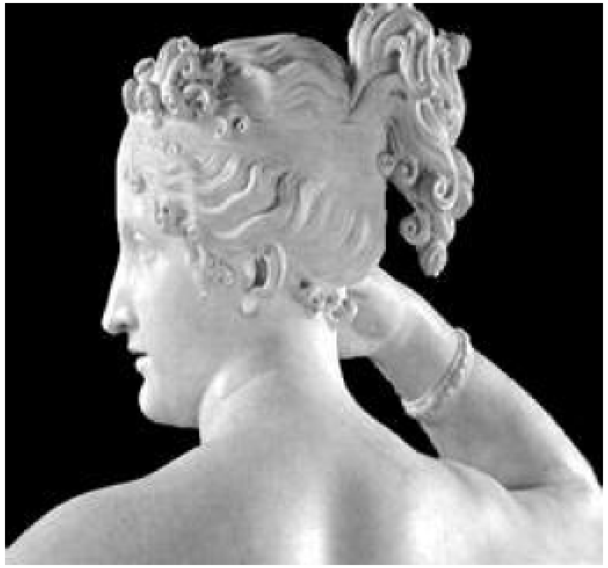


Figure 2.14 A simplified model of photometric image formation. Light is emitted by one or more light sources and is then reflected from an object's surface. A portion of this light is directed towards the camera. This simplified model ignores multiple reflections, which often occur in real-world scenes.

Szeliski, R, *Computer Vision: Algorithms and Applications*, 2010, https://szeliski.org/Book/

Reflectance and Shading



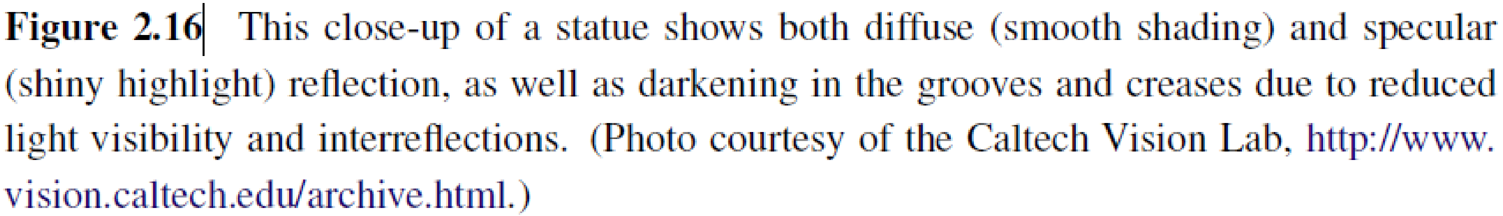


Figure 2.16 This close-up of a statue shows both diffuse (smooth shading) and specular (shiny highlight) reflection, as well as darkening in the grooves and creases due to reduced light visibility and interreflections. (Photo courtesy of the Caltech Vision Lab, http://www.vision.caltech.edu/archive.html.)

Szeliski, R, *Computer Vision: Algorithms and Applications*, 2010, https://szeliski.org/Book/

Edge Detection from Gradients

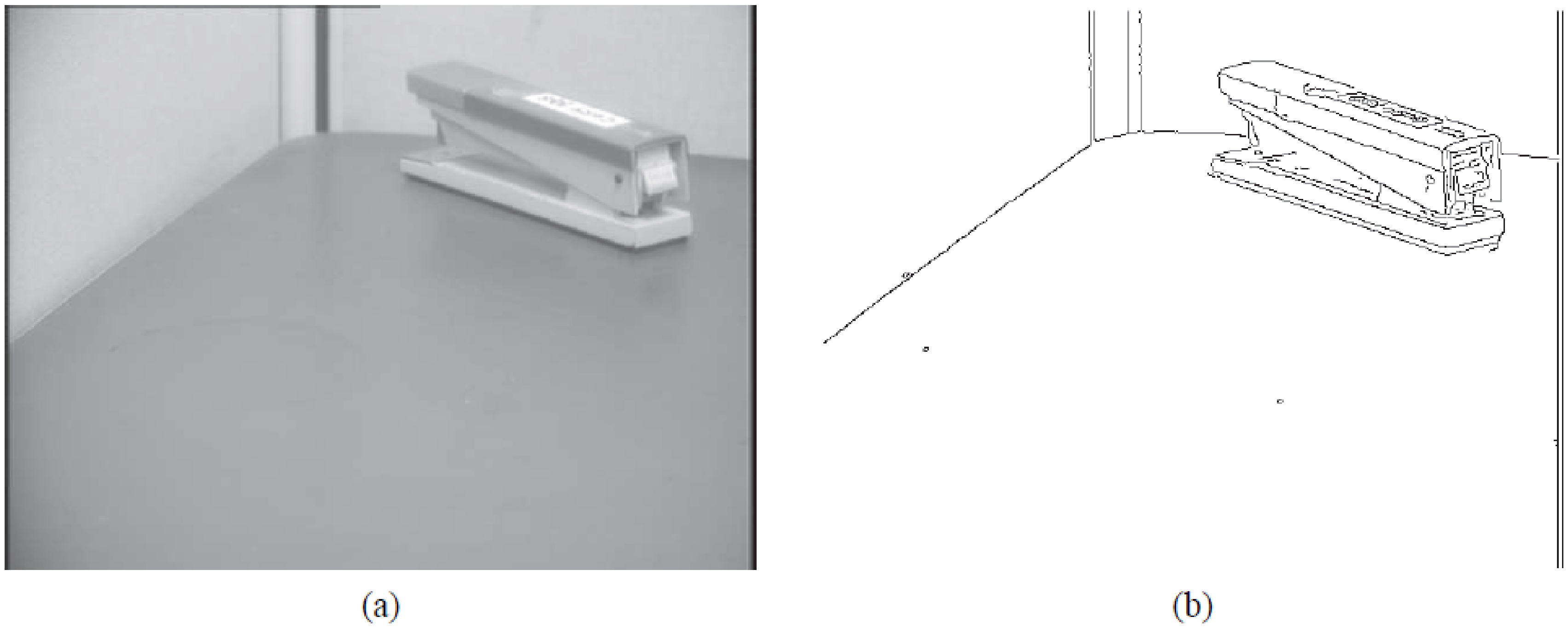

(a)
(b)

Figure 24.7 FILES: figures/stapler1-test.eps (Tue Nov 3 16:23:47 2009) figures/stapler1.edge-test.eps (Tue Nov 3 16:23:47 2009). (a) Photograph of a stapler. (b) Edges computed from (a).

Russell & Norvig, 2020, Chapter 25, p. 892

Derivatives of Intensity For Edge Detection

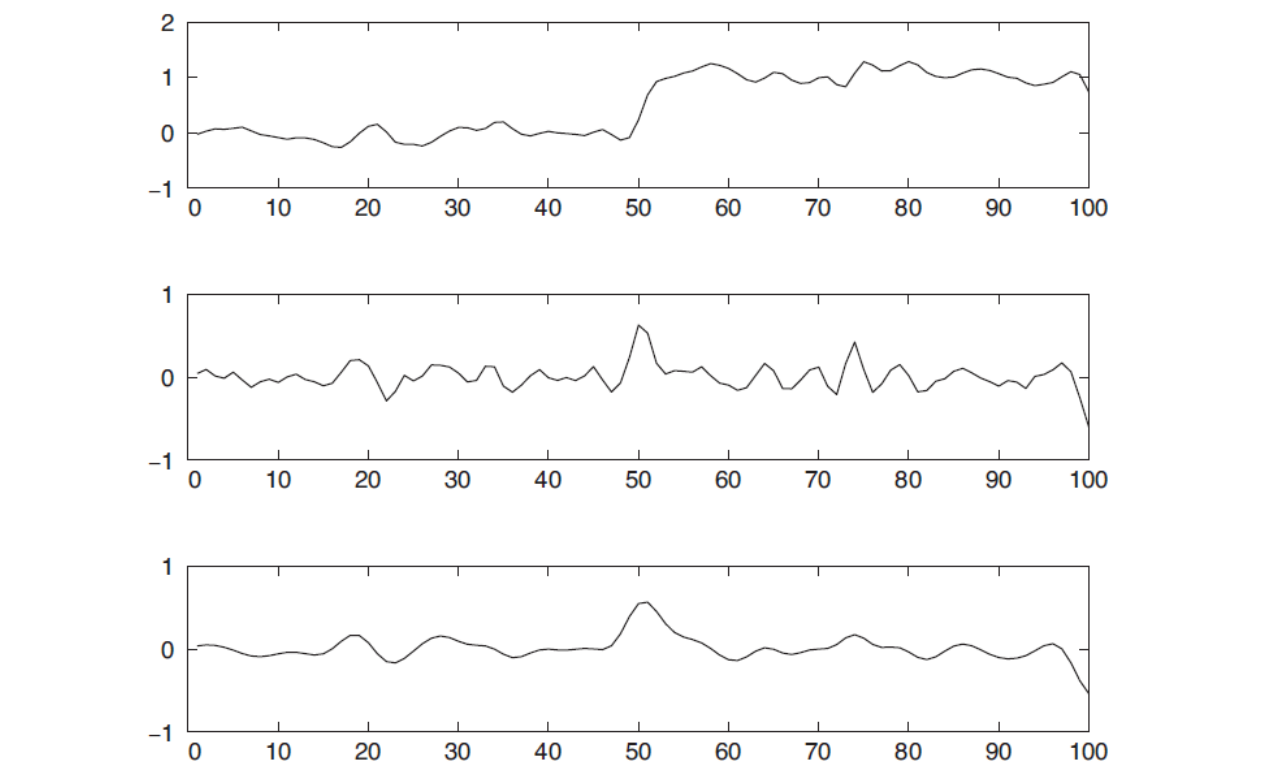


Figure 24.8 **FILES: figures/edgewderiv.eps (Tue Nov 3 16:22:41 2009).** Top: Intensity profile $I(x)$ along a one-dimensional section across an edge at $x = 50$. Middle: The derivative of intensity, $I'(x)$. Large values of this function correspond to edges, but the function is noisy. Bottom: The derivative of a smoothed version of the intensity, $(I * G_\sigma)'$, which can be computed in one step as the convolution $I * G'_\sigma$. The noisy candidate edge at $x = 75$ has disappeared.

