CS 1674/2074: Intro to Computer Vision

PhD. Nils Murrugarra-Llerena
nem177@pitt.edu

University of Pittsburgh

What is Computer Vision?

Done?

"We see with our brains, not with our eyes" (Oliver Sacks and others)

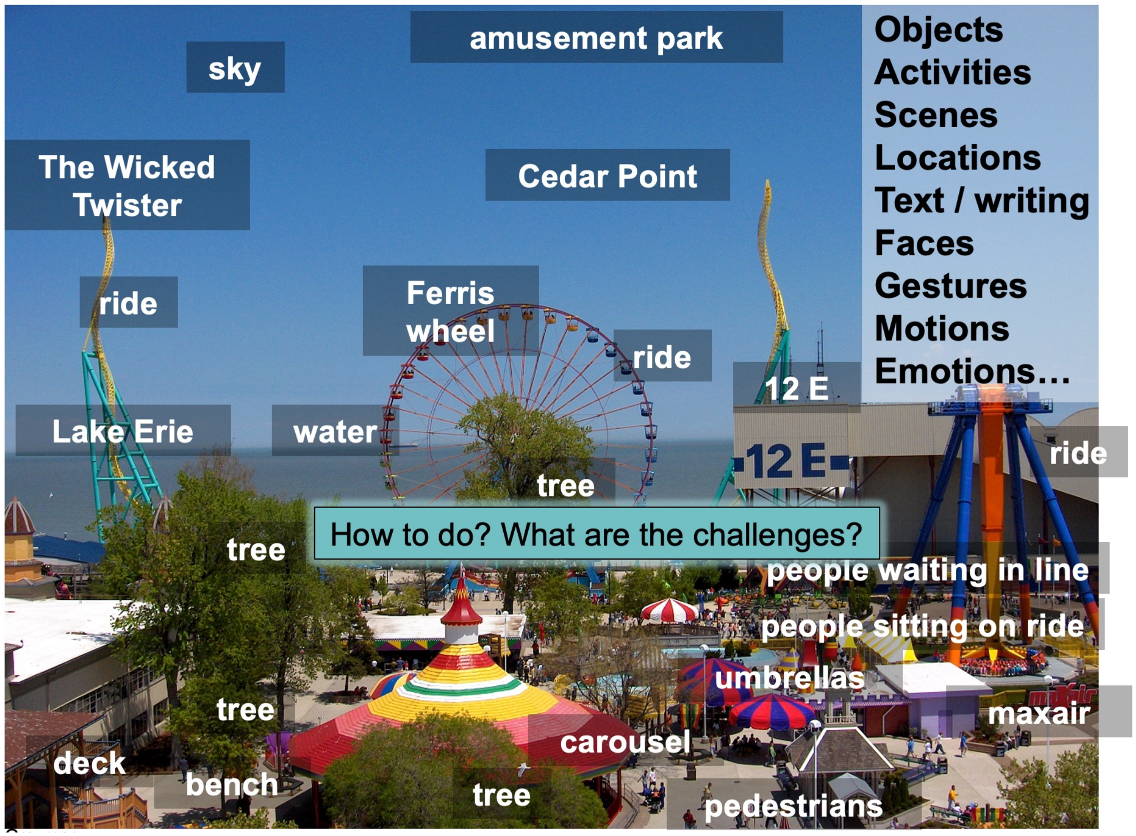
Adapted from Kristen Grauman

What is Computer Vision?

- Automatic understanding of images and video
  - Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities
  - Algorithms to mine, search, and interact with visual data
  - Computing properties and navigating within the 3D world using visual data
  - Generating realistic synthetic visual data

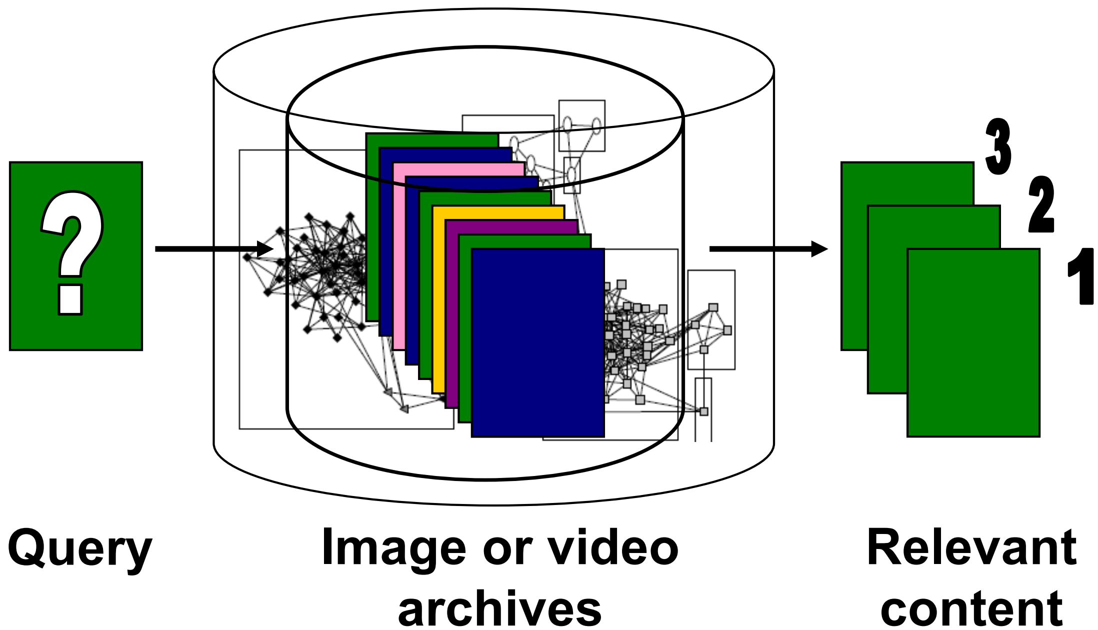
Adapted from Kristen Grauman

What is Computer Vision?



Kristen Grauman

Understanding: Visual search, organization

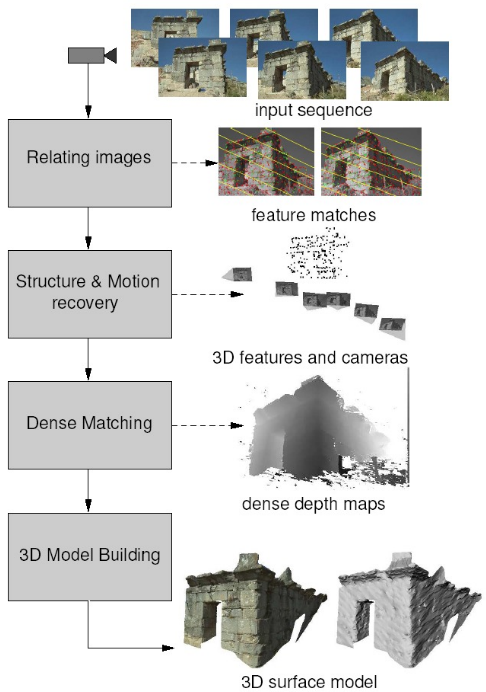


Understanding: Measurement

Real-time stereo

Pollefeys et al.

Structure from motion



Multi-view stereo for community photo collections

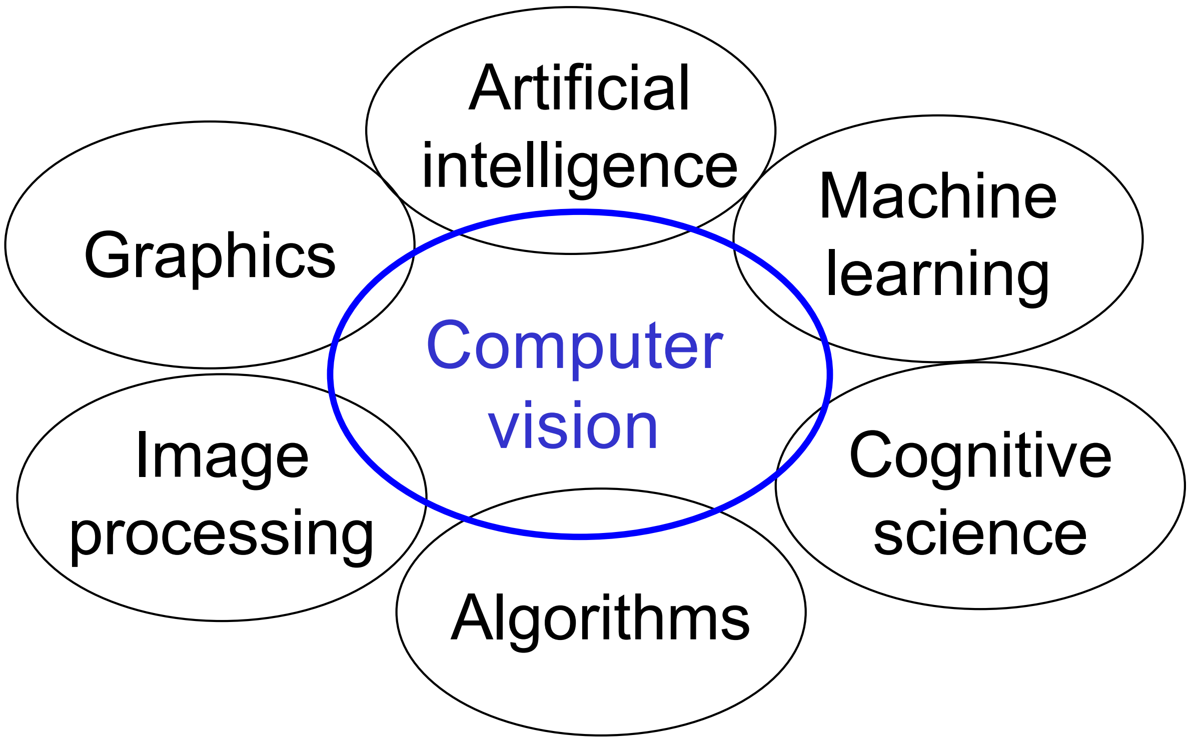
Goesele et al.

Slide credit: L. Lazebnik

Understanding: Generation

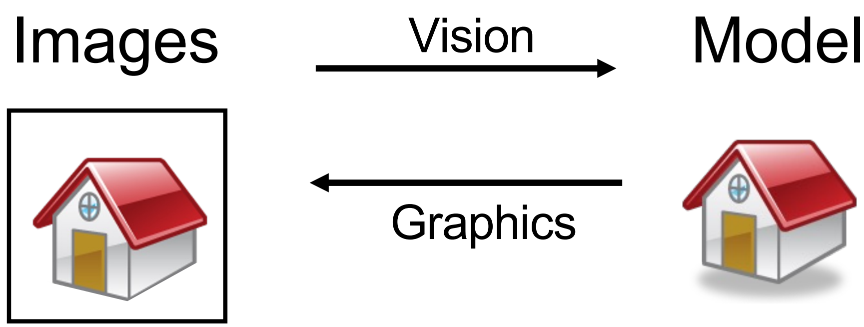
Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation", ICLR 2018

Understanding: Related Disciplines



Kristen Grauman

Understanding: Vision and graphics



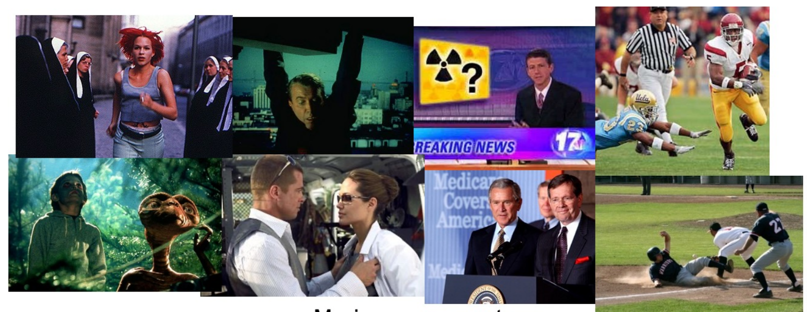
Inverse problems: analysis and synthesis.

