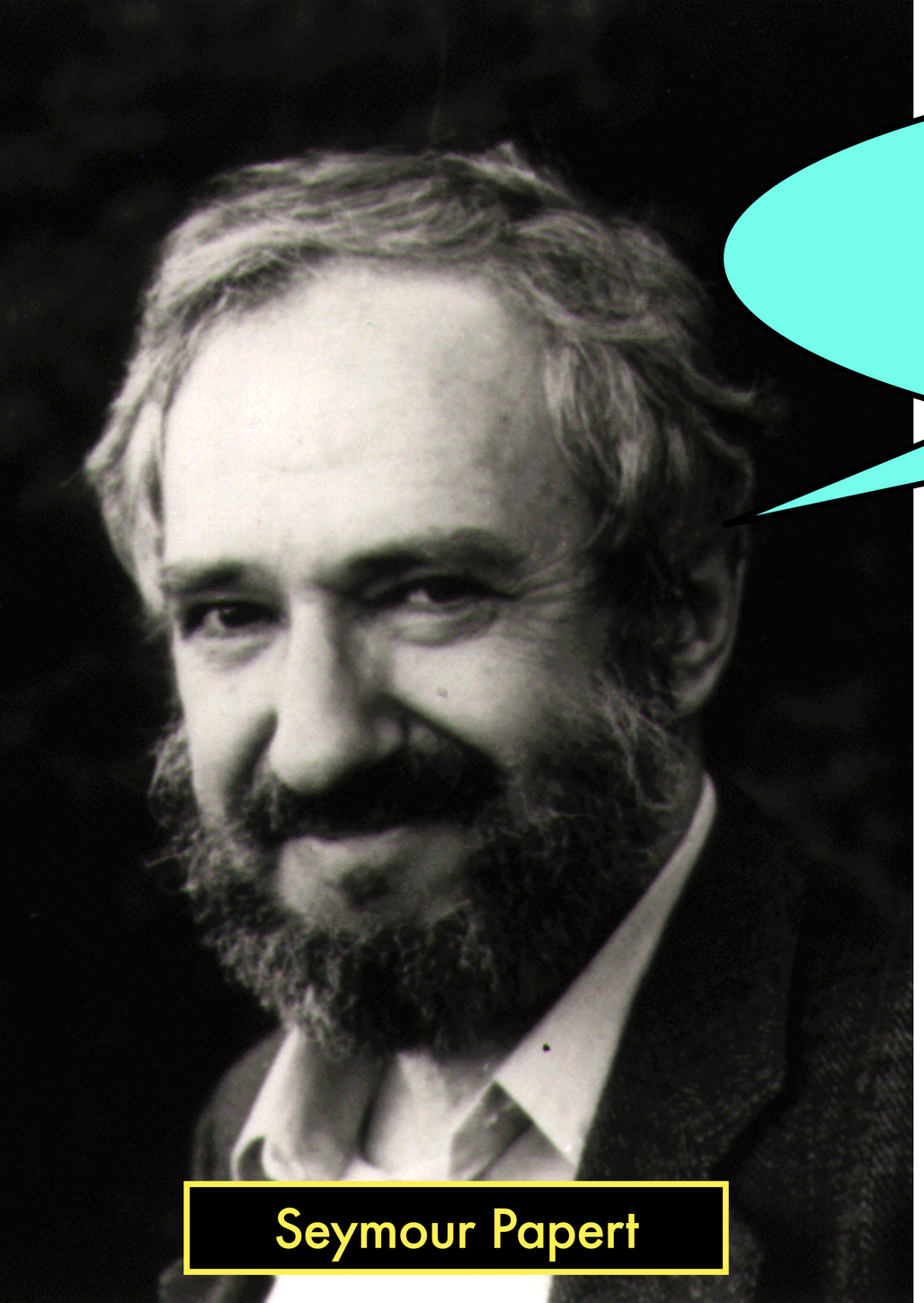


Seymour Papert

"The *summer vision project* is an attempt to use our summer workers... in the construction a *significant part of a visual system*" (1966)