Russell & Norvig, 2020, Chapter 25

Optical Flow for Motion Detection

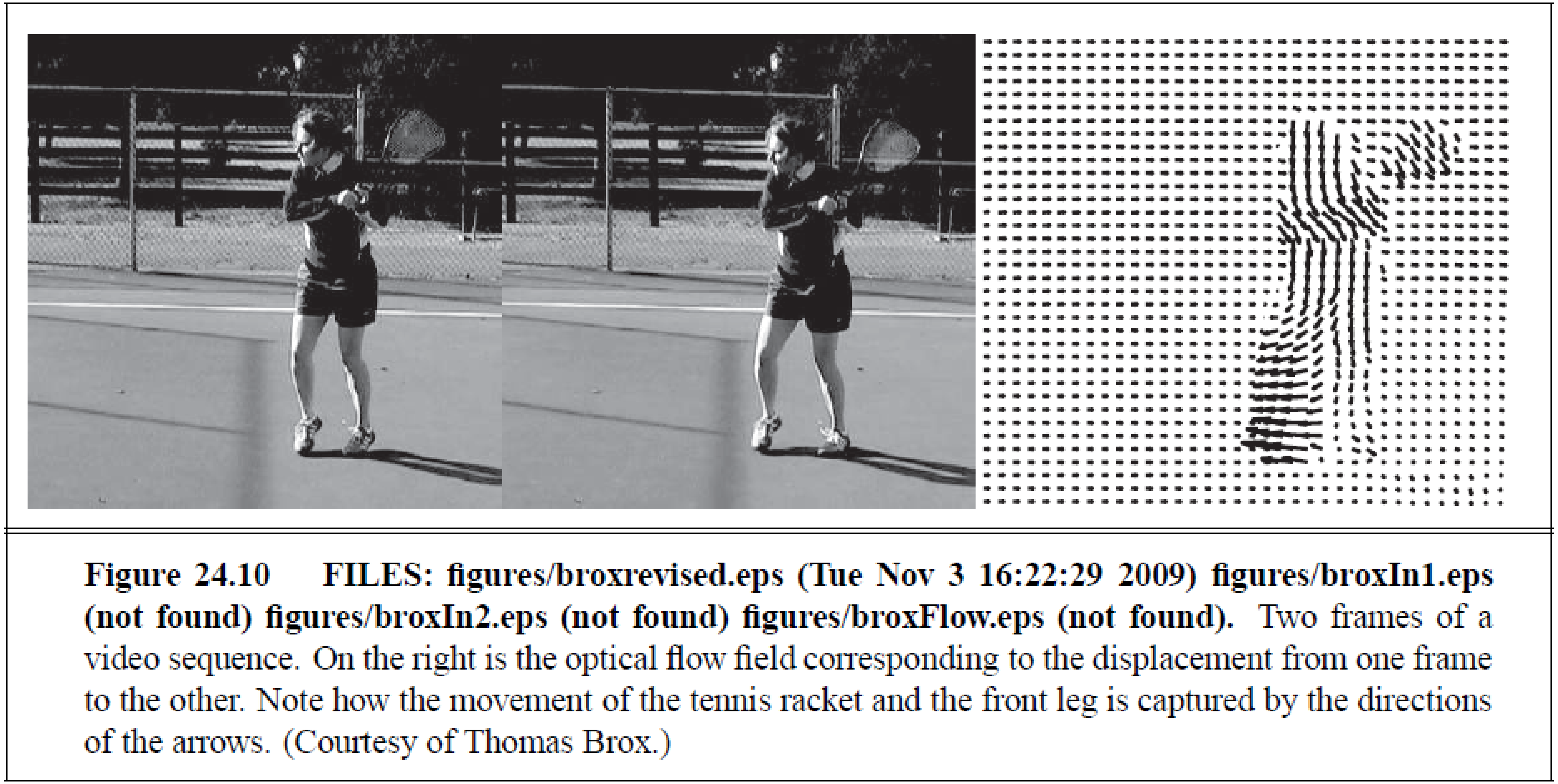


Figure 24.10 FILES: figures/broxrevised.eps (Tue Nov 3 16:22:29 2009) figures/broxIn1.eps (not found) figures/broxIn2.eps (not found) figures/broxFlow.eps (not found). Two frames of a video sequence. On the right is the optical flow field corresponding to the displacement from one frame to the other. Note how the movement of the tennis racket and the front leg is captured by the directions of the arrows. (Courtesy of Thomas Brox.)

Russell & Norvig, 2020, Chapter 25.3, p. 893

Object Detection

- Object detection finds multiple objects of known class in an image
- Need to pinpoint where object is in an image by drawing **bounding boxes**
- Use small sliding windows as potential bounding boxes
  - Decide on a window (often rectangles)
  - A classifier (CNN) decides what is in each window
  - Decide which windows to look at – the **Regions of Interest**
  - Choose which windows to report (depends on overlaps and application)
  - Report locations of objects using these windows

Russell & Norvig, 2020, p.899

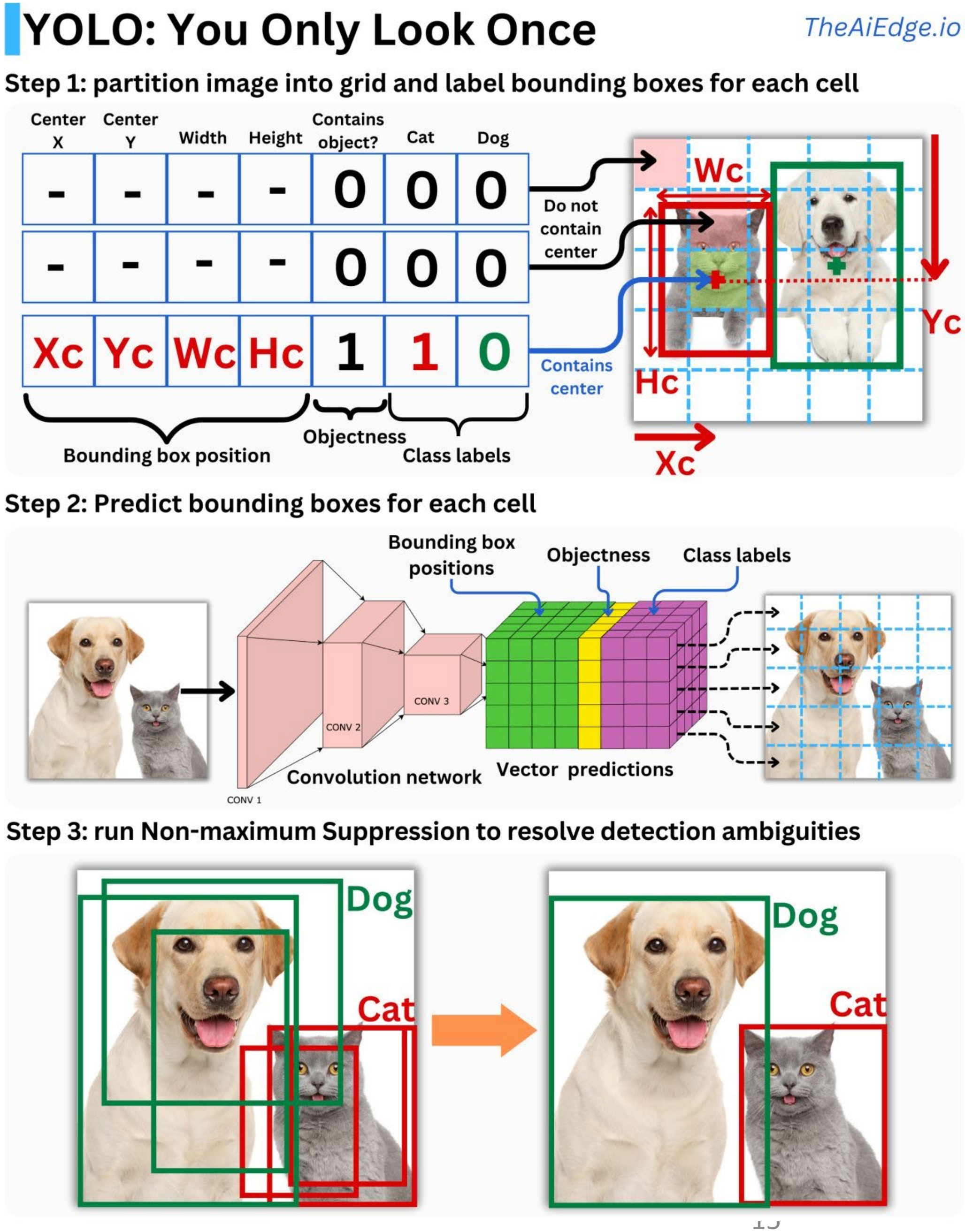
YOLO – You Only Look Once

1) Resize the image to a single size and partition into grid of cells

2) Predict bounding boxes for objects centered in each cell using a CNN

3) Filter out all bounding boxes except those with maximum confidence

J. Redmon, et. al," You Only Look Once: Unified, Real-Time Object Detection, 9 May 2016, http://pjreddie.com/yolo/



YOLO: You Only Look Once

TheAiEdge.io

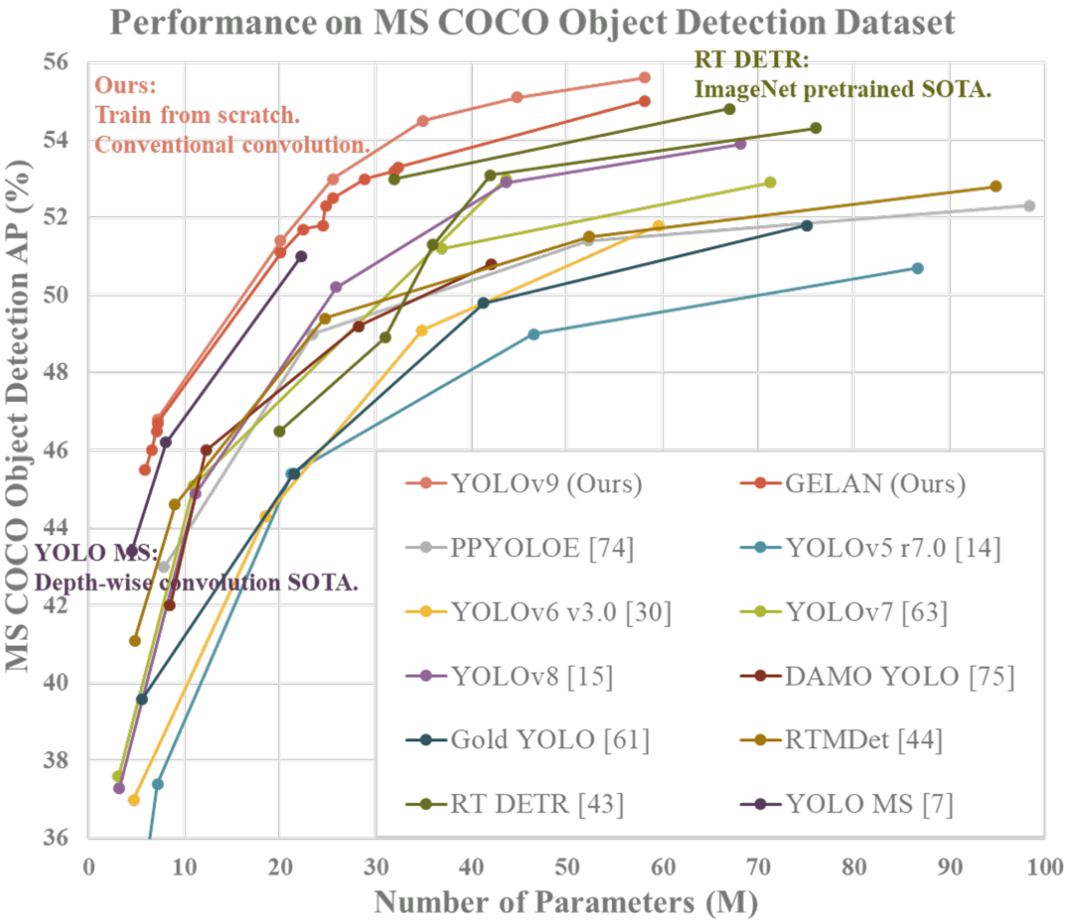
Step 1: partition image into grid and label bounding boxes for each cell