Kristen Grauman

Why Vision?

- Images and video are everywhere!

Personal photo albums

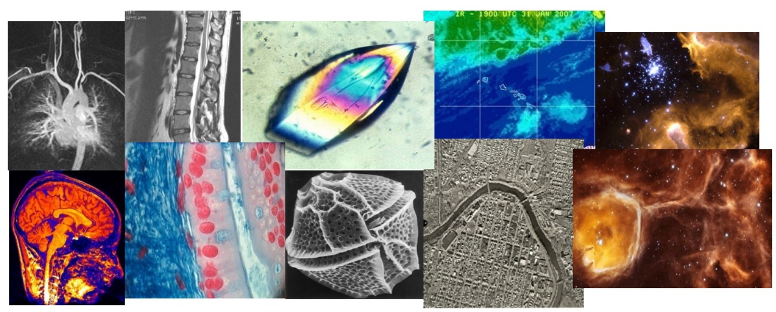


Movies, news, sports

144k hours uploaded to YouTube daily
4.5 mil photos uploaded to Flickr daily
10 bil images indexed by Google

Instagram flickr shutterstock YouTube gettyimages

Surveillance and security



Medical and scientific images

Adapted from Lana Lazebnik

Why Vision?

- As image sources multiply, so do applications
  - Relieve humans of boring, easy tasks
  - Perception for robotics / autonomous agents
  - Organize and give access to visual content
  - Description of content for the visually impaired
  - Human-computer interaction
  - Fun applications (e.g. art styles to my photos)

Adapted from Kristen Grauman

Current Computer Vision Topics: From CVPR, ICCV, and ECCV

CVPR = IEEE/CVF Conference on Computer Vision and Pattern Recognition
ICCV = IEEE/CVF International Conference on Computer Vision
ECCV = European Conference on Computer Vision

Image-text alignment

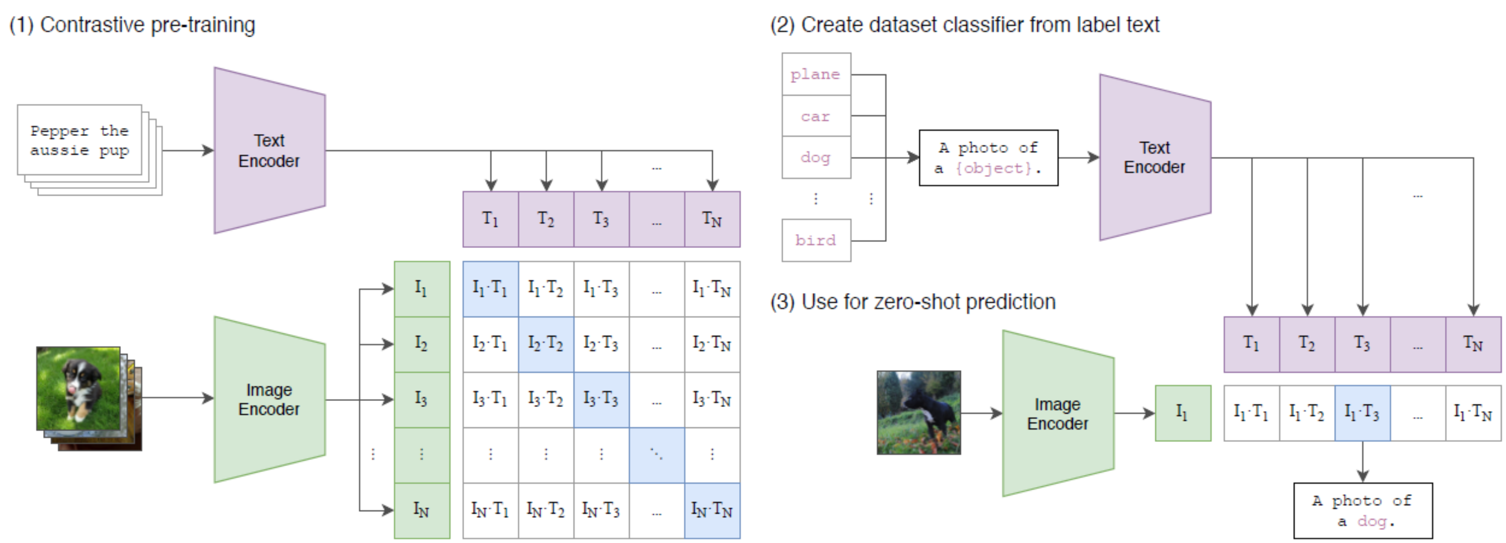


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision." ICML 2021.

Open-vocabulary object detection

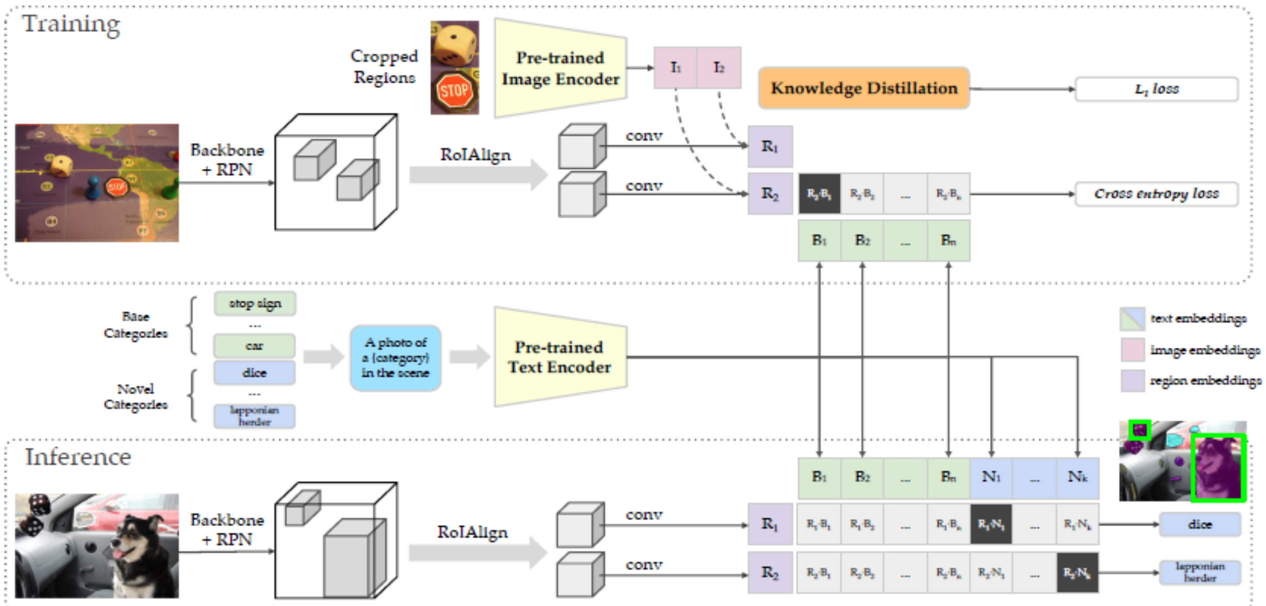


Figure 2: **An overview of using ViLD for open-vocabulary object detection.** ViLD distills the knowledge from a pretrained open-vocabulary image classification model. First, the category text embeddings and the image embeddings of cropped object proposals are computed, using the text and image encoders in the pretrained classification model. Then, ViLD employs the text embeddings as the region classifier (ViLD-text) and minimizes the distance between the region embedding and the image embedding for each proposal (ViLD-image). During inference, text embeddings of novel categories are used to enable open-vocabulary detection.

Gu et al. "Open-vocabulary Object Detection via Vision and Language Knowledge Distillation." ICLR 2021.

How to recognize objects in new modalities

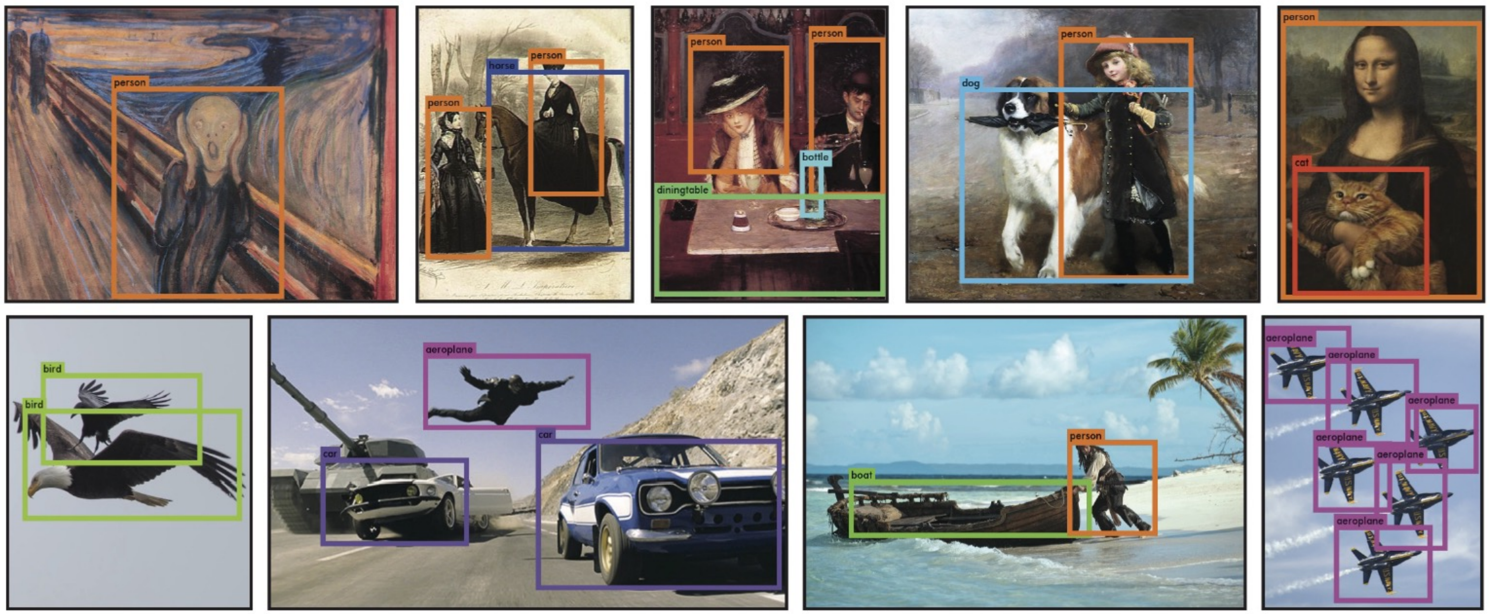
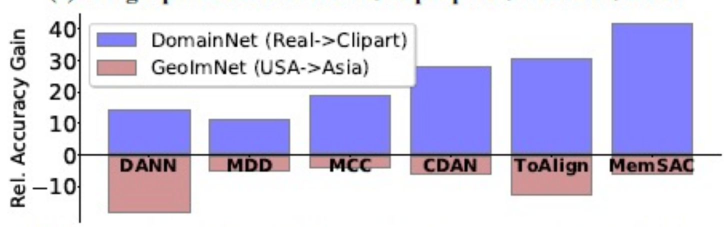


Figure 6: Qualitative Results. YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

Redmon et al., CVPR 2016

How to use models across countries

(a) Geographic bias manifested in proposed GeoNet dataset



(b) Unsupervised domain adaptation does not suffice on GeoNet

Kalluri et al. "GeoNet: Benchmarking Unsupervised Adaptation across Geographies." CVPR 2023.