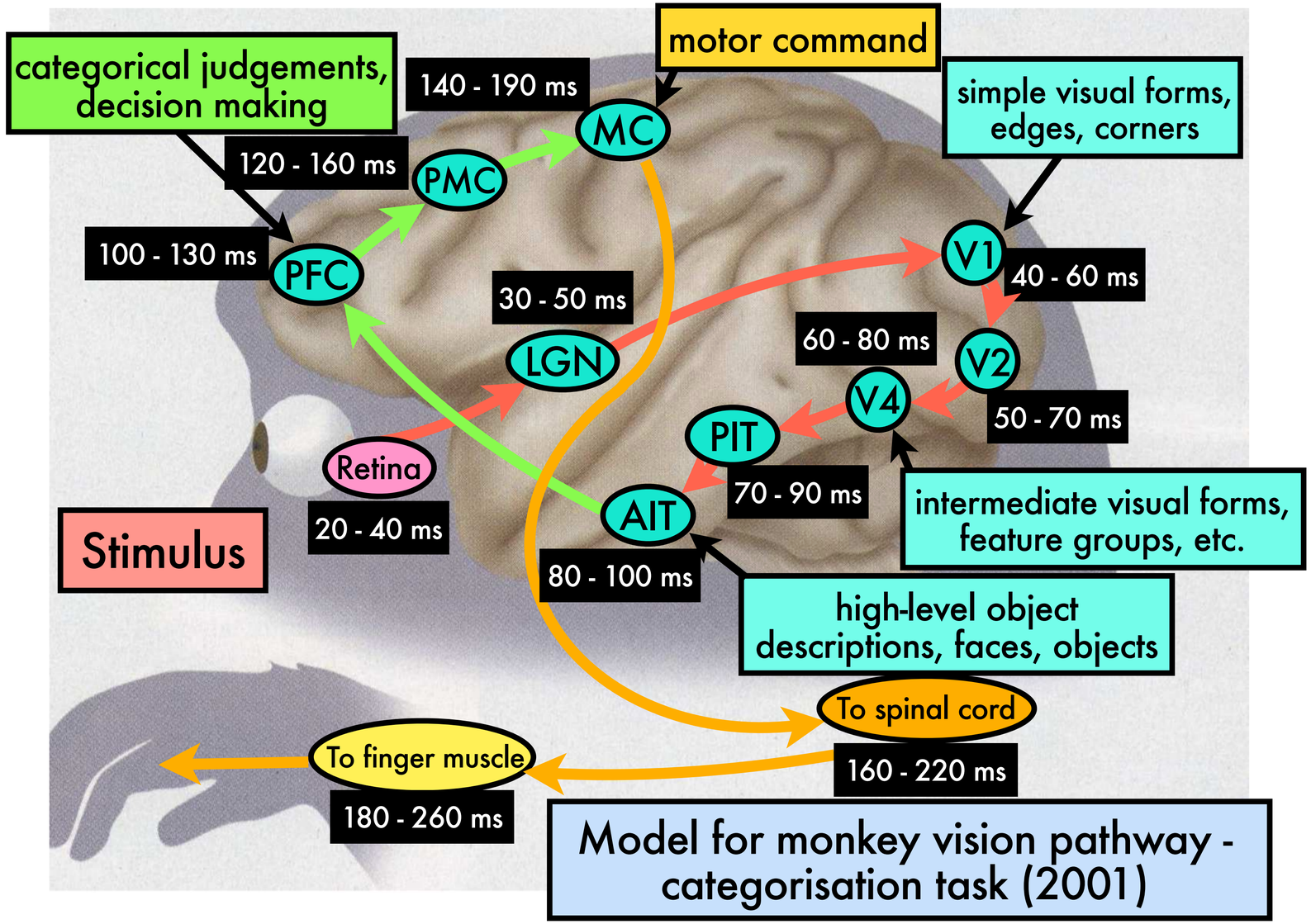
Computer Vision

getting computers to "see"

- object detection
- pose estimation
- ....
- Tasks

What Is Computer Vision

A field that takes inspiration from capabilities of biological vision



Model for monkey vision pathway - categorisation task (2001)

Computer Vision

Human vision is a remarkable thing

Goal: replicate this sense with a computer and a camera

This turns out to be very difficult

Vision is an *inverse problem*: Deduce 3D scene structure, objects and properties from 2D observations (images/video)

Note: biological vision is complex

We'll steal ideas, but we don't need to replicate implementation details

References/Notes/Image credits:
(human vision figure) S. J. Thorpe and M. Fabre-Thorpe, "Seeking categories in the brain", *Science* (2001)

Motivation

What does it mean to "understand" an image or video?

What is this image about?

What kind of scene?

Where might it have been taken?

What objects does it contain?

How many objects?

How are they interacting?

Can we describe it in words?

Image credit: "The Beginning of the Great Migration - Masai Mara, Kenya" by james_hammond (CC BY-ND 2.0)
https://www.flickr.com/photos/36922696@N04/8732132809

Computer Vision - But What Is It Really?

There is general agreement that computer vision involves computers and image data

Beyond this, there isn't a single universally accepted framework/taxonomy

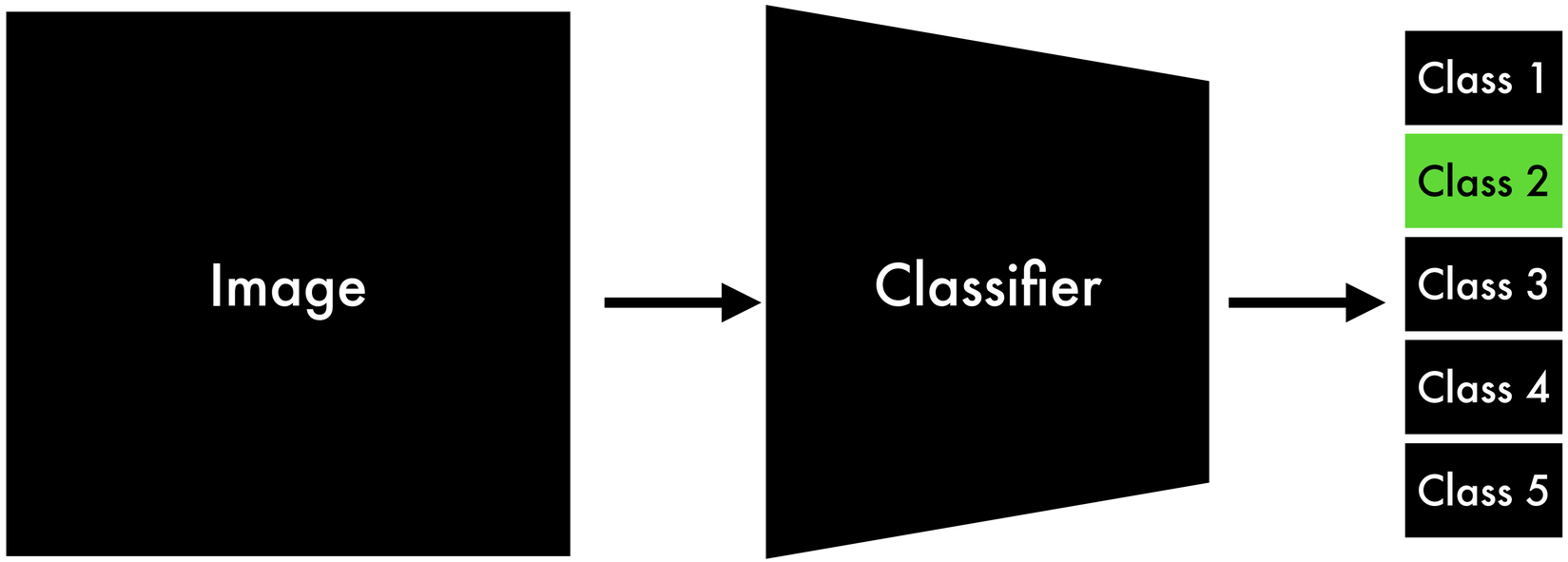
Instead, there is a diverse array of related tasks that are studied from different perspectives since:

- Highly interdisciplinary field: neuroscience | machine learning | psychology | signal processing ...
- the technology is evolving at an absurd pace, so the tasks that can be tackled are also evolving