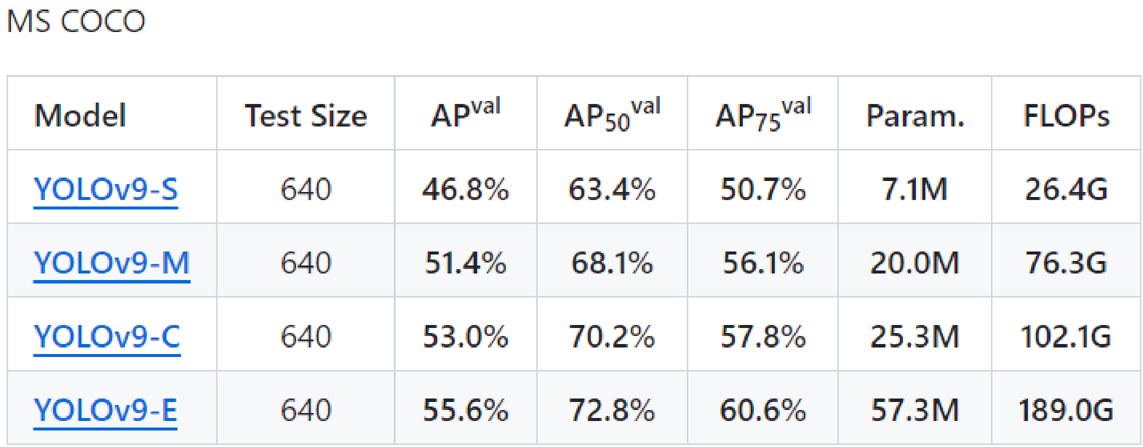
Center X	Center Y	Width	Height	Contains object?	Cat	Dog
-	-	-	-	0	0	0
-	-	-	-	0	0	0
Xc	Yc	Wc	Hc	1	1	0

Step 2: Predict bounding boxes for each cell

Step 3: run Non-maximum Suppression to resolve detection ambiguities

Evolution of YOLO to YOLOv9



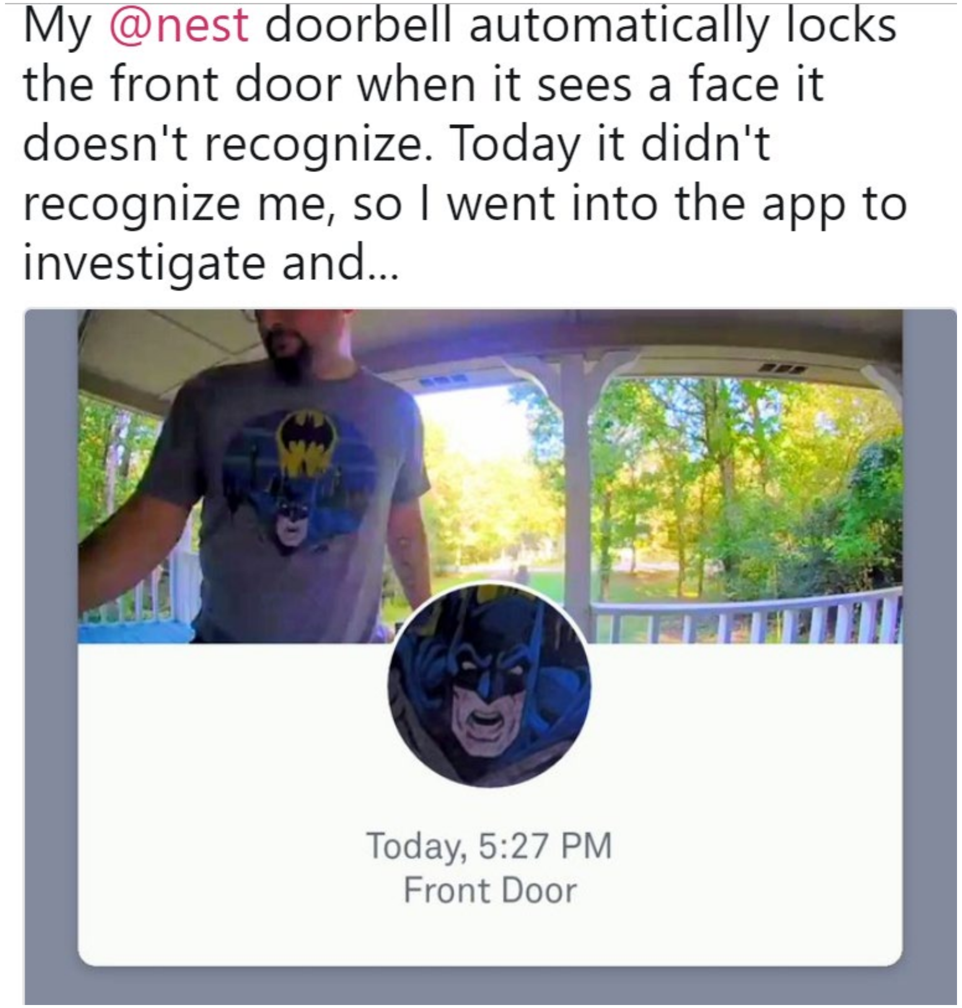


MS COCO

Model	Test Size	AP^{val}	AP_{50}^{val}	AP_{75}^{val}	Param.	FLOPs
YOLOv9-S	640	46.8%	63.4%	50.7%	7.1M	26.4G
YOLOv9-M	640	51.4%	68.1%	56.1%	20.0M	76.3G
YOLOv9-C	640	53.0%	70.2%	57.8%	25.3M	102.1G
YOLOv9-E	640	55.6%	72.8%	60.6%	57.3M	189.0G

https://github.com/WongKinYiu/yolov9?tab=readme-ov-file , last updated Feb 26, 2024

The Importance of Context in Face Detection



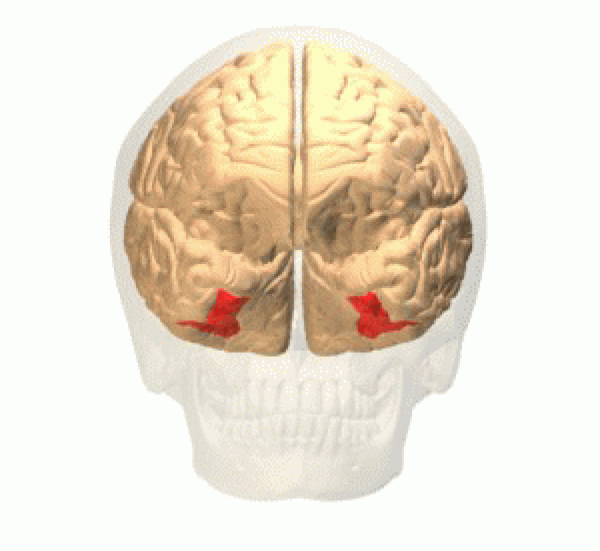
My @nest doorbell automatically locks the front door when it sees a face it doesn't recognize. Today it didn't recognize me, so I went into the app to investigate and…

Today, 5:27 PM
Front Door

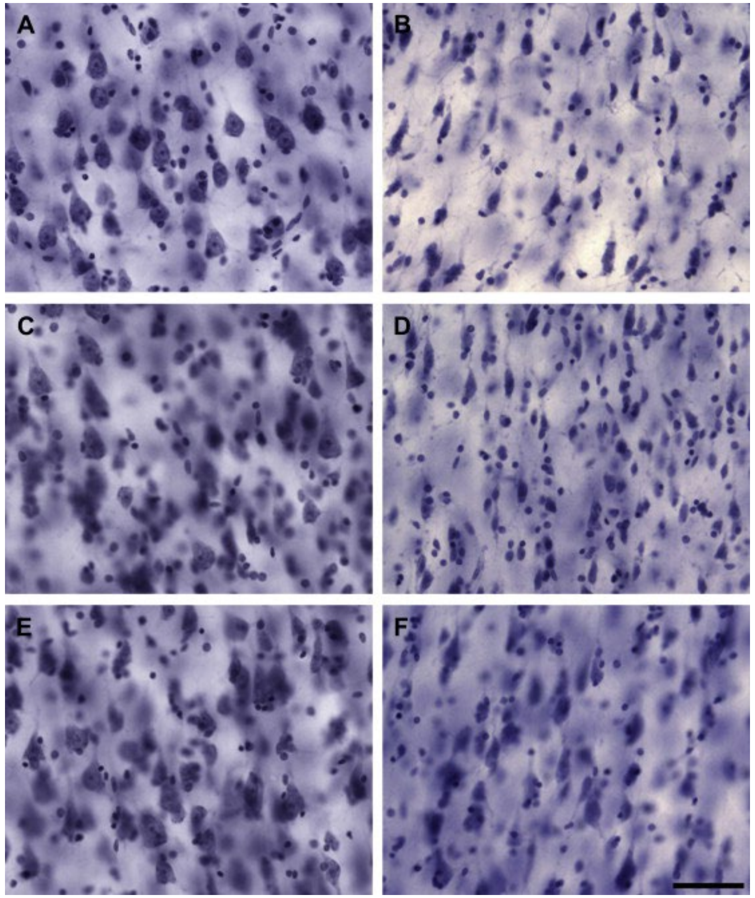
https://twitter.com/bjmay/status/1041833853852311553?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1041833853852311553&ref_url=https%3A%2F%2Fwww.independent.ie%2Fworld-news%2Fand-finally%2Fthis-man-was-locked-out-of-home-when-his-smart-doorbell-thought-he-was-batman-37329890.html

How Do Humans Recognize Faces?