How to query vision-language models

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**91.83**

(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**94.51**

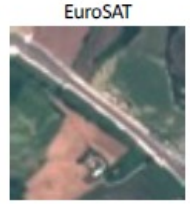
(b)

Describable Textures (DTD)

Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT



Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**83.53**

(d)

Fig. 1 Prompt engineering vs Context Optimization (CoOp). The former needs to use a held-out validation set for words tuning, which is inefficient; the latter automates the process and requires only a few labeled images for learning.

Zhou et al. "Learning to Prompt for Vision-Language Models." IJCV 2022.

How to query vision-language models

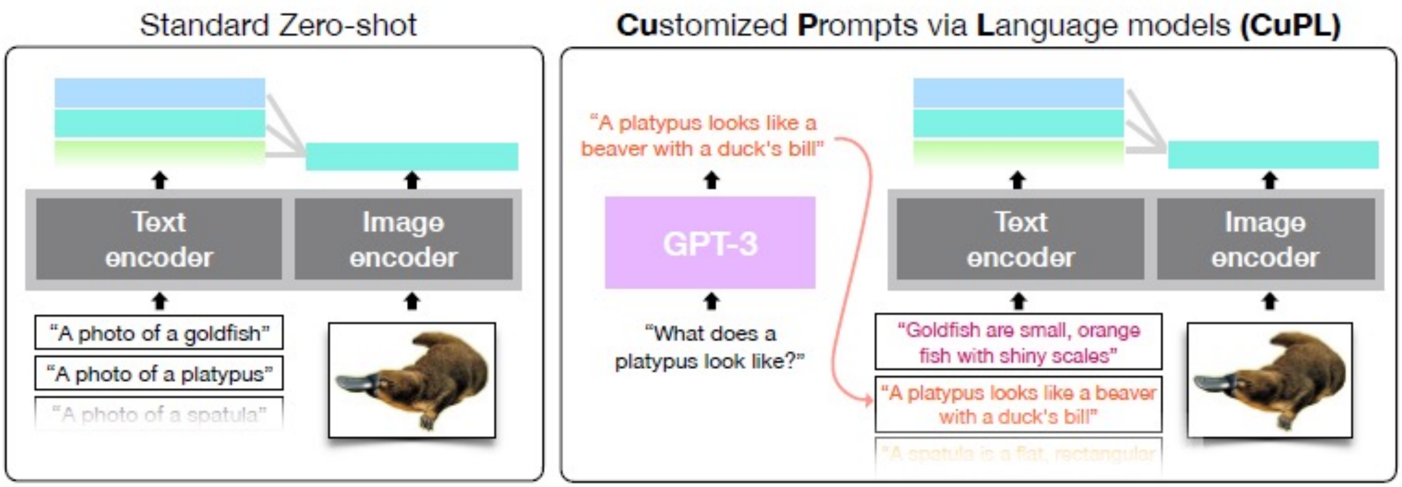


Figure 1: **Schematic of the method.** (Left) The standard method of a zero-shot open vocabulary image classification model (e.g., CLIP (Radford et al., 2021)). (Right) Our method of CuPL. First, an LLM generates descriptive captions for given class categories. Next, an open vocabulary model uses these captions as prompts for performing classification.

Pratt et al. "What does a platypus look like? Generating customized prompts for zero-shot image classification." ICCV 2023.

How to integrate modalities (audio)

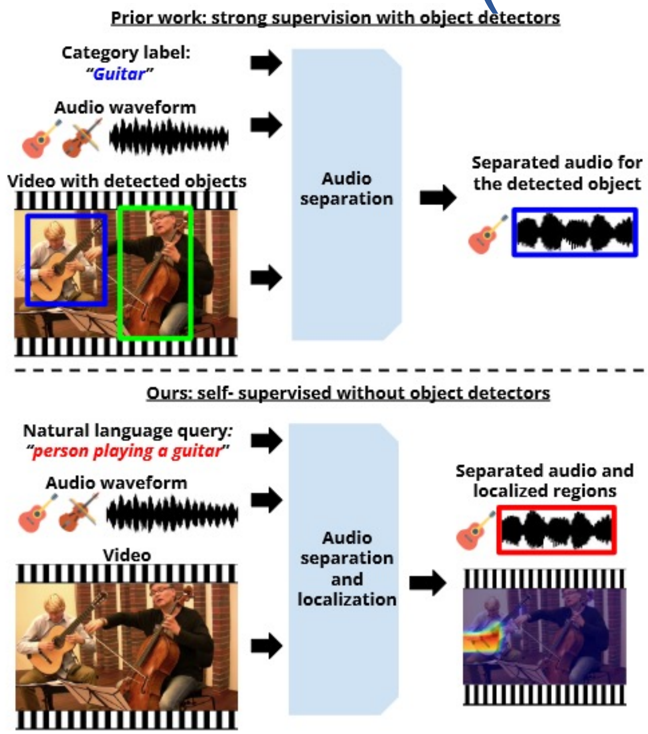


Figure 1. We propose to separate and localize audio sources based on a natural language query, by learning to align the modalities on completely unlabeled videos. In comparison, prior audio-visual sound separation approaches require object label supervision.

Tan et al. "Language-Guided Audio-Visual Source Separation via Trimodal Consistency." CVPR 2023.

How to represent everyday activities

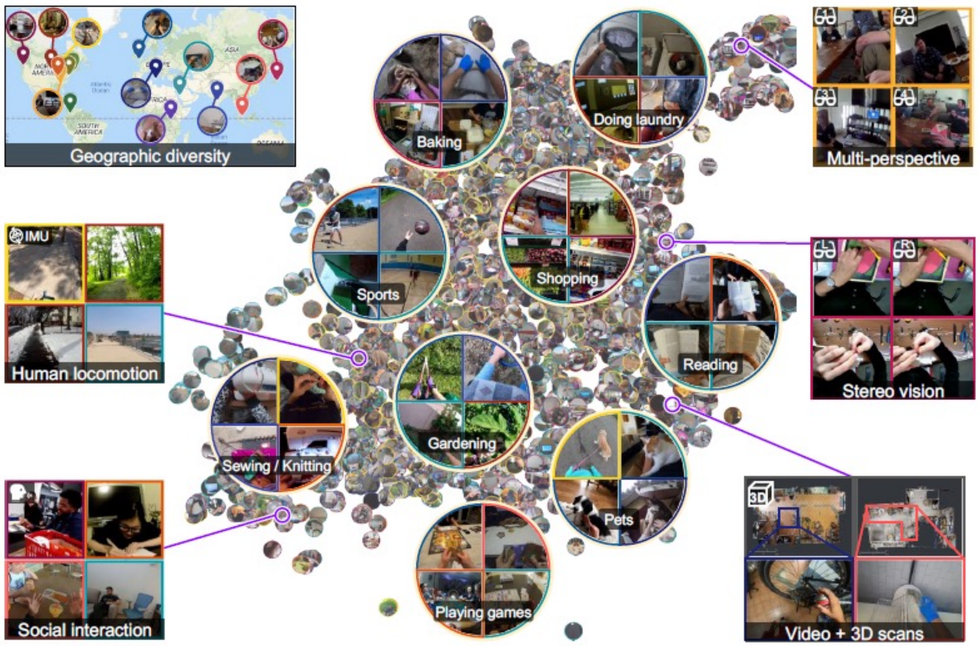


Figure 1. Ego4D is a massive-scale egocentric video dataset of daily life activity spanning 74 locations worldwide. Here we see a snapshot of the dataset (5% of the clips, randomly sampled) highlighting its diversity in geographic location, activities, and modalities. The data includes social videos where participants consented to remain unblurred. See https://ego4d-data.org/fig1.html for interactive figure.

Grauman et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video." CVPR 2022.

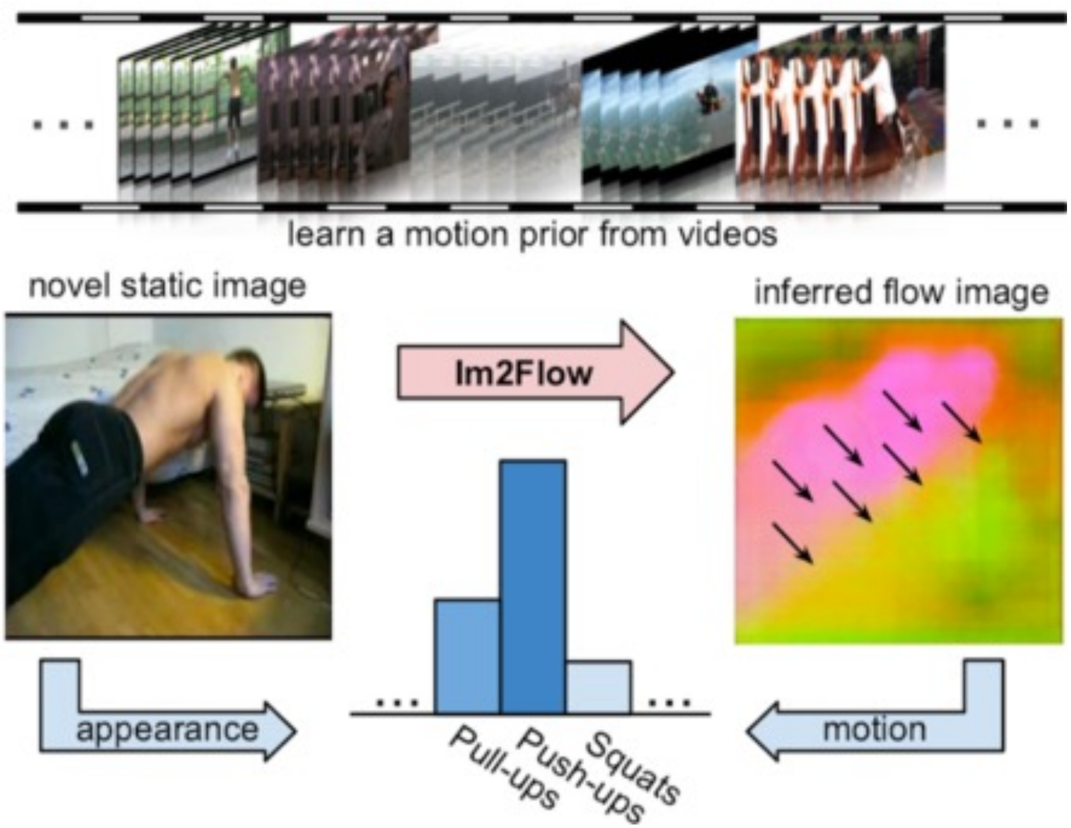
How to understand activities and intents