Let's look at some tasks

Image Classification
Task definition

Objective: assign a class label to the *whole image*



Note: class labels form a finite, discrete set

Image Classification
Example

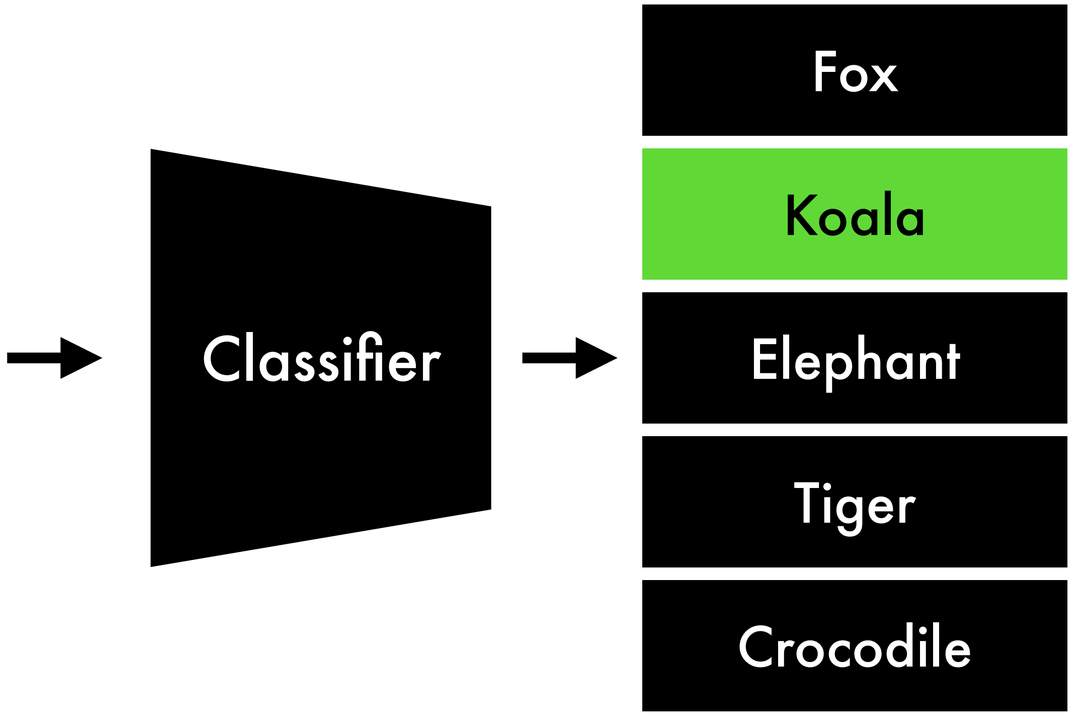


Image credit: "Koala" by jcoterhals (CC BY-NC-ND 2.0),