- The **fusiform gyrus** is an area of the brain that responds specifically to faces and not other objects
- Humans recognize faces in ways we have not been able to replicate in AI

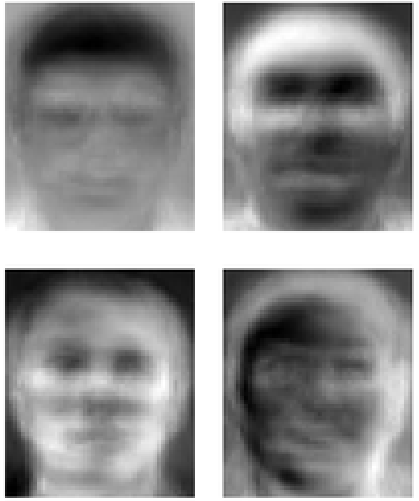


https://en.wikipedia.org/wiki/Fusiform_gyrus



Discrete Cortical Neuropathology in Autism Spectrum Disorders
Neha Uppal, Patrick R. Hof, in The Neuroscience of Autism Spectrum Disorders, 2013

Facial Recognition Using Computers



- 1960s: Recognition based on human-annotated features
- 1970: Recognition without human annotation (CMU!)
- Early 1990s: Eigenfaces: Recognition based on PCA-based features
- Late 1990s: First commercial recognition systems, based on skin segmentation
- 2000s: Real time face detection in videos, using AdaBoost
- 2010s: Use of thermal imagery for facial recognition
- 2018: Skynet in China incorporates imagery from 20 million cameras, matches against full Chinese population in 1 second, world population in 2 seconds
- 2022: Ukraine uses US-based Clearview facial recognition to identify 582 dead Russian soldiers

https://en.wikipedia.org/wiki/Facial_recognition_system

Steps in Facial Recognition

A Geitgey, “Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning,” July 24, 2016, https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

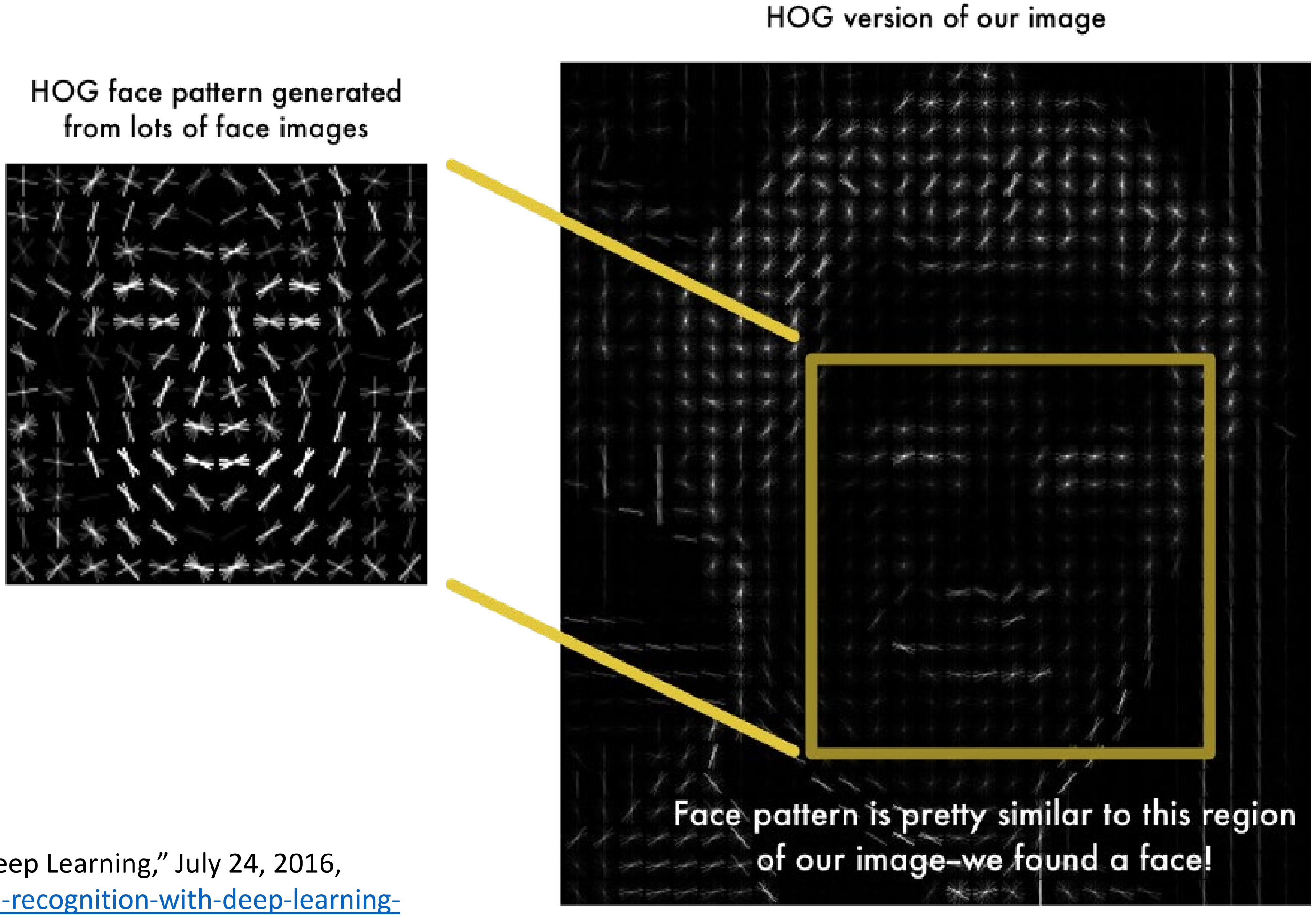
Facial Recognition Using OpenFace

1. Find the faces
   - Histogram of Oriented Gradients
2. Orient the face
   - Face landmark estimation
3. Encode the face
   - Pre-trained CNN from OpenFace
4. Identify the face
   - SVM on a database of encoded faces

A Geitgey, "Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning," July 24, 2016, https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

Histogram of Oriented Gradients (HOG)

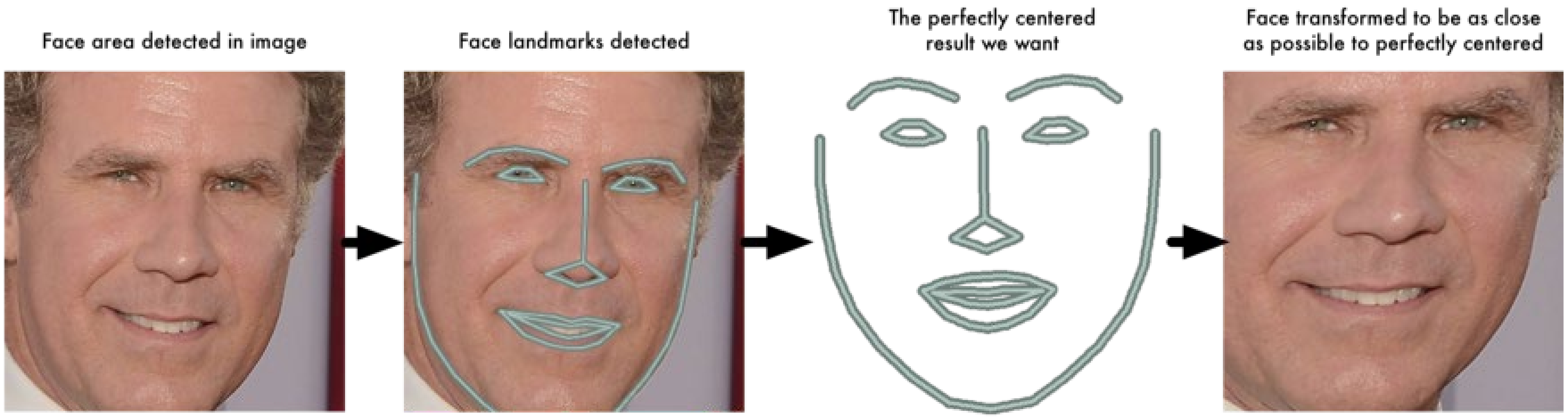
- Detecting faces pixel by pixel takes too long and gets confused by images of different brightness
- Measure changes in brightness over 16x16 pixel regions
- Arrows show direction of changes in brightness


HOG version of our image
HOG face pattern generated from lots of face images
Face pattern is pretty similar to this region of our image–we found a face!

A Geitgey, “Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning,” July 24, 2016, https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

Frontalization: Posing and Orienting the Face

- Find 68 points that every face has (top of the chin, outside of eyes)
- Transform the image so eyes and mouth are centered

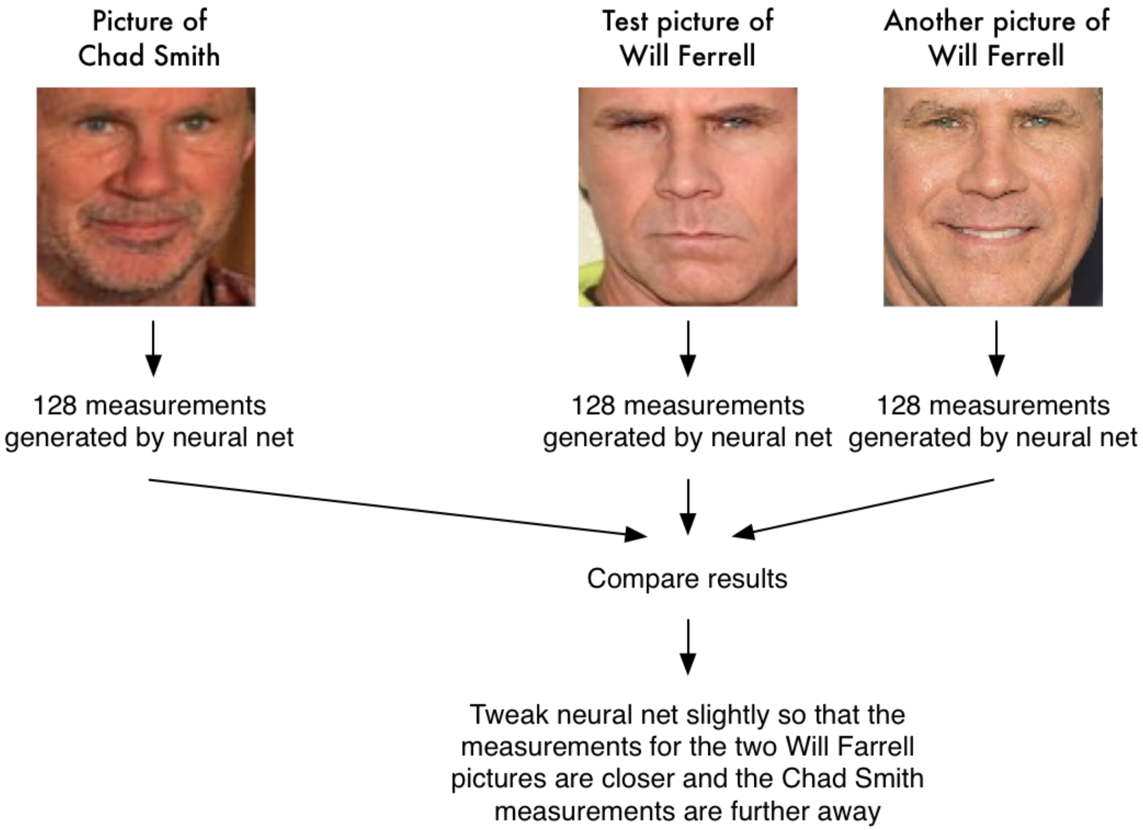

Face area detected in image
Face landmarks detected
The perfectly centered result we want
Face transformed to be as close as possible to perfectly centered