Bagautdinov et al., CVPR 2017

How to grade how well an activity is performed

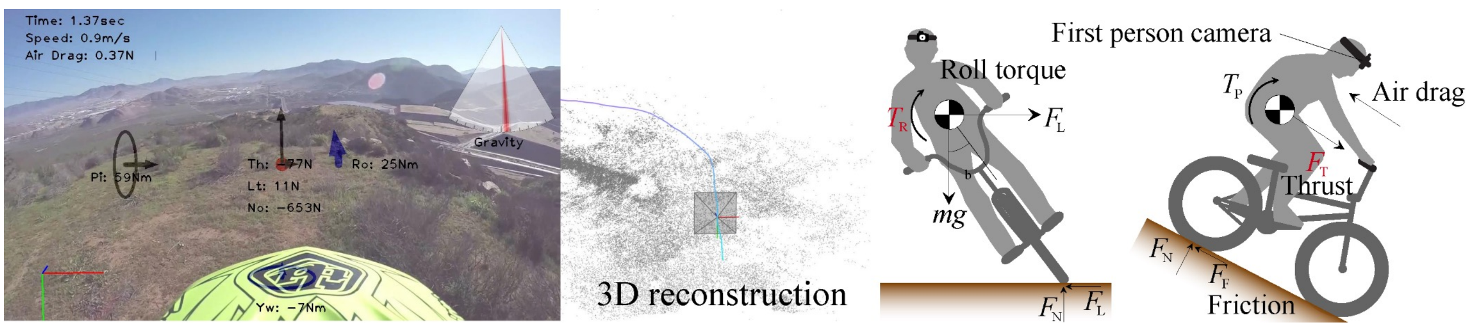
Pirsiavash et al., ECCV 2014

How to imagine motion in static images



Gao et al., CVPR 2018

How to decode physics from video



Park et al., CVPR 2016

How to perform high-level reasoning

Figure 1. **VISPROG is a modular and interpretable neuro-symbolic system for compositional visual reasoning.** Given a few examples of natural language instructions and the desired high-level programs, VISPROG generates a program for any new instruction using *in-context learning* in GPT-3 and then executes the program on the input image(s) to obtain the prediction. VISPROG also summarizes the intermediate outputs into an interpretable *visual rationale* (Fig. 4). We demonstrate VISPROG on tasks that require composing a diverse set of modules for image understanding and manipulation, knowledge retrieval, and arithmetic and logical operations.

Gupta and Kembhavi. "Visual Programming: Compositional visual reasoning without training." CVPR 2023.

How to understand stories in film

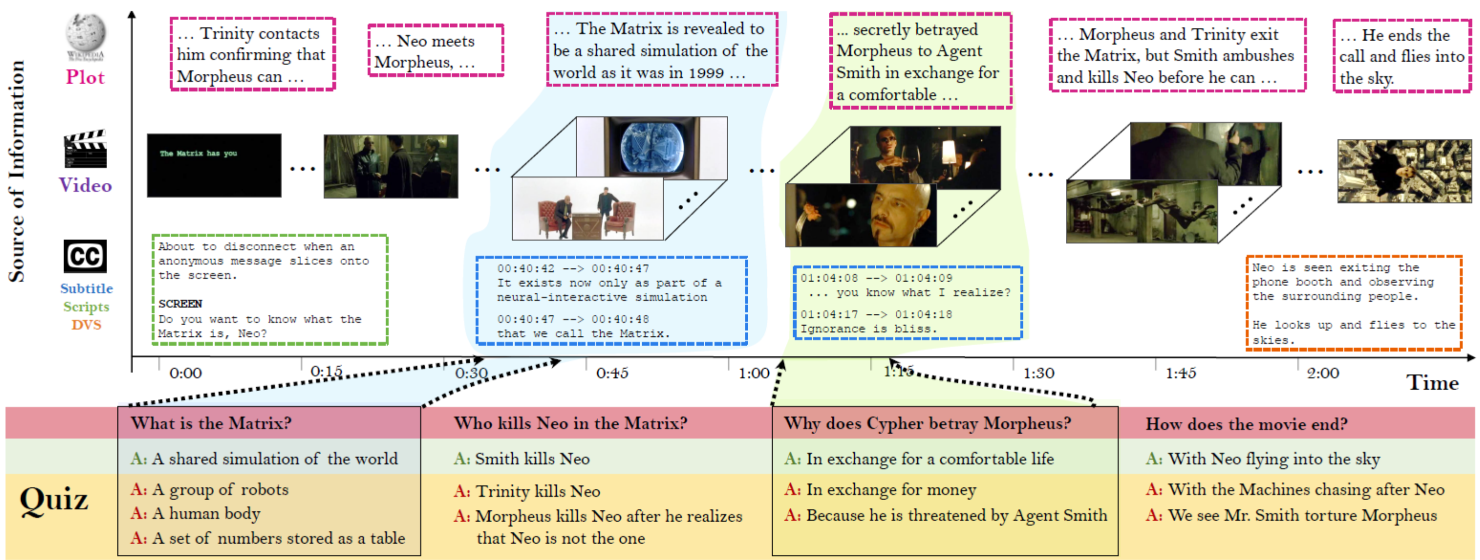
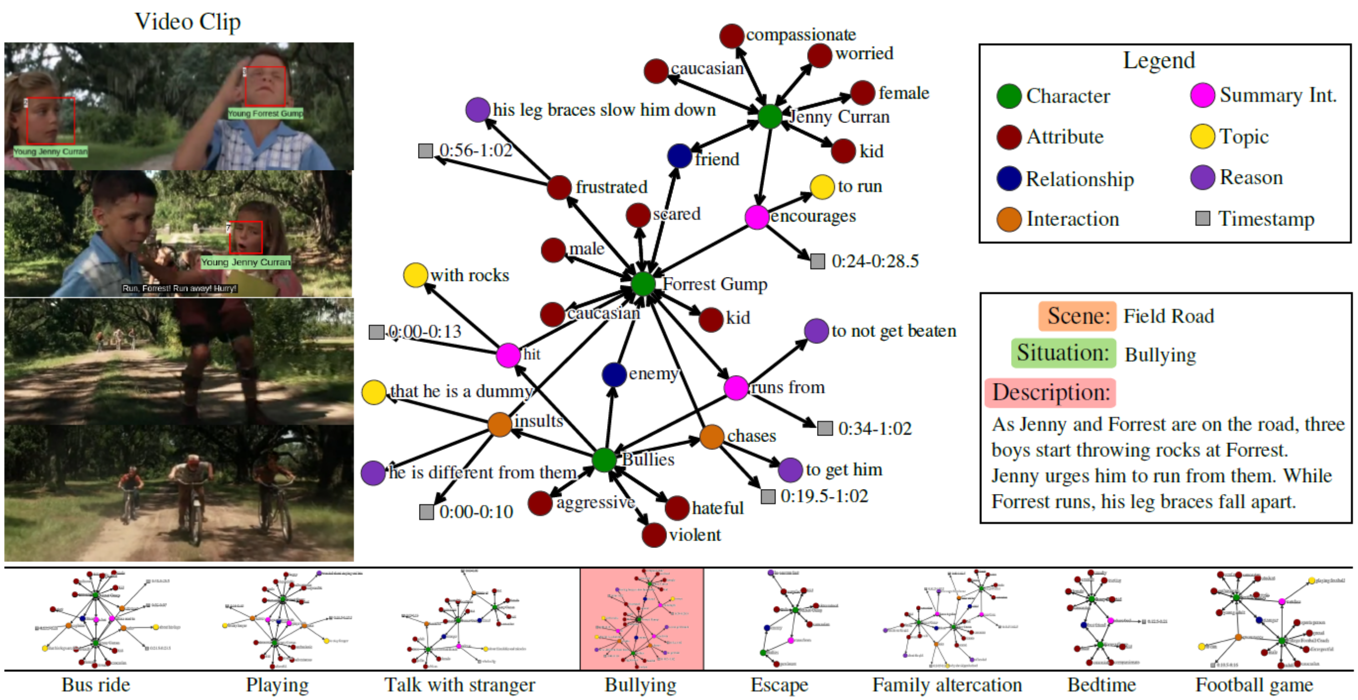


Figure 1: Our MovieQA dataset contains 14,944 questions about 408 movies. It contains multiple sources of information: plots, subtitles, video clips, scripts, and DVS transcriptions. In this figure we show example QAs from *The Matrix* and localize them in the timeline.

Tapaswi et al., CVPR 2016

How to understand roles in film



Vicol et al., CVPR 2018

How to understand media persuasion

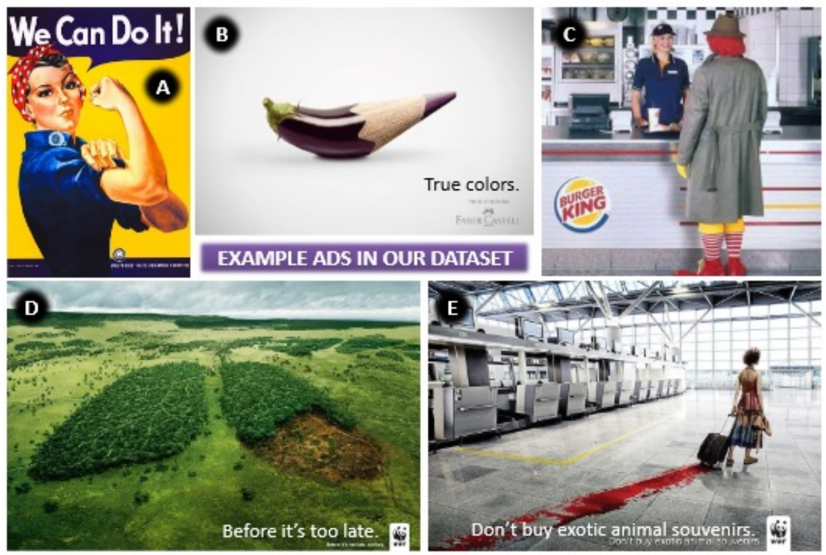


Fig. 1: Example advertisements from our dataset that require challenging visual recognition and reasoning. Despite the potential applications of understanding the messages of ads, this problem has not been tackled in computer vision.

Ye et al. "Interpreting the Rhetoric of Visual Advertisements." TPAMI 2019.

Automatic Understanding of Image and Video Advertisements

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, Adriana Kovashka

University of Pittsburgh

Understanding advertisements is more challenging than simply recognizing physical content from images, as ads employ a variety of strategies to persuade viewers.

We collect an advertisement dataset containing 64,832 images and 3,477 videos, each annotated by 3-5 human workers from Amazon Mechanical Turk.

Image	Topic	204,340	Strategy	20,000
	Sentiment	102,340	Symbol	64,131
	Q+A Pair	202,090	Slogan	11,130
Video	Topic	17,345	Fun/Exciting	15,380
	Sentiment	17,345	English?	17,374
	Q+A Pair	17,345	Effective	16,721

Here are some sample annotations in our dataset.

What's being advertised in this image?

Cars, automobiles

What sentiments are provoked in the viewer?

Amused, Creative, Impressed, Youthful, Conscious

What strategies are used to persuade viewer?

Symbolism, Contrast, Straightforward, Transferred qualities

What should the viewer do, and why should they do this?

- I should buy Volkswagen because it can hold a big bear.
- I should buy VW SUV because it can fit anything and everything in it.
- I should buy this car because it can hold everything I need.