https://www.flickr.com/photos/28745942@N05/4485682479

Image Retrieval
Task definition

Objective: rank a pool of images according to how well they match a query

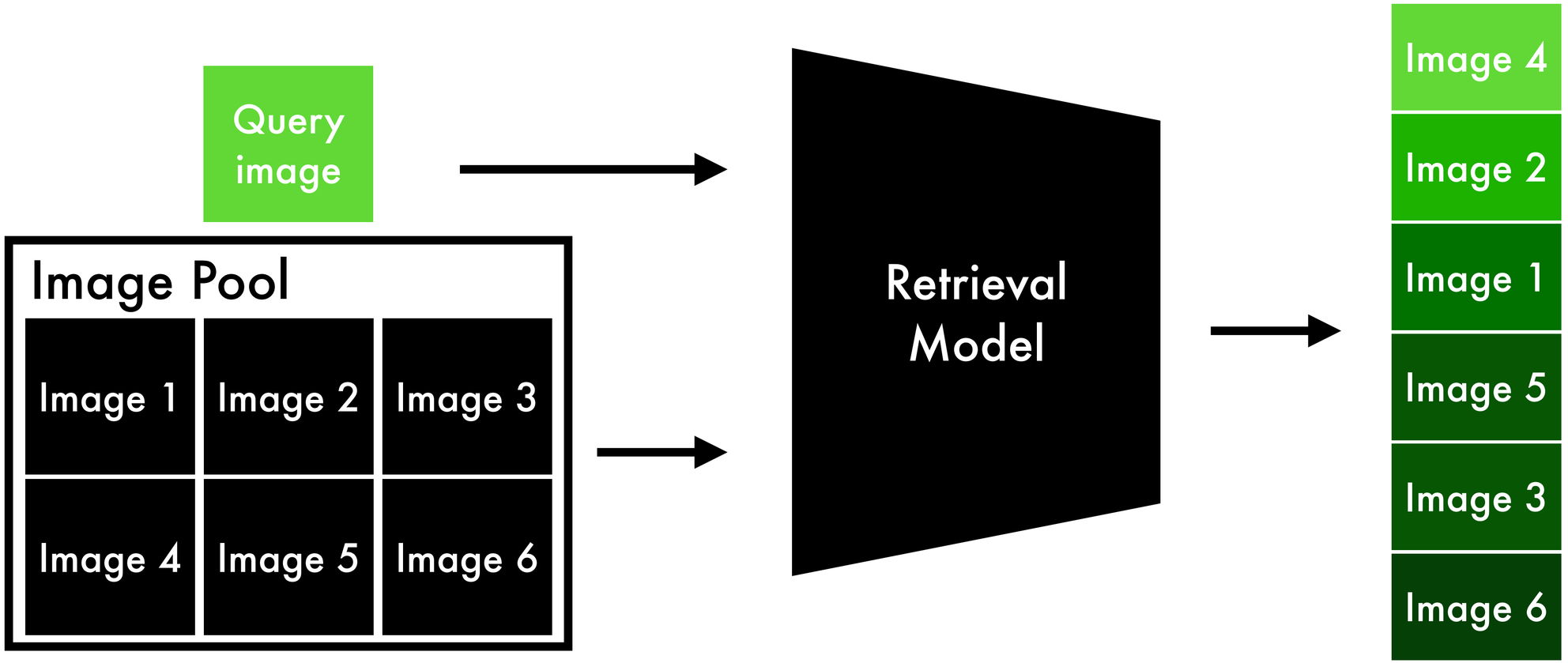
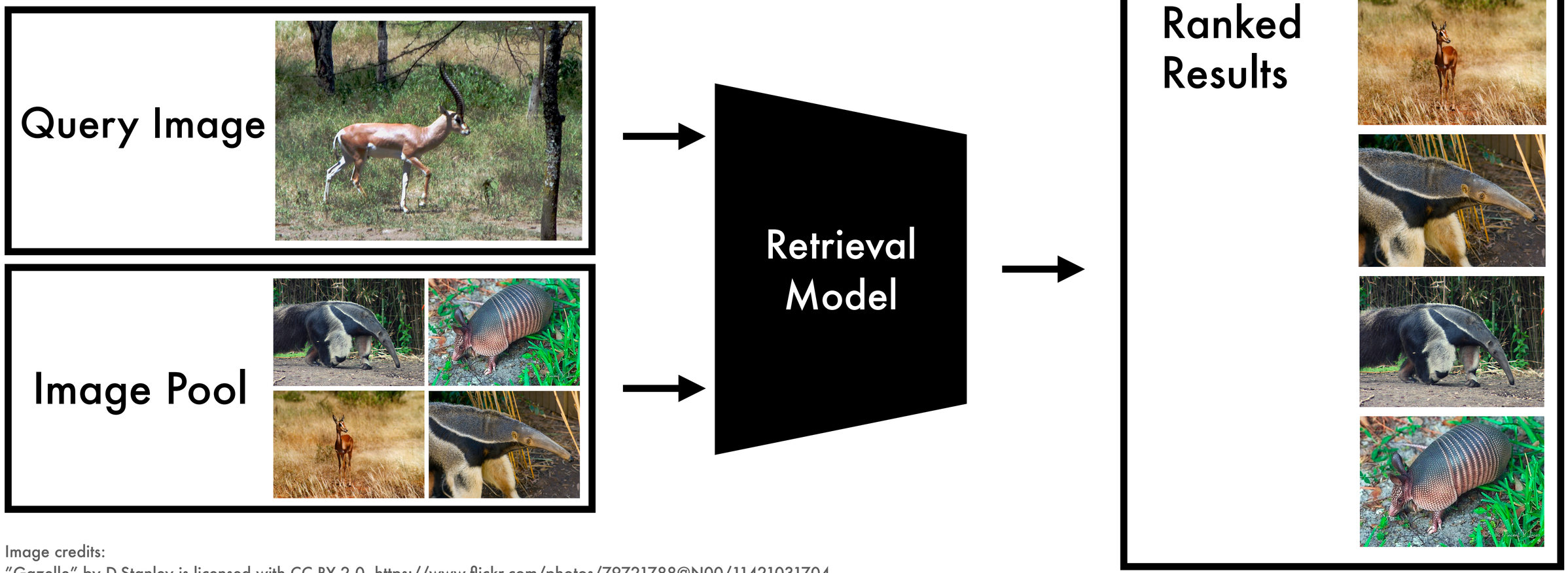
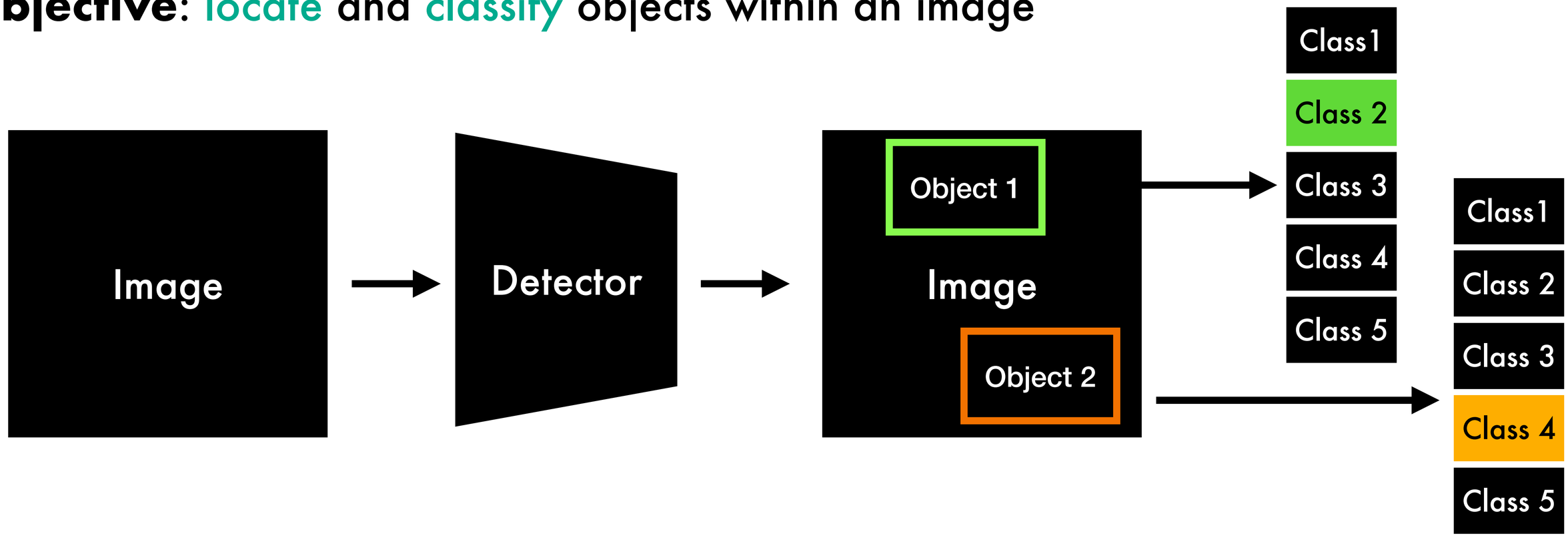