A Geitgey, “Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning,” July 24, 2016, https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

Encoding the Face

- Use CNN to compute face embeddings as 128-element vector
- Need to do this millions of times for millions of images
- Build on pretrained networks such as OpenFace (CMU!) https://github.com/cmusatyalab/openface

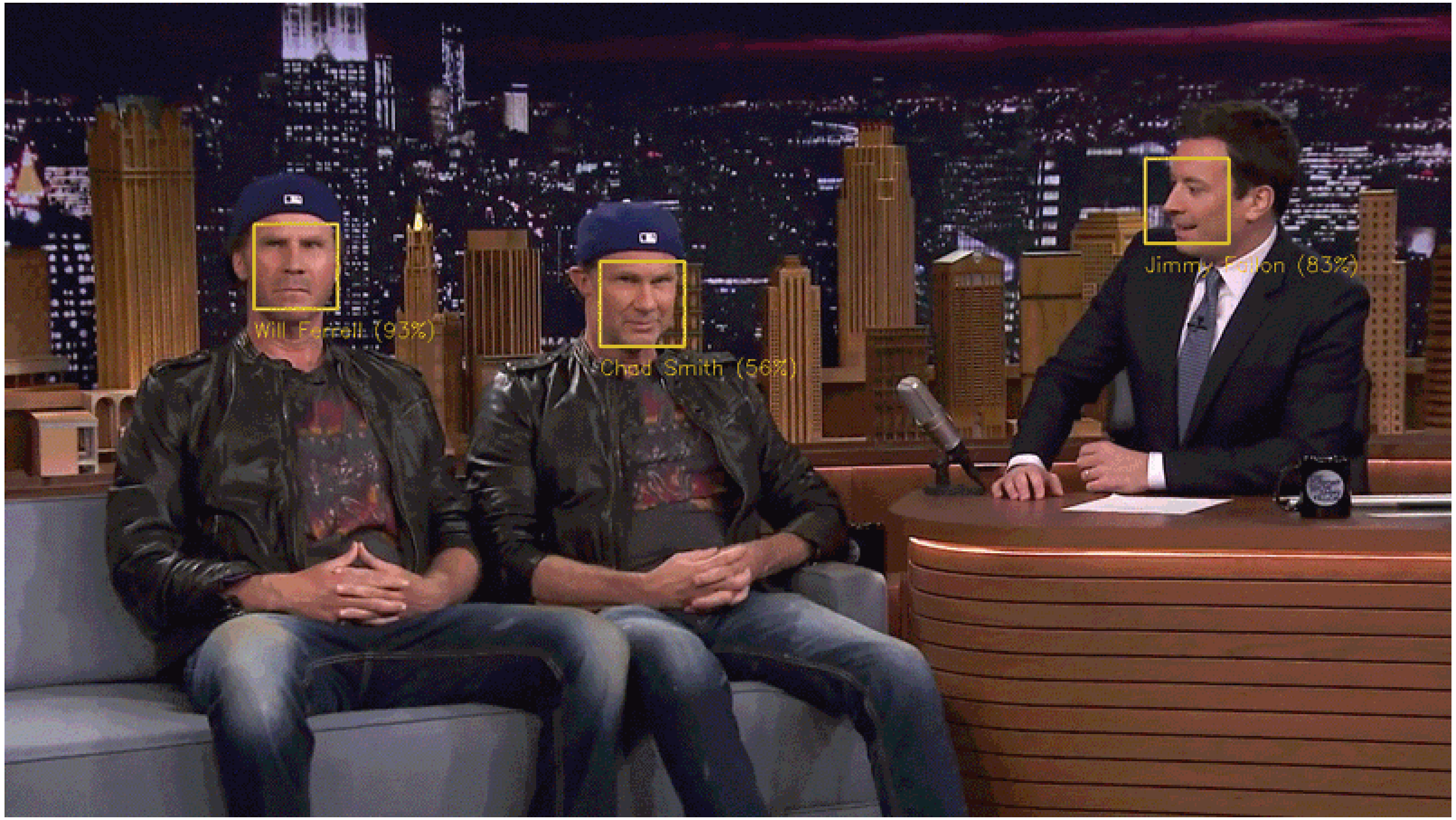
A single 'triplet' training step:


Picture of Chad Smith
Test picture of Will Ferrell
Another picture of Will Ferrell
128 measurements generated by neural net
128 measurements generated by neural net
128 measurements generated by neural net
Compare results

Tweak neural net slightly so that the measurements for the two Will Farrell pictures are closer and the Chad Smith measurements are further away

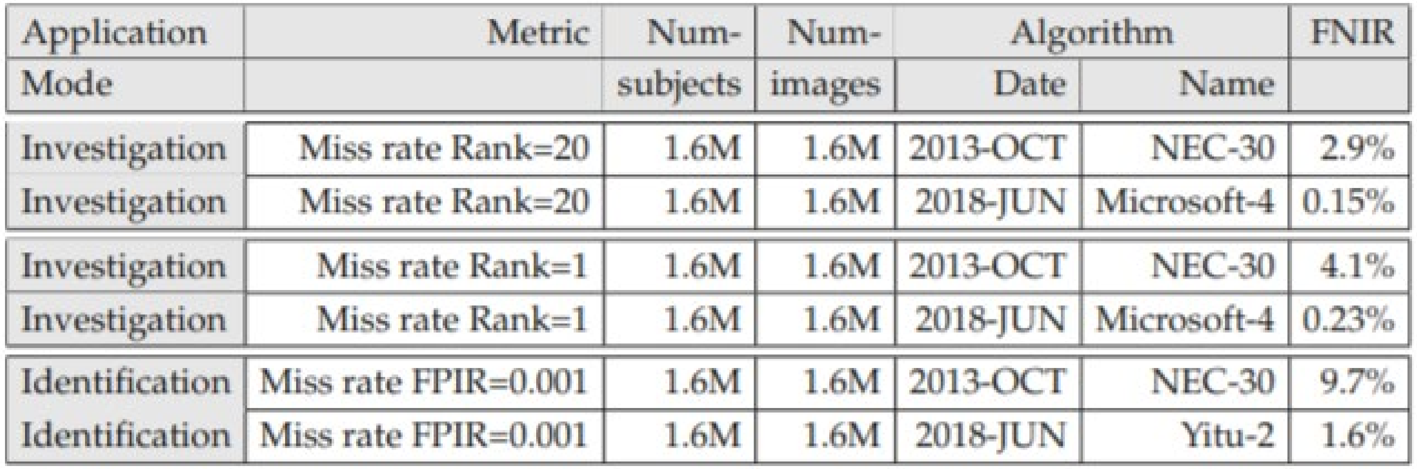
A Geitgey, “Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning,” July 24, 2016, https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

Identify the Face(s) with SVM Classifier



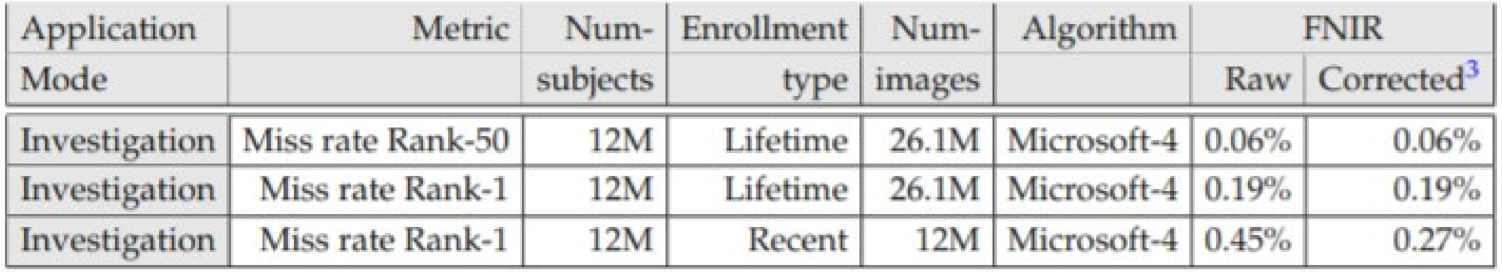
A Geitgey, “Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning,” July 24, 2016, https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

Facial Recognition Errors Have Decreased 20x in 5 years



Application	Metric	Num-	Num-	Algorithm		FNIR
Mode		subjects	images	Date	Name	
Investigation	Miss rate Rank=20	1.6M	1.6M	2013-OCT	NEC-30	2.9%
Investigation	Miss rate Rank=20	1.6M	1.6M	2018-JUN	Microsoft-4	0.15%
Investigation	Miss rate Rank=1	1.6M	1.6M	2013-OCT	NEC-30	4.1%
Investigation	Miss rate Rank=1	1.6M	1.6M	2018-JUN	Microsoft-4	0.23%
Identification	Miss rate FPIR=0.001	1.6M	1.6M	2013-OCT	NEC-30	9.7%
Identification	Miss rate FPIR=0.001	1.6M	1.6M	2018-JUN	Yitu-2	1.6%

Table 1: Accuracy gains since 2013.



Application	Metric	Num-	Enrollment	Num-	Algorithm	FNIR	
Mode		subjects	type	images		Raw	Corrected[3]
Investigation	Miss rate Rank-50	12M	Lifetime	26.1M	Microsoft-4	0.06%	0.06%
Investigation	Miss rate Rank-1	12M	Lifetime	26.1M	Microsoft-4	0.19%	0.19%
Investigation	Miss rate Rank-1	12M	Recent	12M	Microsoft-4	0.45%	0.27%

Table 2: Absolute accuracy 2018.