More information available at http://cs.pitt.edu/~kovashka/ads

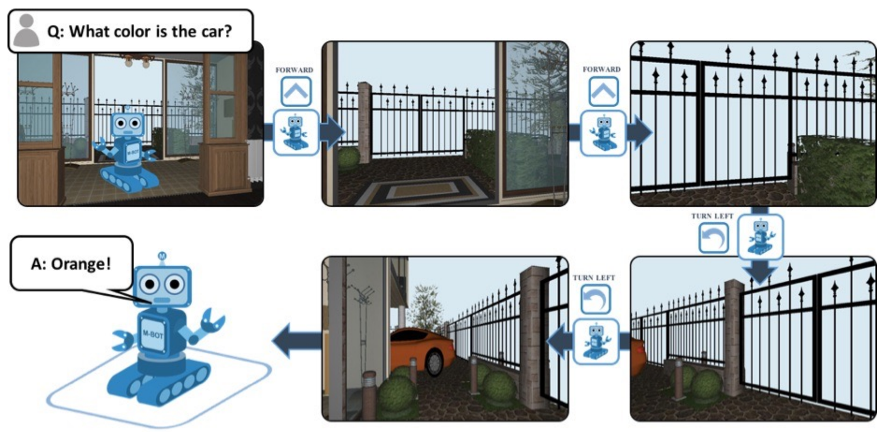
Hussein et al., CVPR 2017

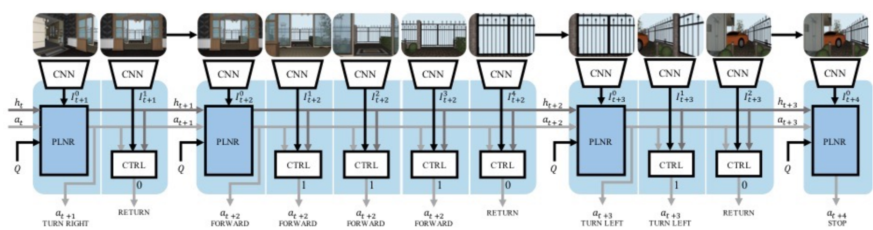
How to generate arbitrary content

Figure 1. Make-A-Scene: Samples of generated images from text inputs (a), and a text and scene input (b). Our method is able to both generate the scene (a, bottom left) and image, or generate the image from text and a simple sketch input (b, center).

Gafni et al. "Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors." ECCV 2022.

How to reason and act





Das et al., CVPR 2018

How to use language models for robotics tasks

Figure 1: LLMs have not interacted with their environment and observed the outcome of their responses, and thus are not grounded in the world. SayCan grounds LLMs via value functions of pretrained skills, allowing them to execute real-world, abstract, long-horizon commands on robots.

Ahn et al. "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances." CoRL 2023.

Computer vision is not solved

- Deep learning makes excellent use of massive data (labeled for the task of interest?)

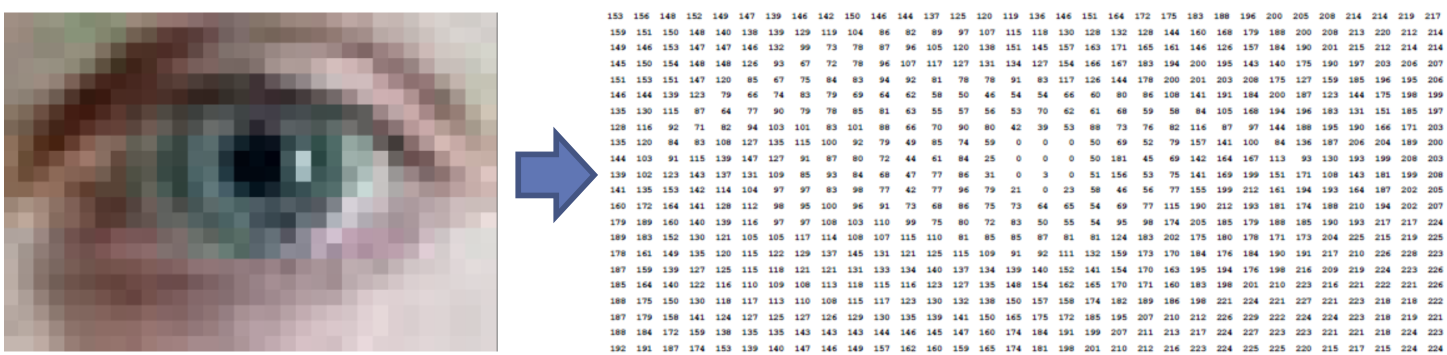
  - But it's hard to understand *how* it does so, makes it hard to fix when it doesn't work well
  - It doesn't work well when massive data is not available and your task is different than tasks for which data is available
  - We can recognize objects with 97% accuracy but reasoning about relationships and intent is harder

Why is Vision difficult?

- Ill-posed problem: real world much more complex than what we can measure in images
  - 3D → 2D
  - Motion → static
- Impossible to literally "invert" image formation process with limited information
  - Need information outside of this particular image to generalize what image portrays (e.g. to resolve occlusion)

Adapted from Kristen Grauman

What the computer see?



Why is this problematic?

Adapted from Kristen Grauman and Lana Lazebnik

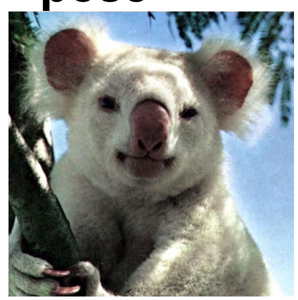
Challenges: many nuisance parameters

Illumination

Object pose

Clutter

Occlusions



Intra-class appearance

Viewpoint

Kristen Grauman

Think again about the pixels…

Challenges: intra-class variation





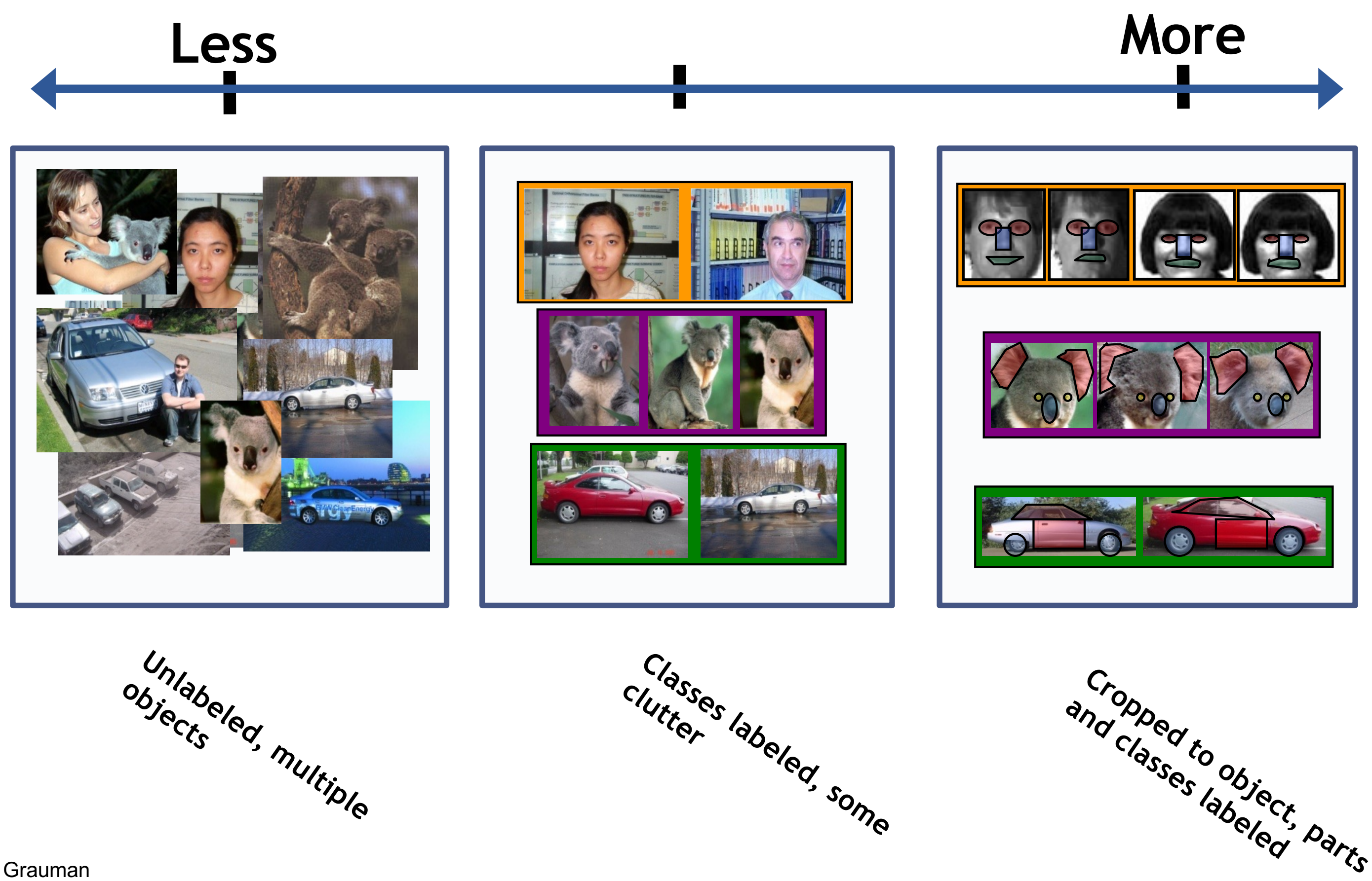

slide credit: Fei-Fei, Fergus & Torralba

Challenges: Complexity

- Thousands to millions of pixels in an image
- 3,000-30,000 human recognizable object categories
- 30+ degrees of freedom in the pose of articulated objects (humans)
- Billions of images indexed by Google Image Search
- 1.424 billion smart camera phones sold in 2015
- About half of the cerebral cortex in primates is devoted to processing visual information [Felleman and van Essen 1991]

Kristen Grauman

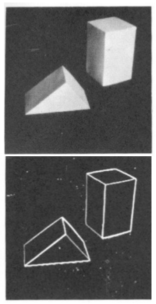
Challenges: Limited supervision

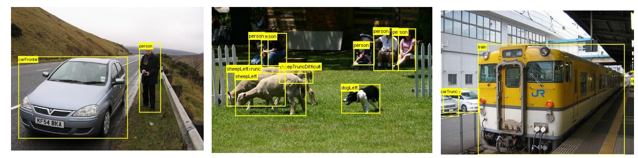


Kristen Grauman

Challenges: Evolution of datasets

- Challenging problem → active research area





PASCAL:
20 categories, 12k images

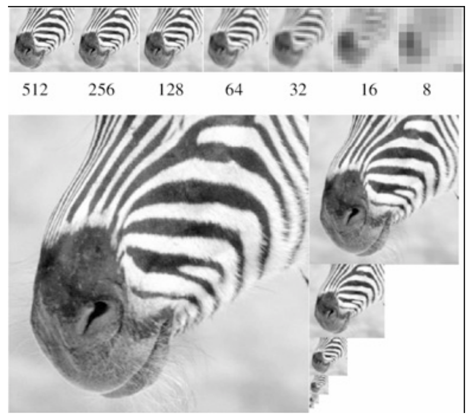
ImageNet:
22k categories, 14mil images

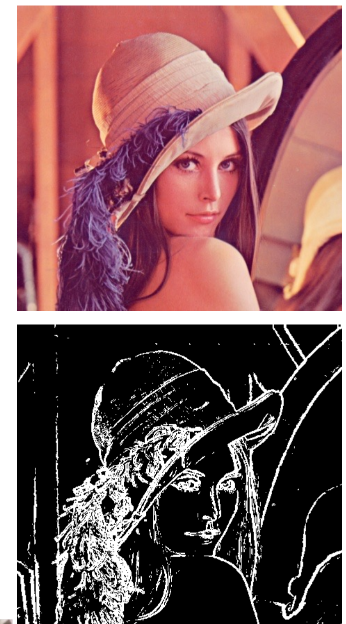
Microsoft COCO:
80 categories, 300k images