Image Retrieval
Example



Object Detection
Task definition

Objective: locate and classify objects within an image



Object Detection
Example

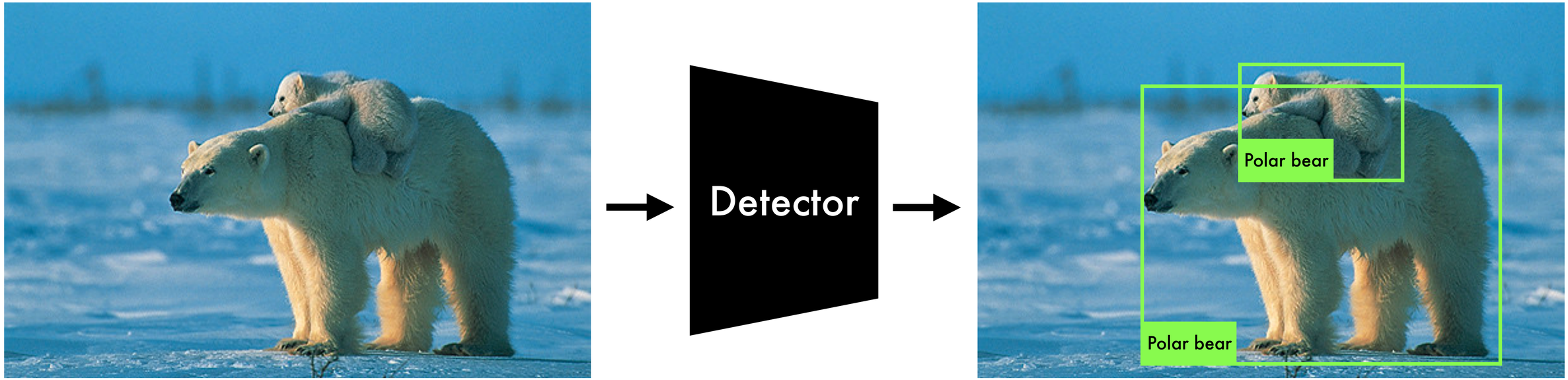
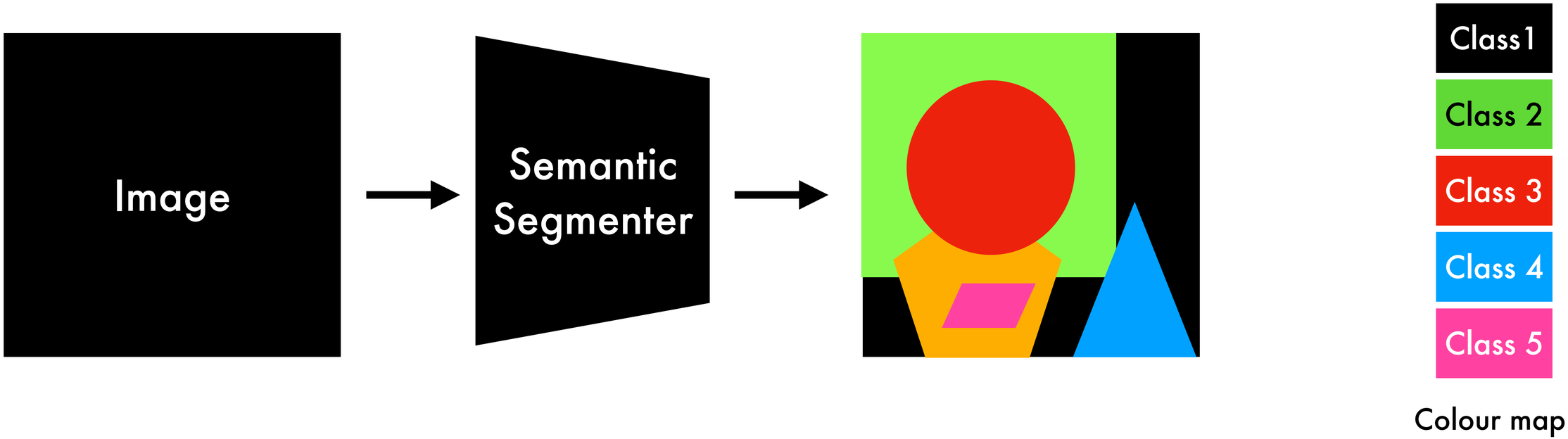