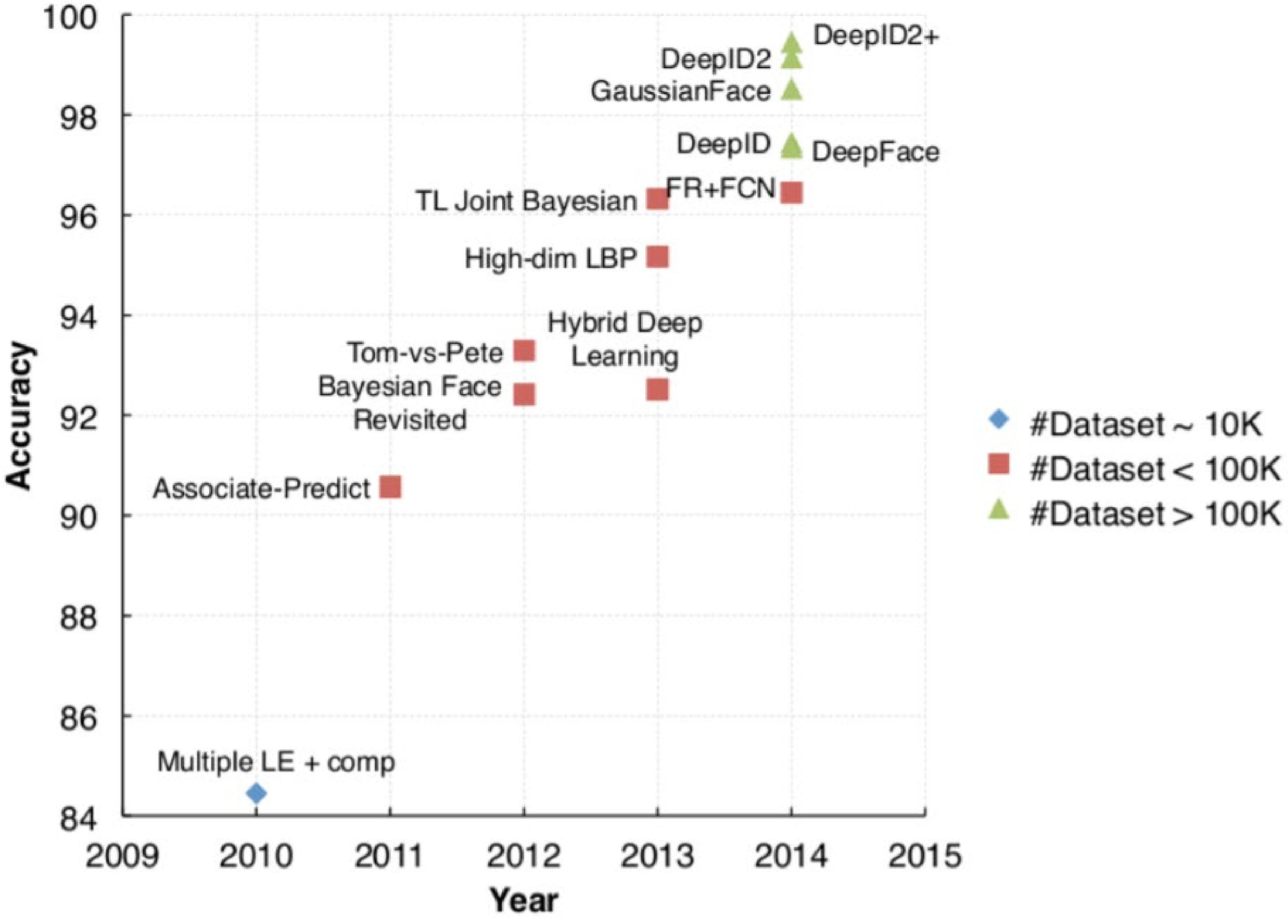
Grother, P. , Ngan, M. , Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification", November 2018, https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8238.pdf

Accuracy of Facial Recognition and Training Set Sizes

- Google's FaceNet trained on over 260 million images
- Clearview trained on 3 billion images!

https://www.cnn.com/2020/02/10/tech/clearview-ai-ceo-hoan-ton-that/index.html

http://cs.wellesley.edu/~vision/slides/Qianli_summary_deep_face_models.pdf



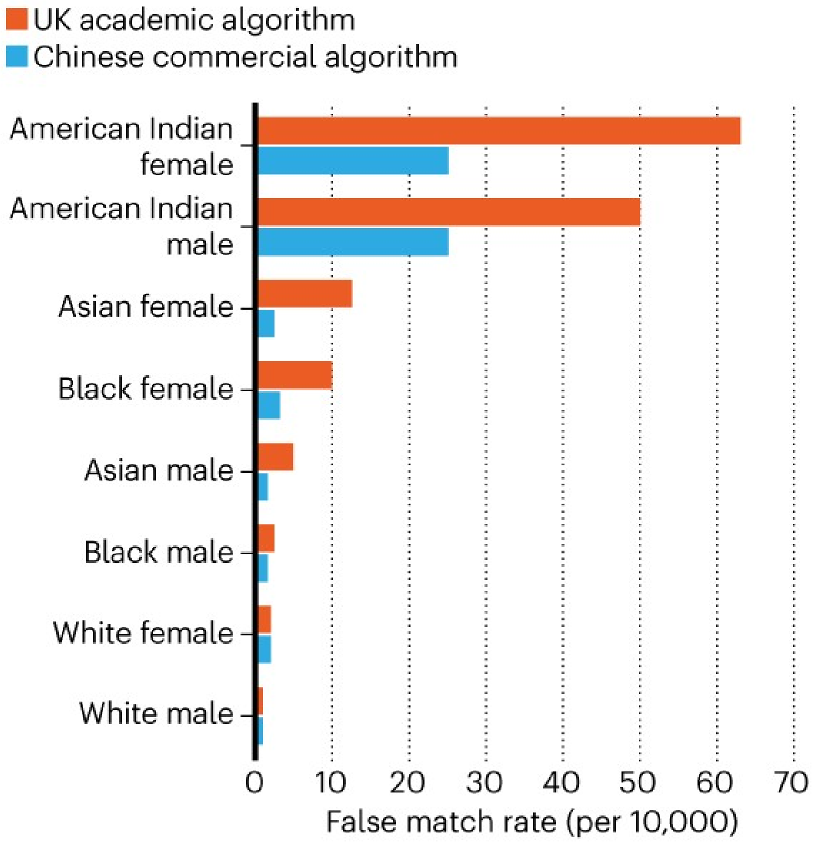
Bias in Facial Recognition

- African-American or Asian faces were 10-100 times more likely to be misidentified than whites by commercial facial recognition software

- D. Castelvecchi, “Is facial recognition too biased to be let loose?”, 18 Nov 2020, https://www.nature.com/articles/d41586-020-03186-4

MISTAKEN IDENTITY

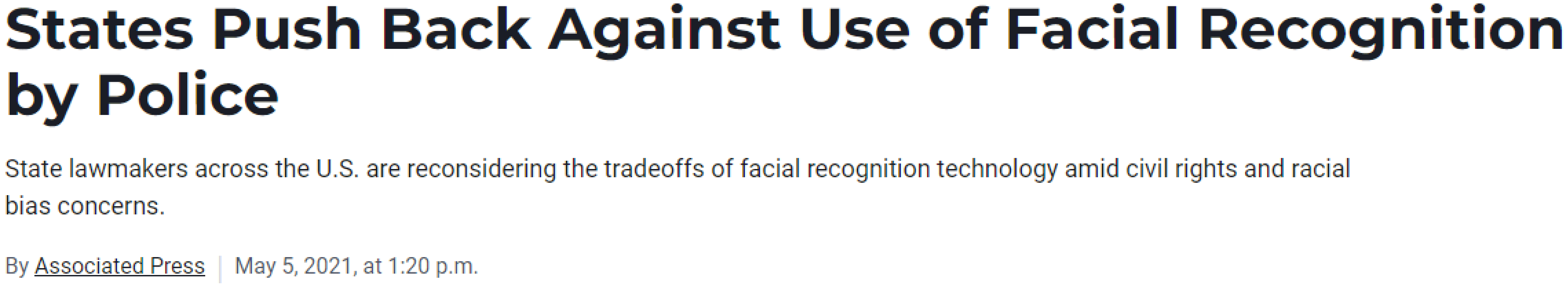
A 2019 review of facial-recognition algorithms shows the chance of false positives* — incorrectly finding matches between two faces — when comparing high-quality US mugshots of different people of the same gender and race†. The rate is highest for female faces of people of colour, but differs across algorithms (shown in two examples).



*Algorithm’s confidence threshold for a ‘match’ was set so as to ensure the false-positive rate for white males was 1 per 10,000; others used same threshold. †Ethnicities as described in ref. 5.

©nature

Moves to Limit Use of Facial Recognition



States Push Back Against Use of Facial Recognition by Police

State lawmakers across the U.S. are reconsidering the tradeoffs of facial recognition technology amid civil rights and racial bias concerns.

By Associated Press | May 5, 2021, at 1:20 p.m.

"At least seven states and nearly two dozen cities have limited government use of the technology amid fears over civil rights violations, racial bias and invasion of privacy. Debate over additional bans, limits and reporting requirements has been underway in about 20 state capitals this legislative session, according to data compiled by the Electronic Privacy Information Center."

REVISE: **RE**vealing **VIS**ual Bias**E**s in Datasets

- **Object-based**: size, context, or diversity of the depicted objects.
  - Airplane is overrepresented as very large in images, as there are few images of airplanes smaller and flying in the sky
  - Person appears more with unhealthy food like cake (55%) or hot dog (56%) than broccoli (15%) or orange (29%)
- **Person-based**: portrayal of people within the dataset.
  - As the skin tone of the person in an image increases in darkness, the person is more likely to be smaller and further from the center
- **Geography-based**: representation of different geographic locations.
  - Countries in Africa and Asia that are already underrepresented are frequently represented by non-locals rather than locals

A. Wang, et. al, "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets," 23 Jul 21, https://arxiv.org/pdf/2004.07999.pdf