Computer Vision: Summary

Overview of topics

Features and Filters



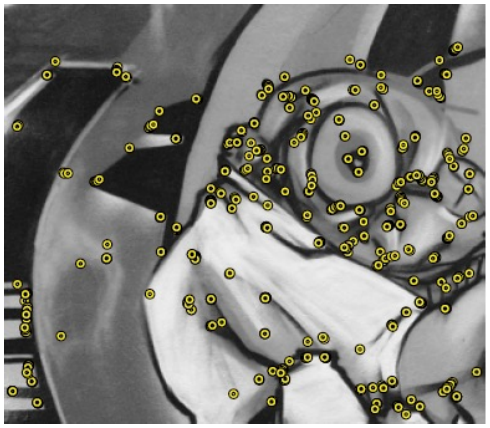


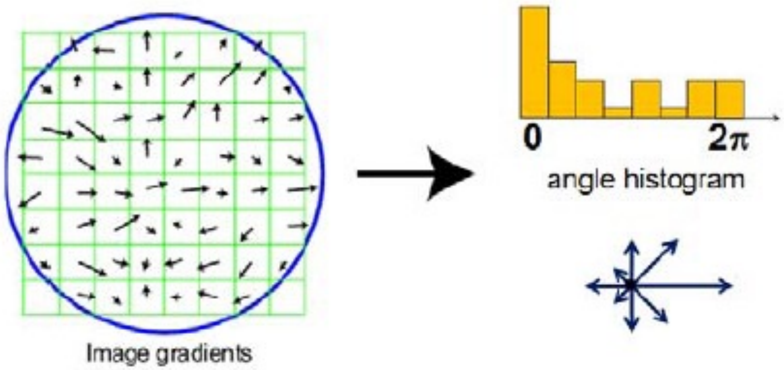
- Describing and transforming textures, colors, edges

Kristen Grauman

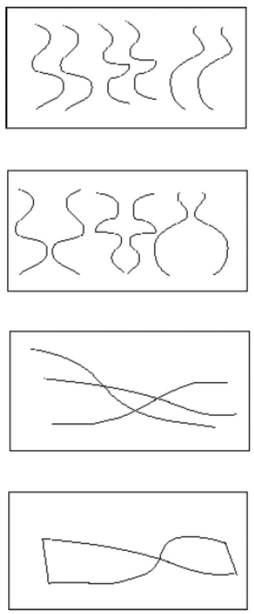
Features and Filters

- Detecting distinctive and repeatable features
- Describing images with local statistics





Grouping



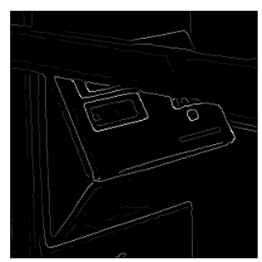
Parallelism

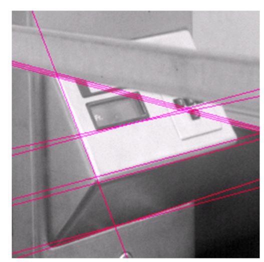
Symmetry

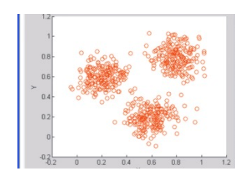
Continuity

Closure

[fig from Shi et al]





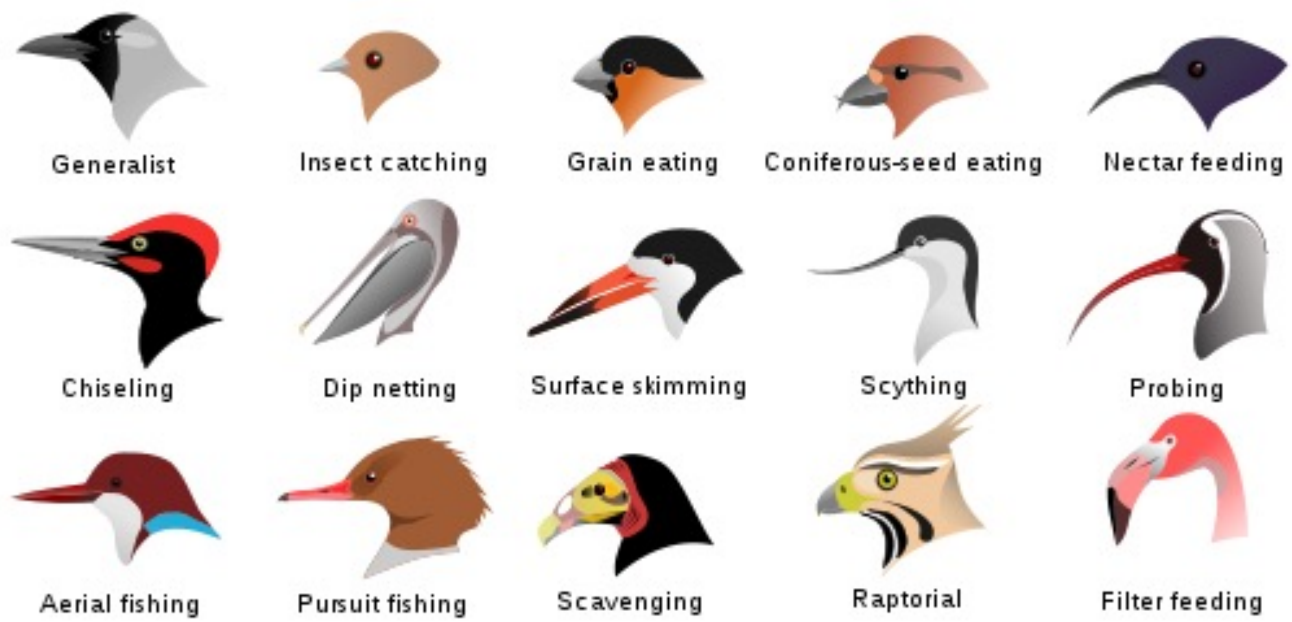


- Segmentation, fitting; what parts belong together?

Kristen Grauman

Image Categorization

- Fine-grained recognition

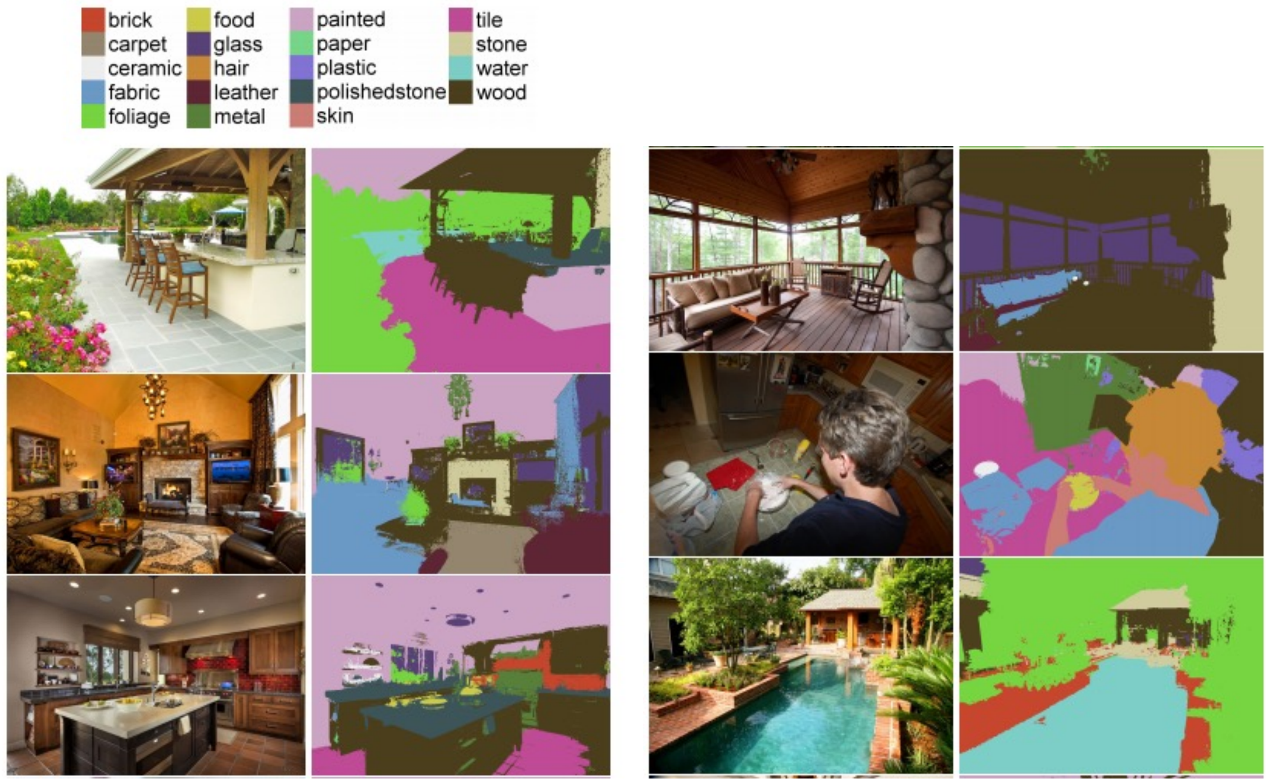


[Visipedia Project]

Slide credit: D. Hoiem

Image Categorization

- Material recognition

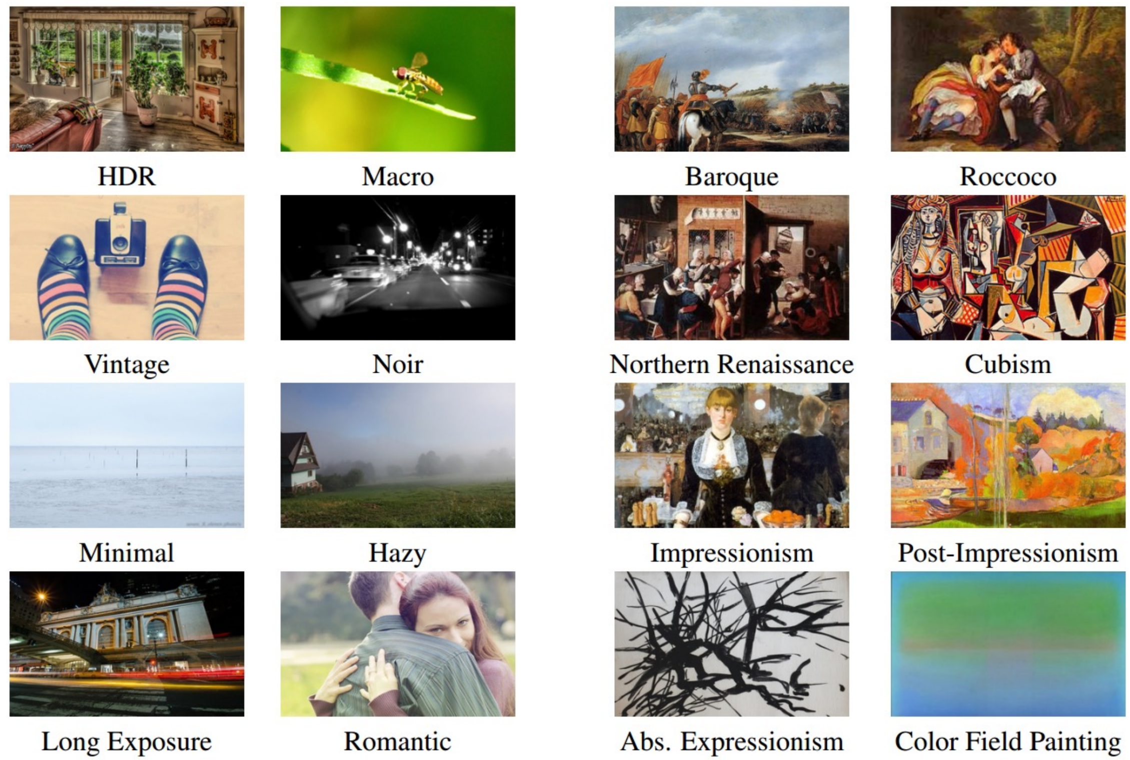


[Bell et al. CVPR 2015]