Image credits: National Geographic Channel, 2004

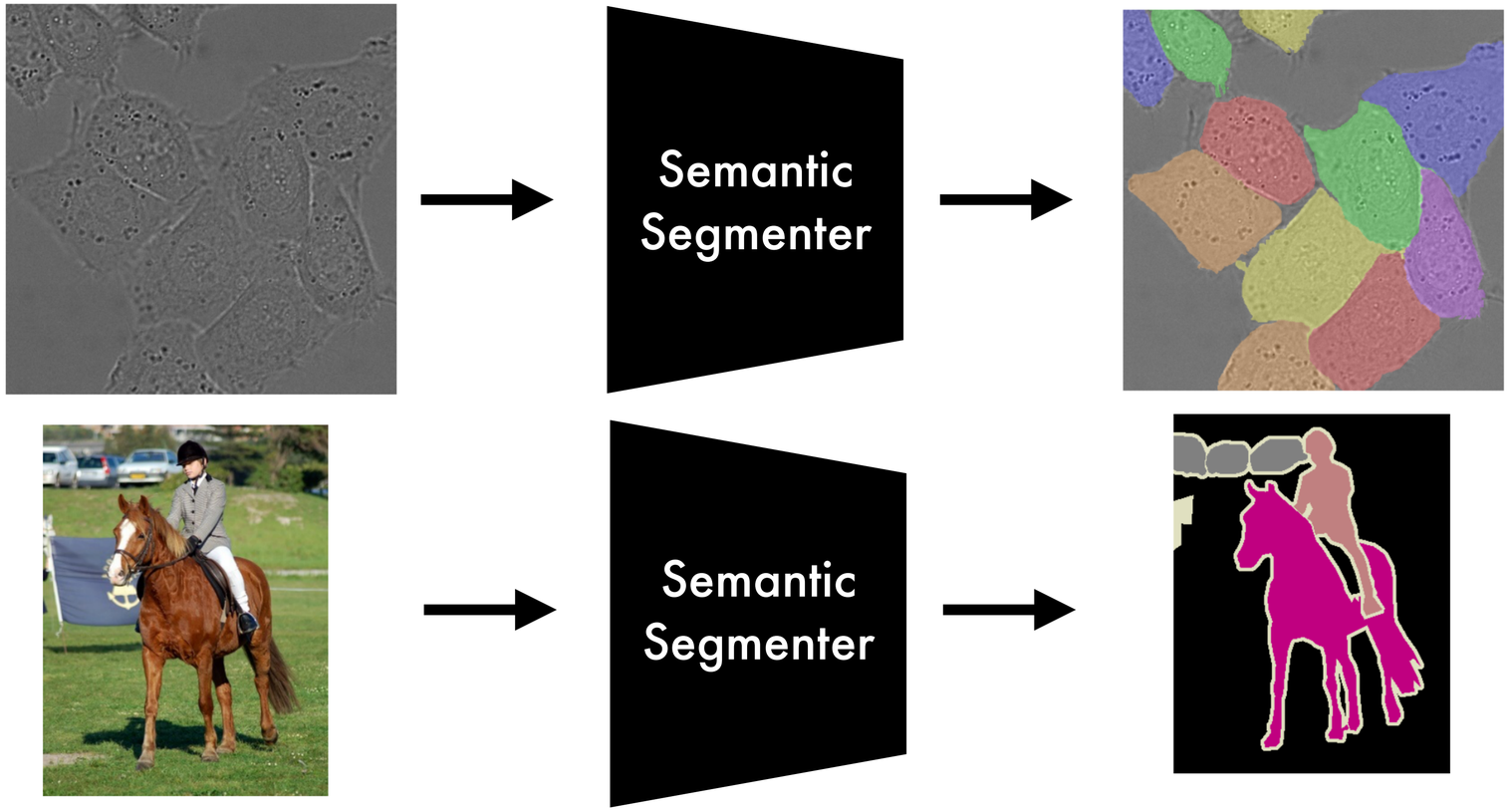
Semantic Segmentation

Task definition

Objective: Assign a class label to every pixel location in the image



Semantic Segmentation
Example

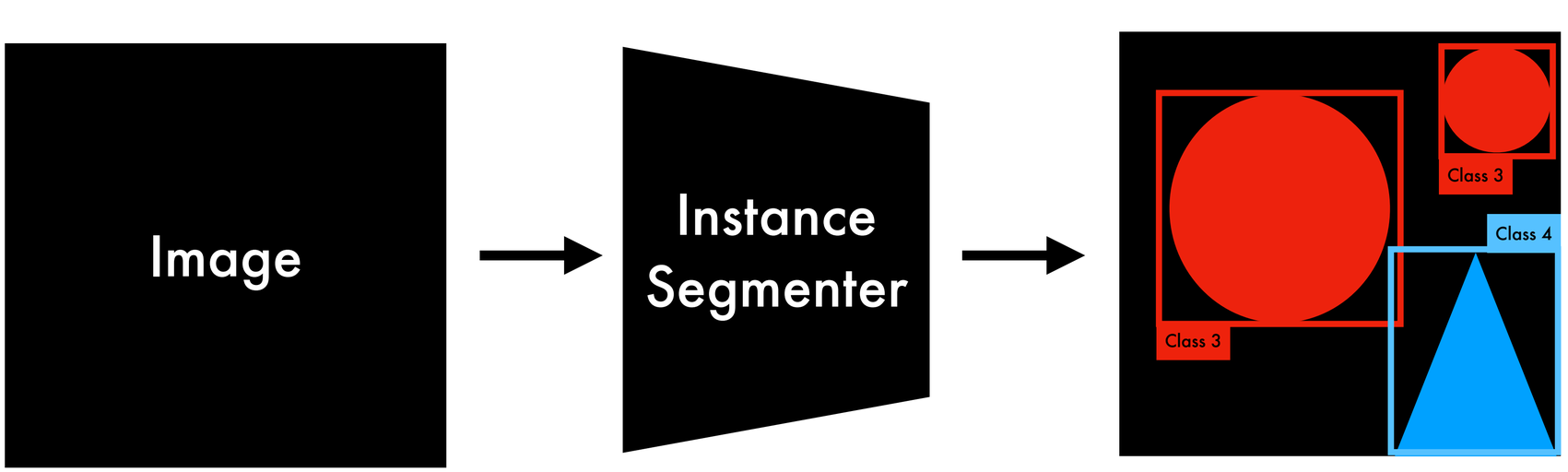


Example images from Fig 6 in J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", *CVPR* (2015)
O. Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation", MICCAI (2015)

Instance Segmentation

Task definition

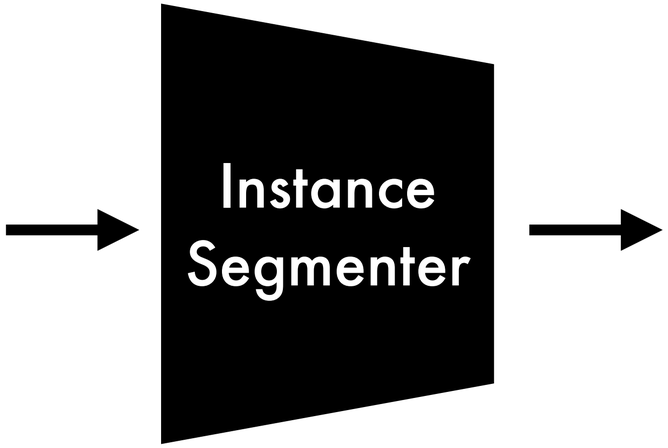
Objective: Detect, classify and segment objects in the image