Photos of People With Musical Instruments

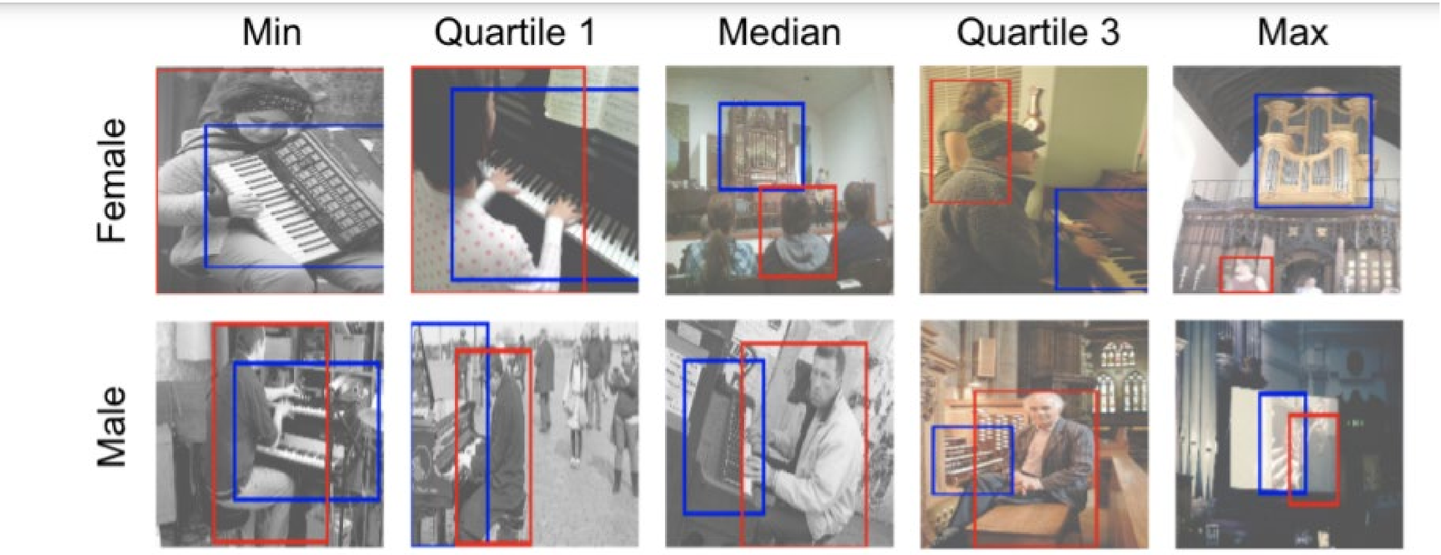


Fig. 9: 5 images from OpenImages for a person (red bounding box) of each gender pictured with an organ (blue bounding box) along the gradient of inferred 3D distances. Males tend to be featured as actually playing the instrument, whereas females are oftentimes merely in the same space as the instrument.

A. Wang, et. al, "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets," 23 Jul 21, https://arxiv.org/pdf/2004.07999.pdf

Challenges in Object Classification

- Lighting (brightness and color)
- Foreshortening (distortion from viewing angle)
- Aspect (shape from viewing angle)
- Occlusion (hidden parts of object)
- Deformation (object changes shape)

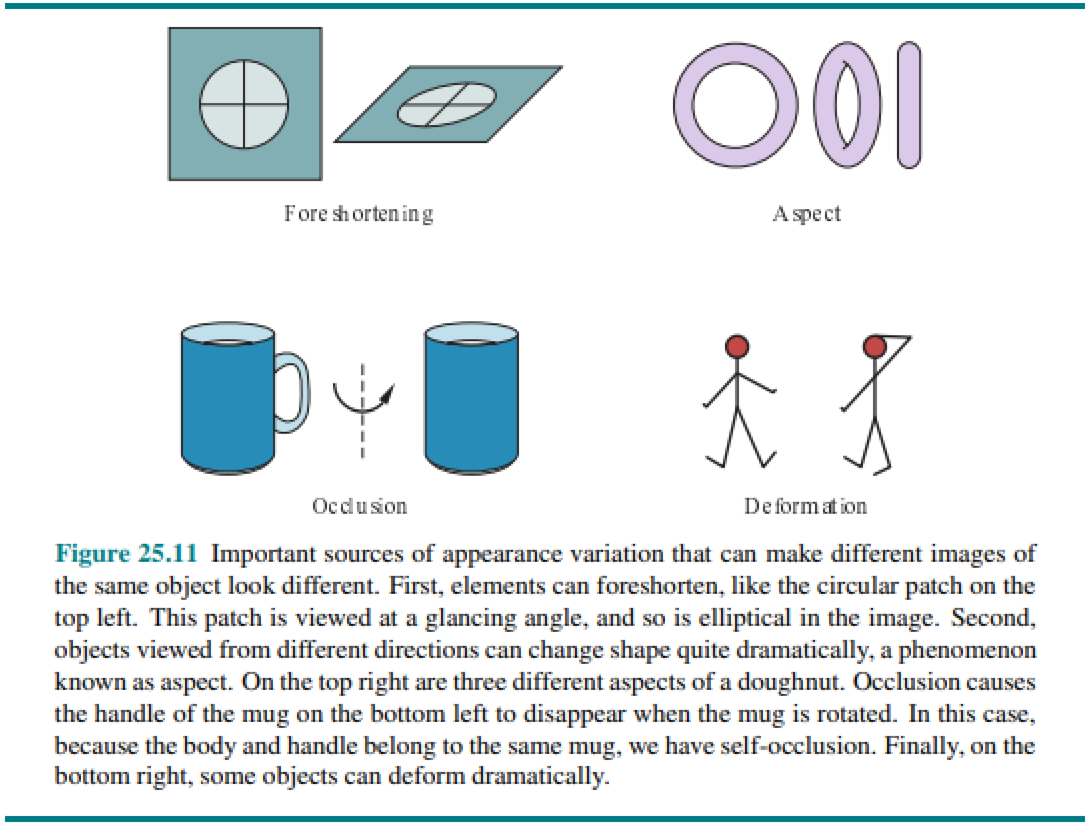
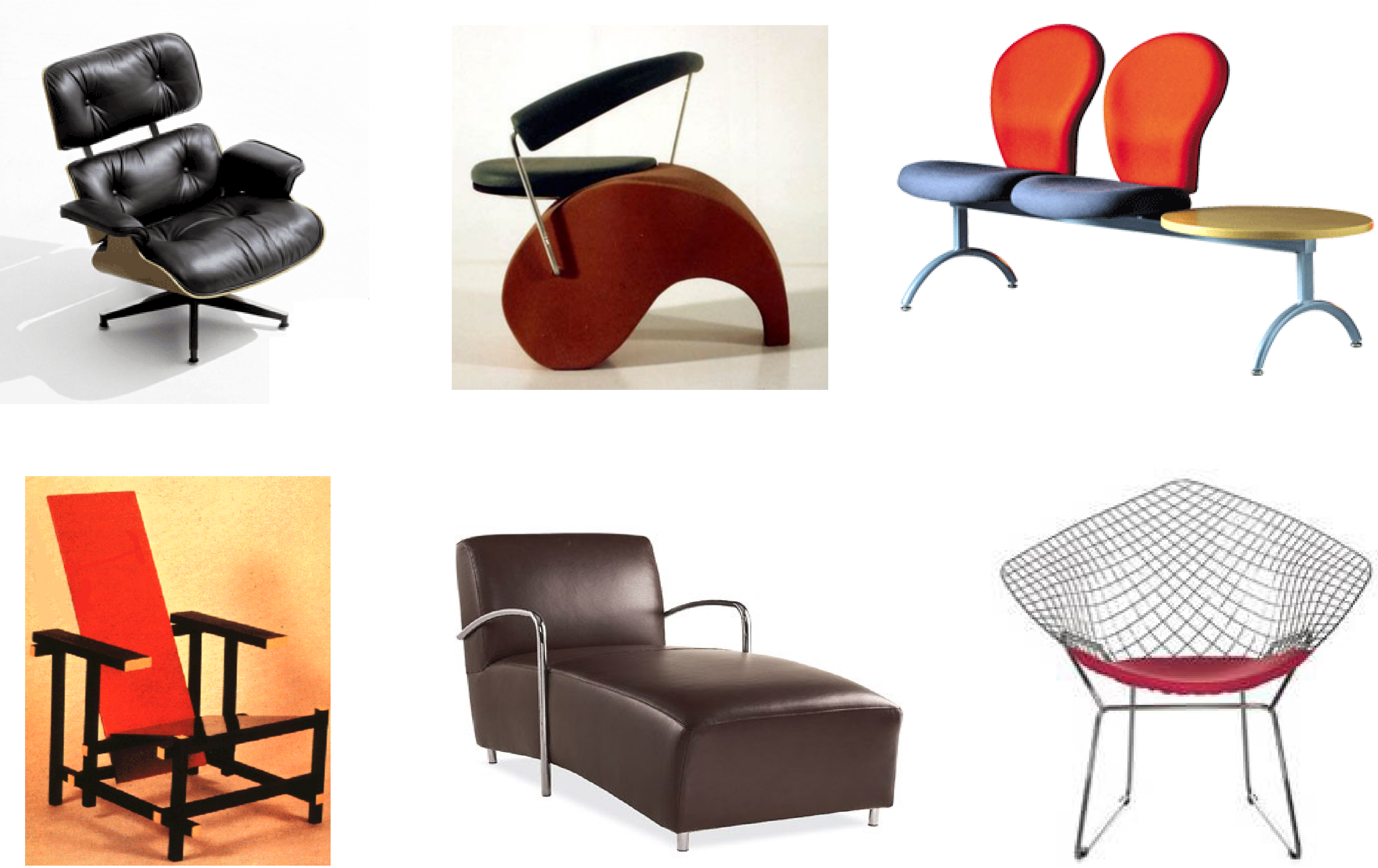


Figure 25.11 Important sources of appearance variation that can make different images of the same object look different. First, elements can foreshorten, like the circular patch on the top left. This patch is viewed at a glancing angle, and so is elliptical in the image. Second, objects viewed from different directions can change shape quite dramatically, a phenomenon known as aspect. On the top right are three different aspects of a doughnut. Occlusion causes the handle of the mug on the bottom left to disappear when the mug is rotated. In this case, because the body and handle belong to the same mug, we have self-occlusion. Finally, on the bottom right, some objects can deform dramatically.