Slide credit: D. Hoiem

Image Categorization

- Image style recognition

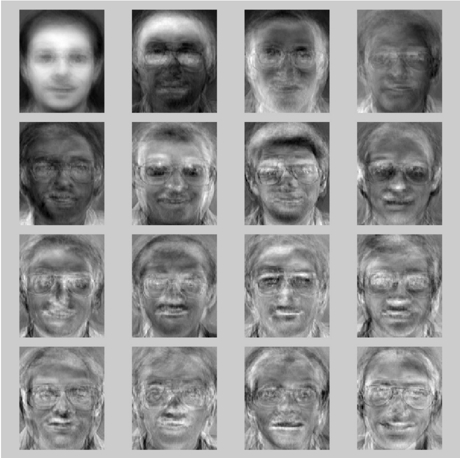


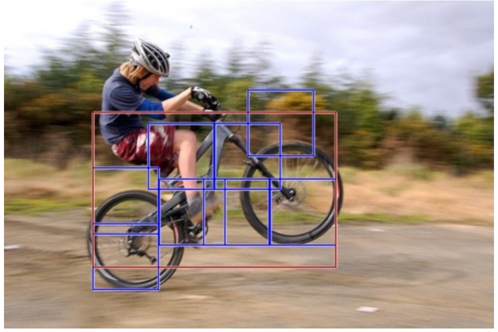
Flickr Style: 80K images covering 20 styles. Wikipaintings: 85K images for 25 art genres.

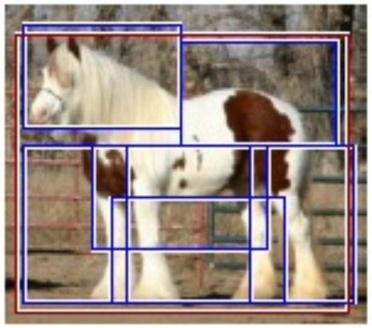
[Karayev et al. BMVC 2014]

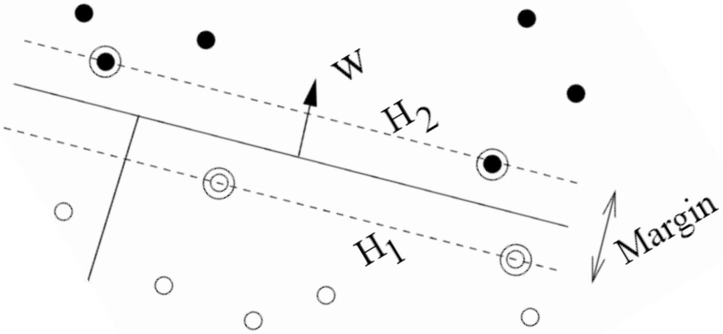
Slide credit: D. Hoiem

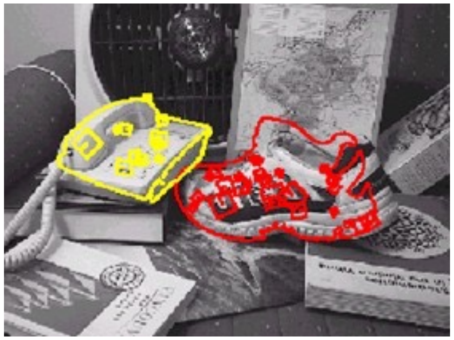
Visual Recognition and SVMs











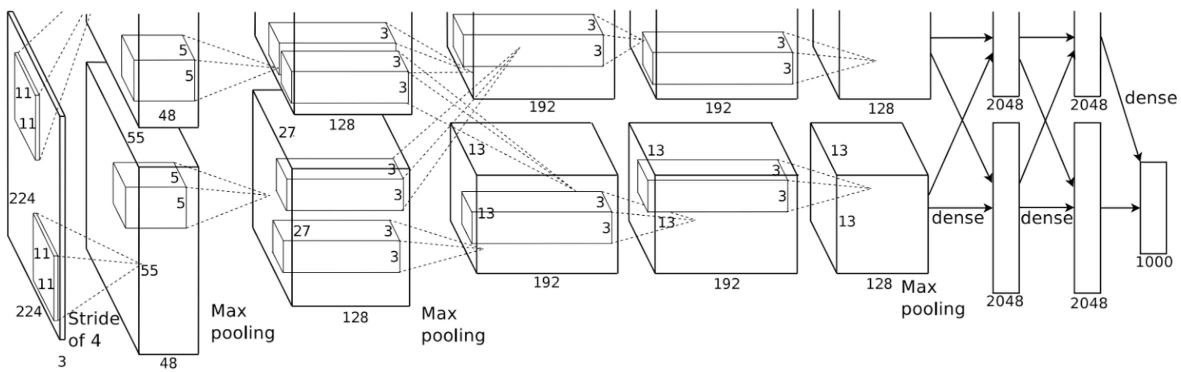
- Recognizing objects and categories, learning techniques

Adapted from Kristen Grauman

Convolutional Neural Networks (CNNs)

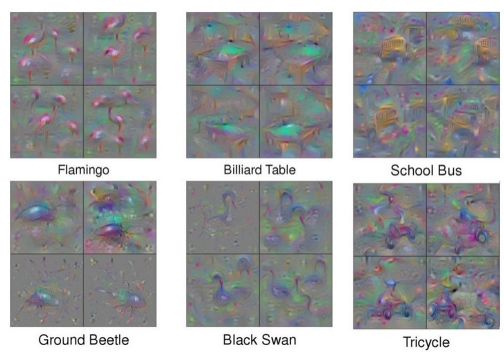
- State-of-the-art on many recognition tasks

Image



Prediction

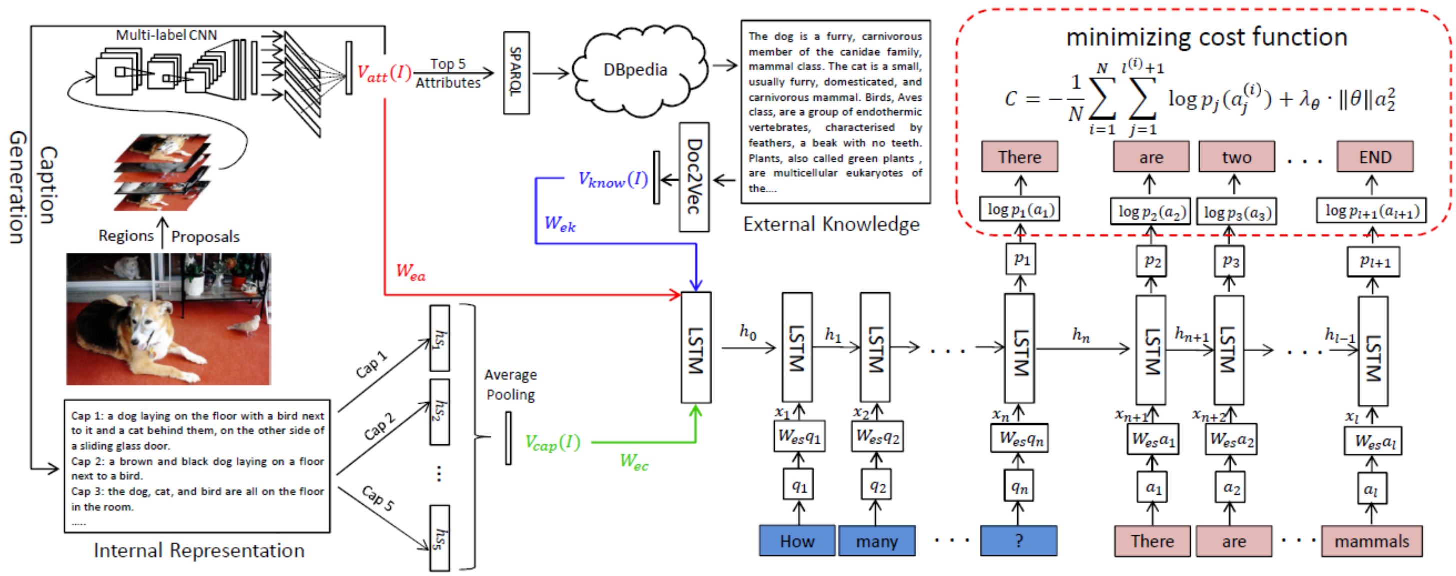
Krizhevsky et al., NIPS 2012



Yosinski et al., ICML DL workshop 2015

Recurrent Neural Networks (RNNs)

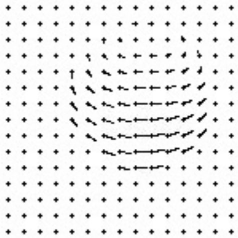
- Sequence processing, e.g. question answering