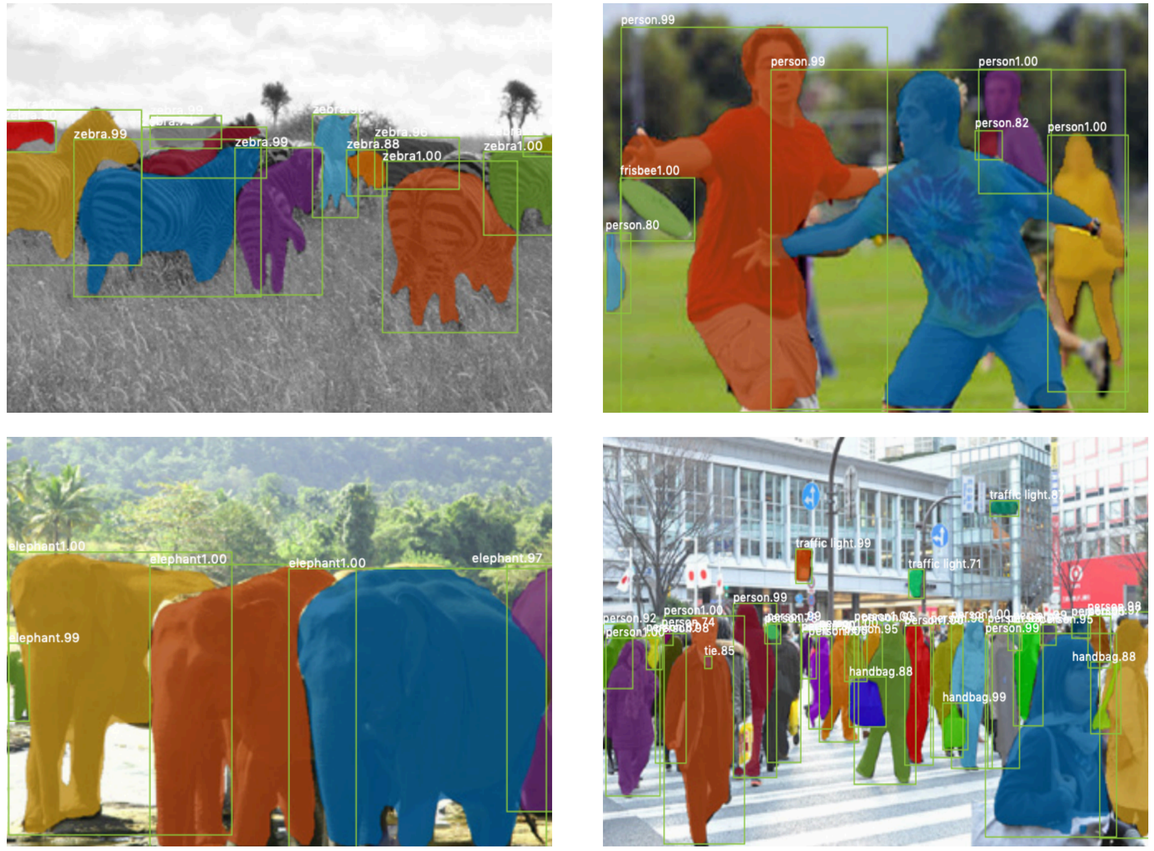


Colour map

Instance Segmentation

Example



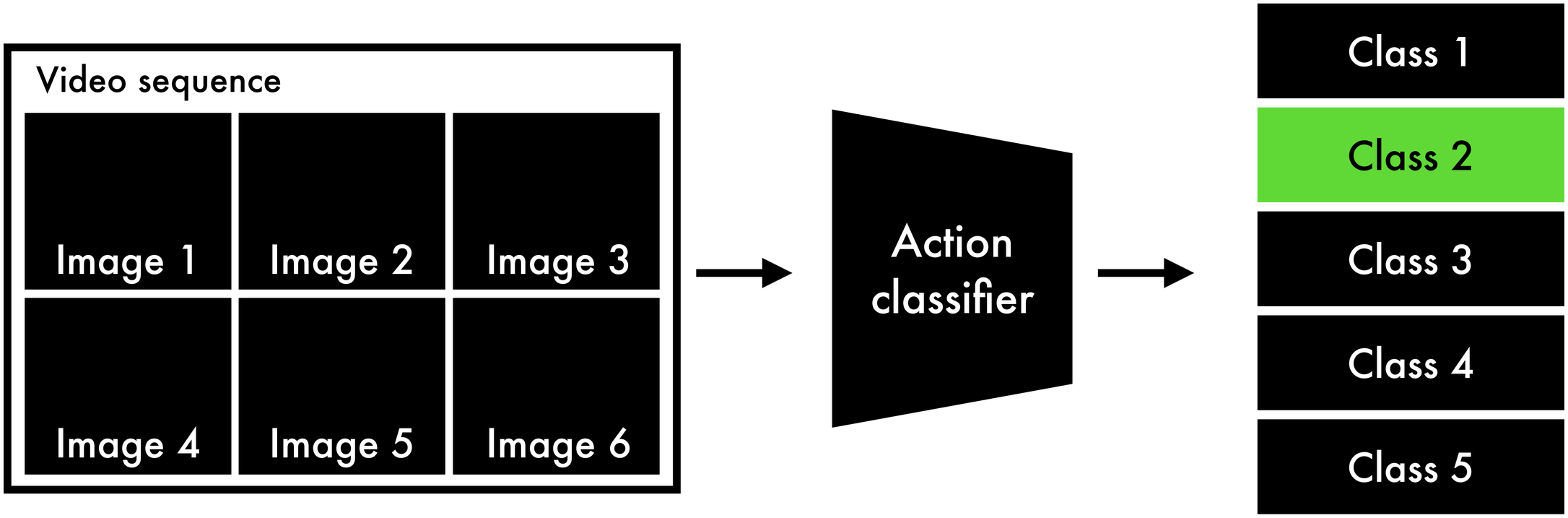


Examples from: K. He et al., "Mask r-cnn", ICCV (2017)

Action Recognition In Video

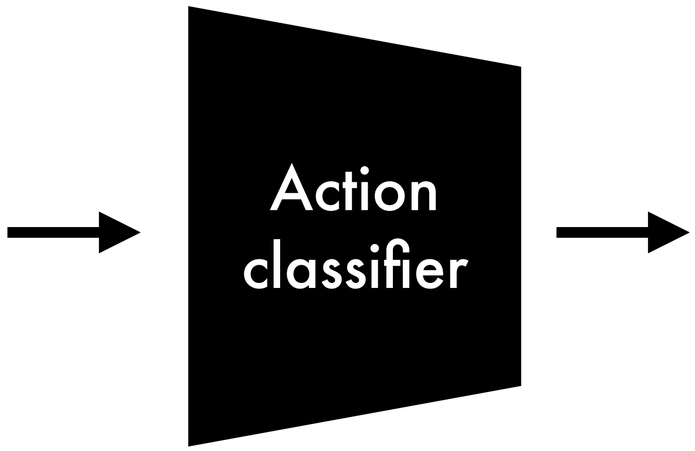
Task definition

Objective: Given a video sequence, predict the dominant action



Action Recognition In Video

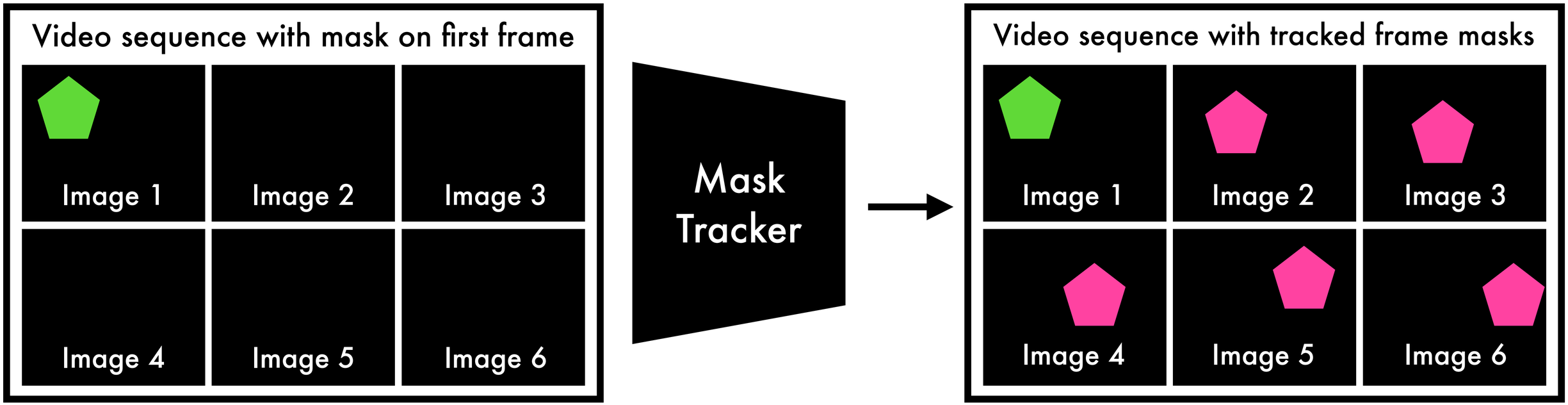
Example



Mask Tracking On Video

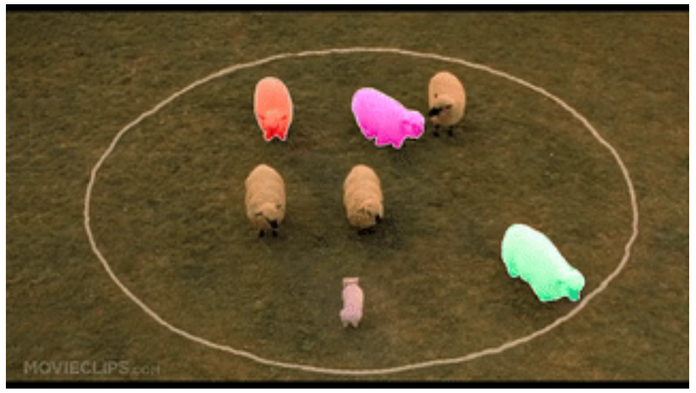
Task definition

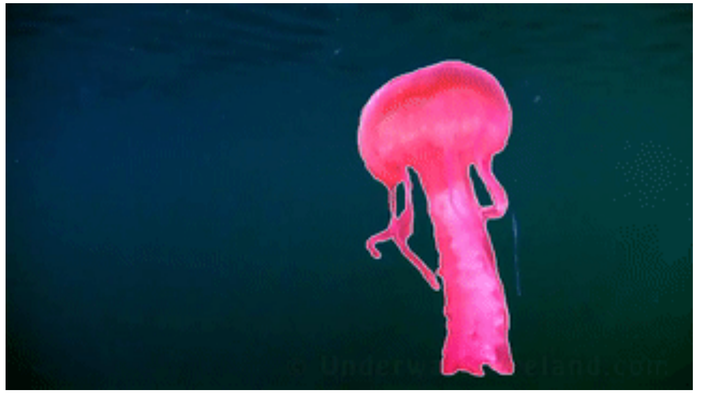
Objective: Given a mask region covering an object, update the mask to ensure it keeps covering the object across frames