Russell & Norvig, 2020, Chapter 25, p. 896

Challenges in Object Classification: Intra-Class Variation



Slide credit: Fei-Fei, Fergus & Torralba

Using CNNs to Classify Images

- Each combination of a convolution with a ReLU activation function is a local pattern detector
- Multiple layers detect patterns of patterns
- Data set augmentation by translation, stretching, rotating, cropping the image improves accuracy

Conclusion

- Computer vision builds on image processing fundamentals
- Simple image processing techniques for edge detection and image segmentation
- CNNs have made many previous computer vision techniques obsolete because of their ability to learn models of many visual tasks from examples, but still require careful design choices
- **Next class (September 25)**
    - Applications of computer vision and image synthesis

Appendix: Image Segmentation

- Partitioning an image into meaningful regions (per pixel).
- Types:
  - **Semantic:** Classify every pixel (e.g., road, car, sky).
  - **Instance:** Distinguish between objects of the same class (e.g., two different dogs).
  - **Panoptic:** Combine both.
- Applications:
  - Self-driving cars (pedestrian/road segmentation)
  - Medicine (tumor boundary detection)
  - Agriculture (crop vs. weed mapping)

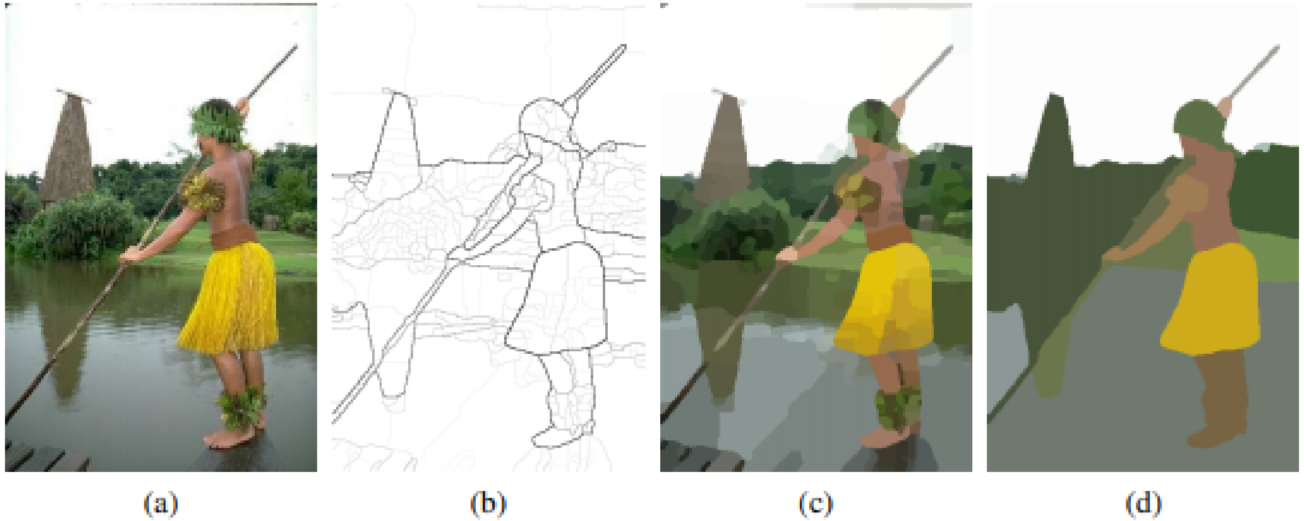


Figure 25.10 (a) Original image. (b) Boundary contours, where the higher the P_b value, the darker the contour. (c) Segmentation into regions, corresponding to a fine partition of the image. Regions are rendered in their mean colors. (d) Segmentation into regions, corresponding to a coarser partition of the image, resulting in fewer regions. (Images courtesy of Pablo Arbelaez, Michael Maire, Charless Fowlkes and Jitendra Malik.)

Russell & Norvig, 2020, Chapter 25, p. 895

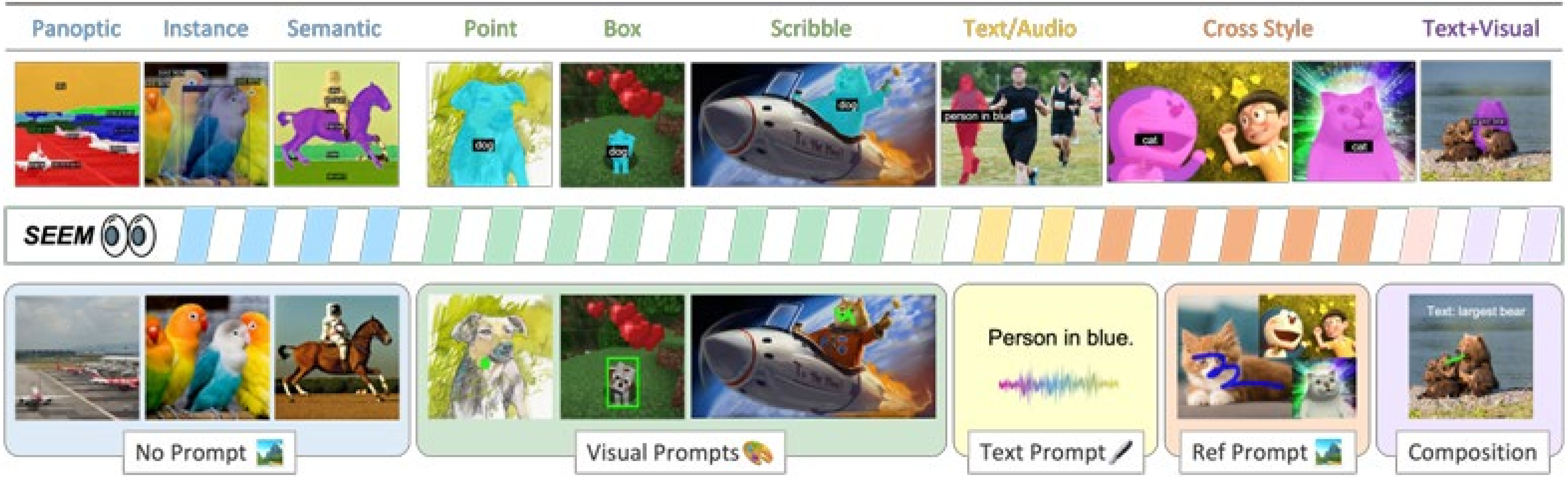
Classical Approaches to Image Segmentation

- Thresholding (Otsu): Detecting the boundaries, using supervised learning
- Edge detection (Canny)
- Finding the regions using clustering on pixels, often represented as nodes in a graph
  - Clustering (k-means, mean-shift)
  - Graph-based (Normalized Cuts)

Deep Learning Approaches to Image Segmentation

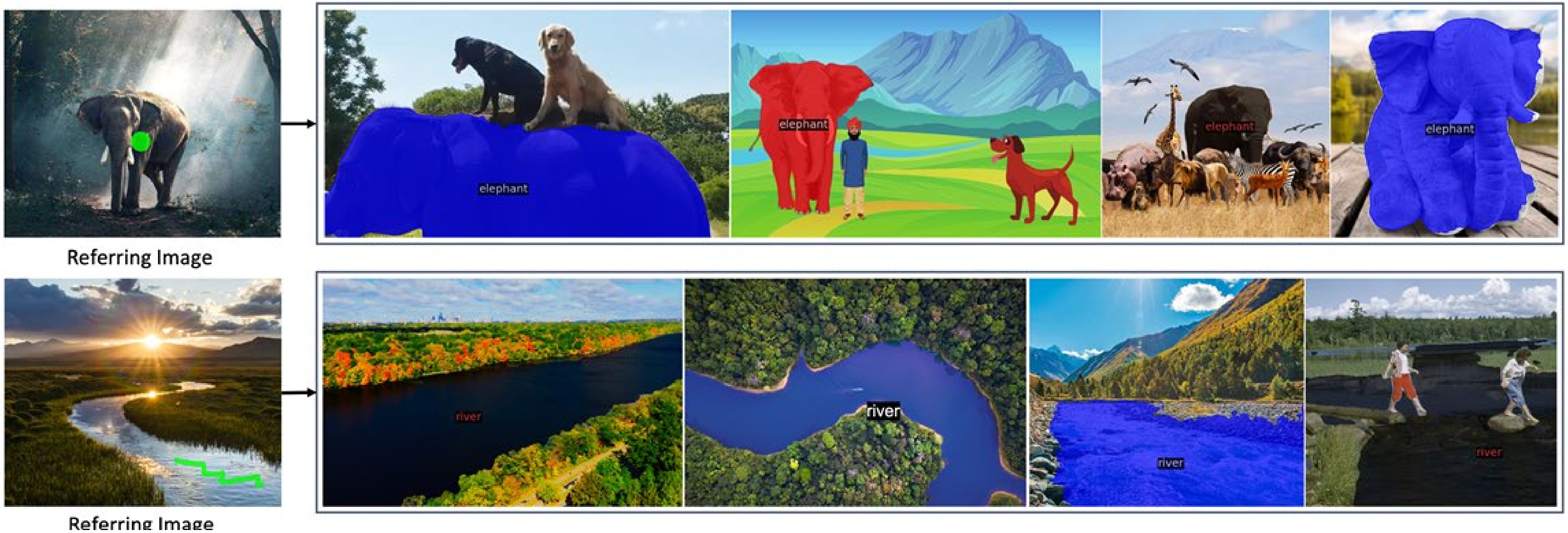
- Fully Convolutional Networks (FCNs)
- U-Net (medical imaging)
- Mask R-CNN (instance segmentation)
- Transformers/SAM (foundation models)

Segmentation in a Joint Visual-Semantic Space



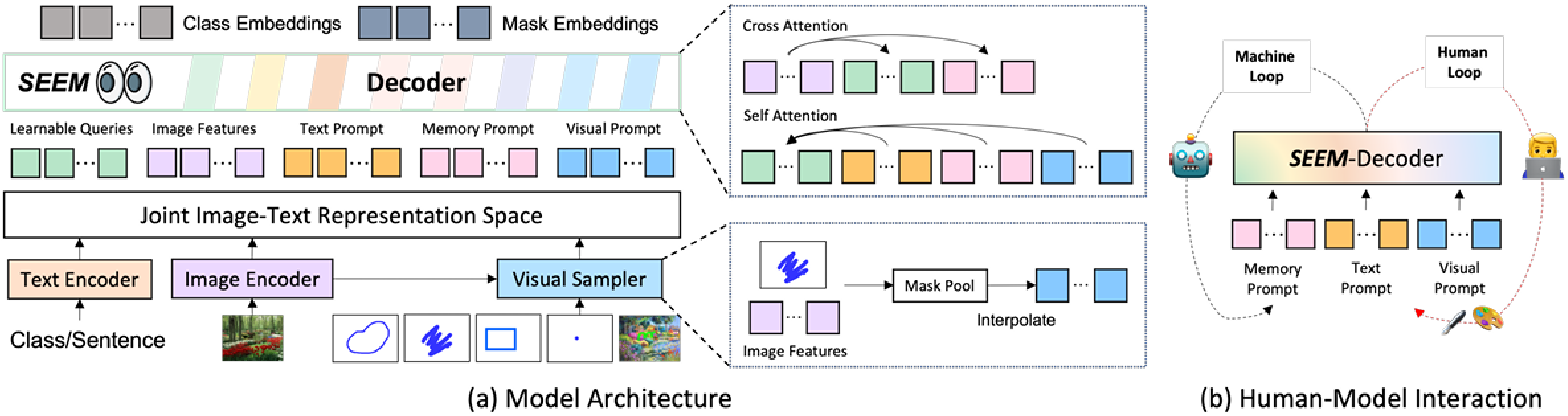
X. Zou, et. al, “Segment Everything Everywhere All at Once,” July 11, 2023, https://arxiv.org/abs/2304.06718

Zero-Shot Visual Referring Segmentation



X. Zou, et. al, “Segment Everything Everywhere All at Once,” July 11, 2023, https://arxiv.org/abs/2304.06718

SEEM Decoder



X. Zou, et. al, “Segment Everything Everywhere All at Once,” July 11, 2023, https://arxiv.org/abs/2304.06718

Evaluation Metrics for Image Segmentation

- IoU (Intersection over Union)
- Dice coefficient
- Pixel accuracy