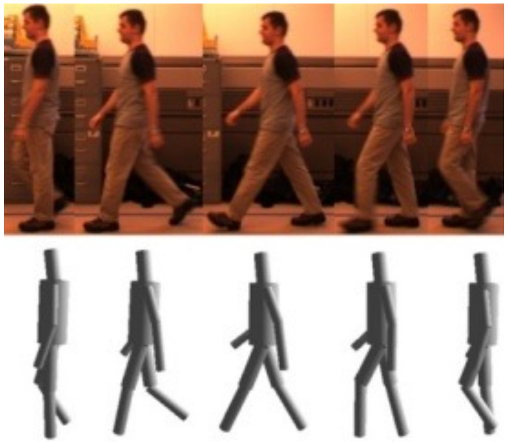


Wu et al., CVPR 2016

Motion and tracking

- Tracking objects, video analysis

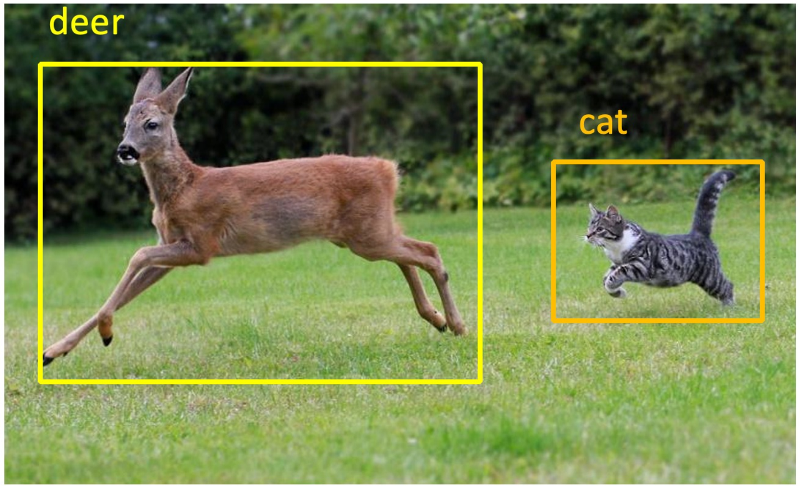




Tomas Izo

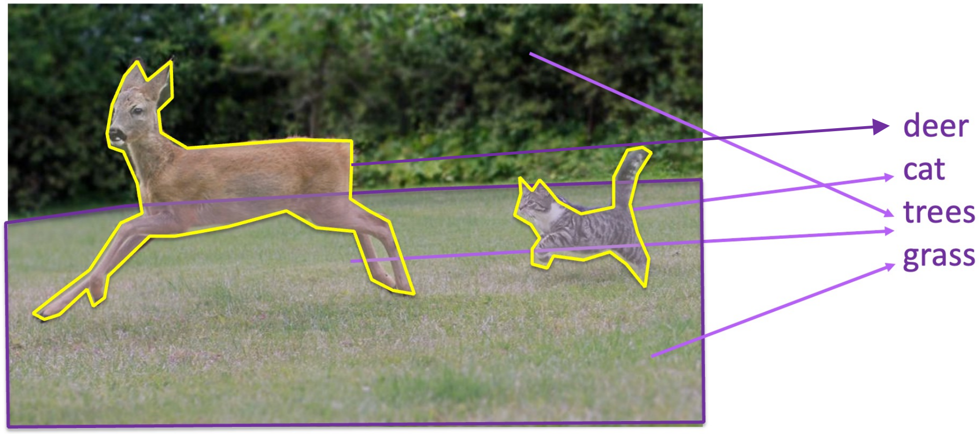
Kristen Grauman

Object Recognition



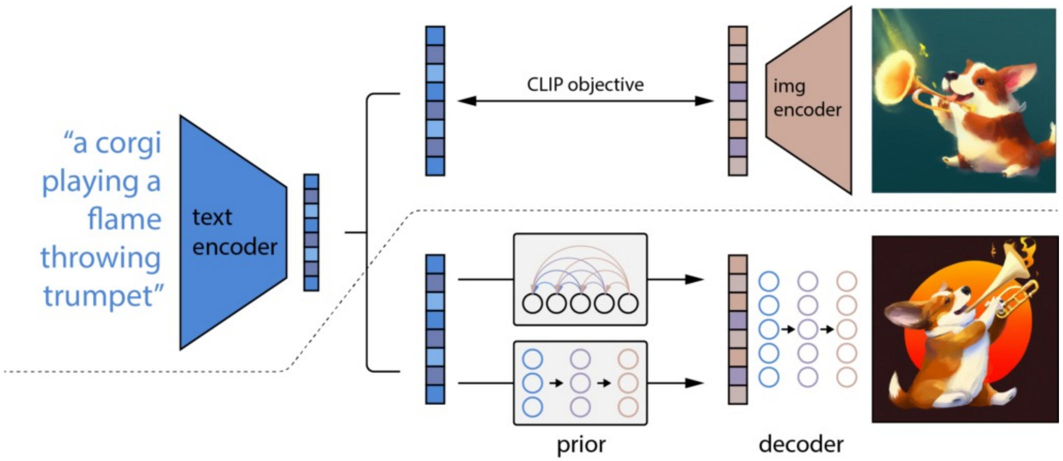
Adapted from Vicente Ordoñez

Image Segmentation

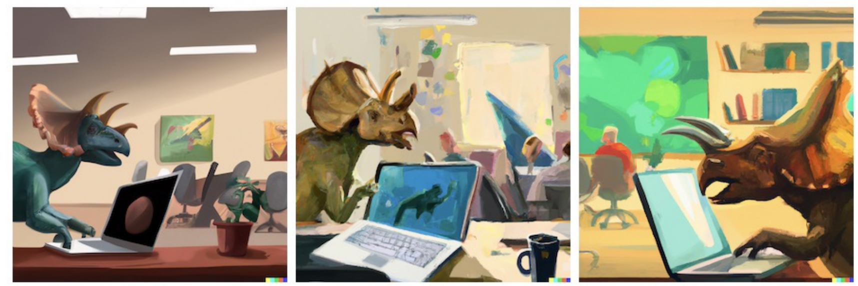


Adapted from Vicente Ordoñez

Generative AI



Dall.e 2: https://learnopencv.com/mastering-dall-e-2/



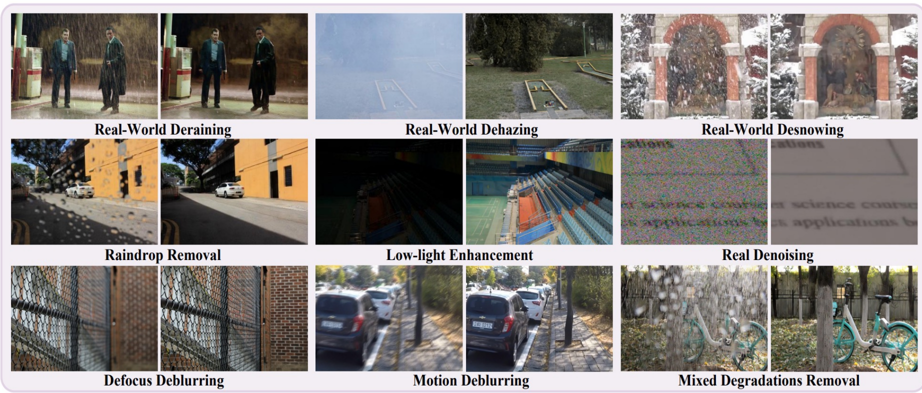
Stable Diffusion: "Triceratops programming on a MacBook in a startup office"

"a man with red hair"

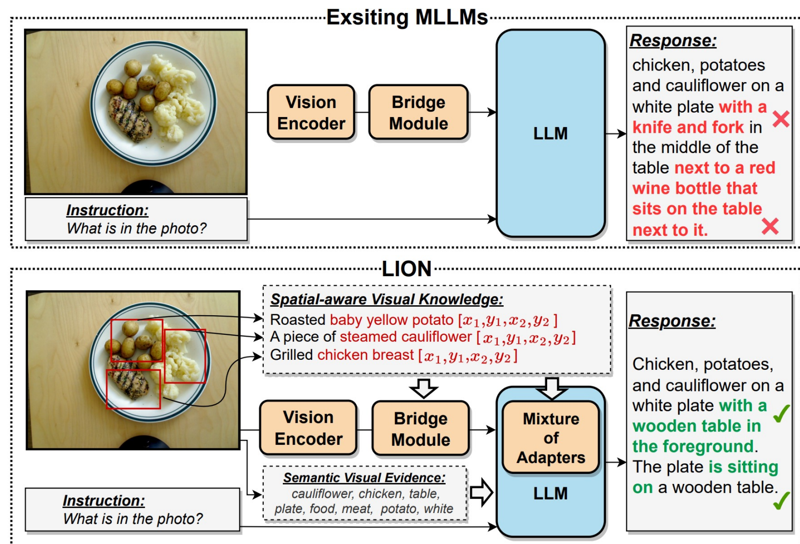
"a vase of flowers"

Text-conditional image-inpainting [ref]

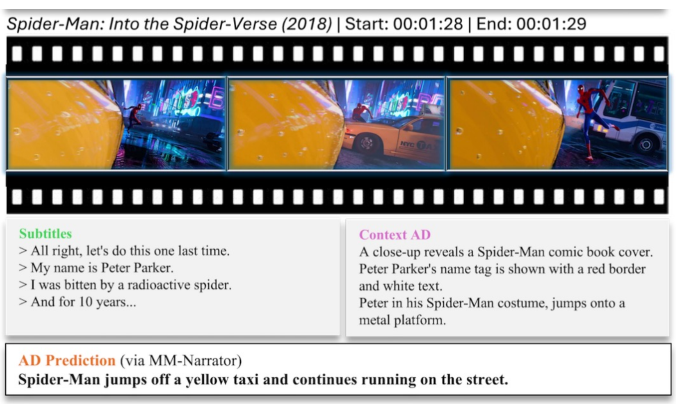
Multimodal Generative AI



Multimodal Prompt Perceiver: Empower Adaptiveness, Generalizability and Fidelity for All-in-One Image Restoration [CVPR]

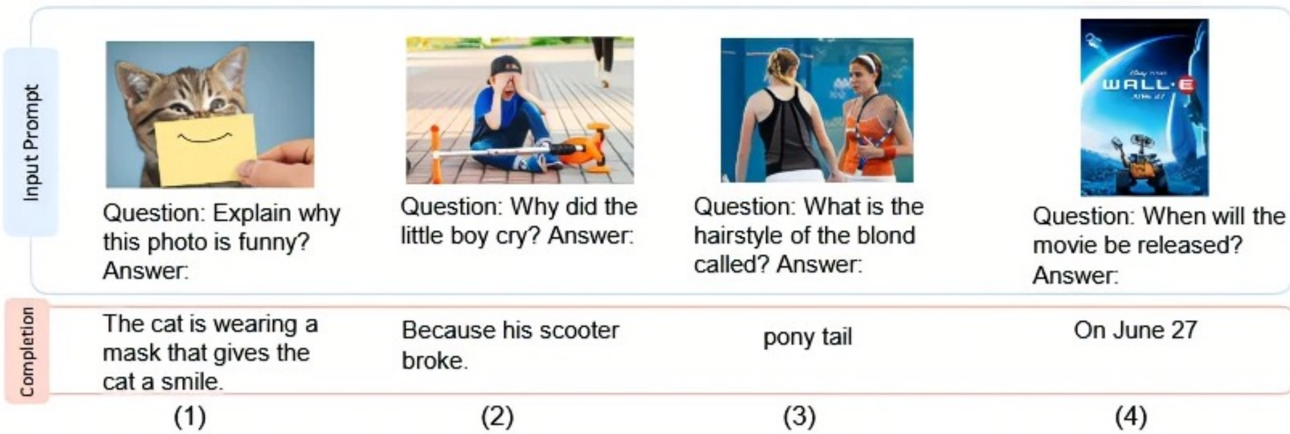


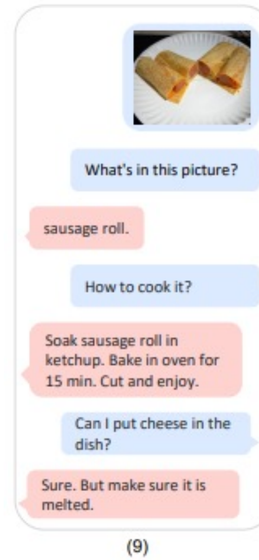
LION : Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge [CVPR]



MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning [CVPR]

Large Language Models





Gpt-4: https://medium.com/@amol-wagh/whats-new-in-gpt-4-an-overview-of-the-gpt-4-architecture-and-capabilities-of-next-generation-ai-900c445d5ffe