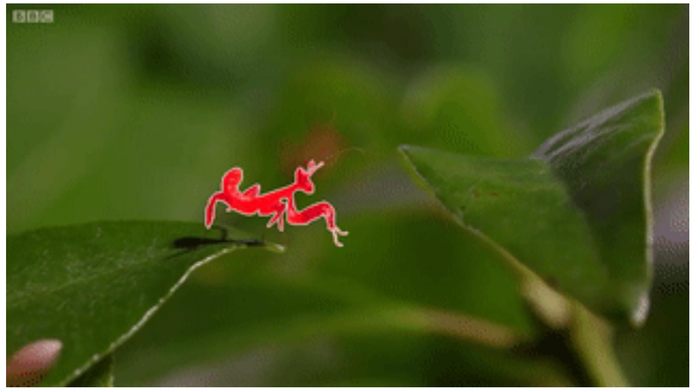


Mask Tracking On Video

Example



Video sequences: DAVIS-2017 Video Segmentation and Youtube-VOS 2018 Video Segmentation
Method: Z. Lai et al., "MAST: A Memory-Augmented Self-supervised Tracker", CVPR (2020)

Image Captioning
Task definition

Objective: provide an accurate text description of an image

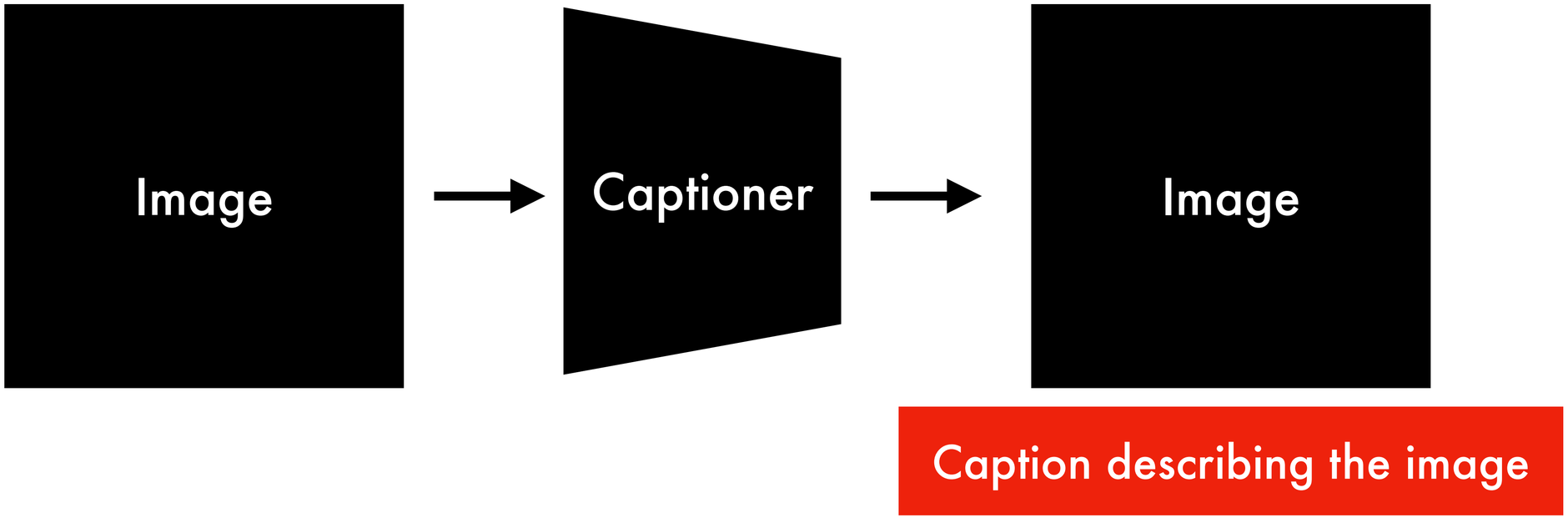
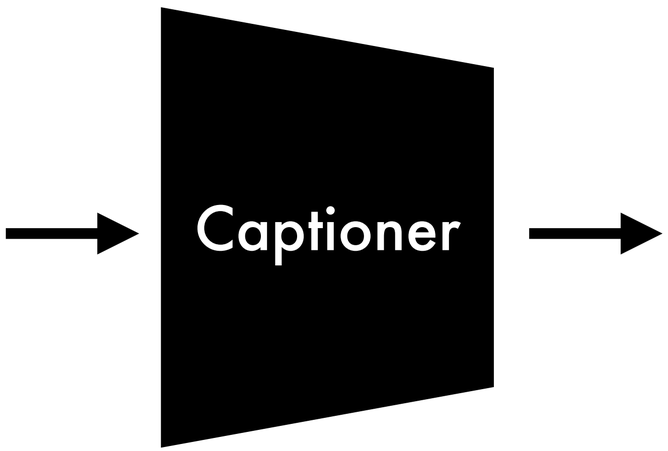


Image Captioning
Example



The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Image credit: Fig 1 from X. Chen et al., "Microsoft coco captions: Data collection and evaluation server", arxiv (2015)

Computer Vision - A Whirlwind Tour Of Further Themes

Image processing

De-noising 

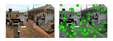
Super-resolution 

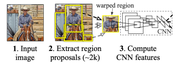
Inpainting 

Outpainting 

Feature extraction

Feature detection 

Hand-crafted descriptors 

Deep features 

Building blocks for other tasks

Fine-grained estimation

Pose estimation 

Motion analysis/tracking

Optical flow 

References/image credits:
https://en.wikipedia.org/wiki/Total_variation_denoising#/media/File:ROF_Denoising_Example.png
https://en.wikipedia.org/wiki/Inpainting#/media/File:Restoration.jpg
https://en.wikipedia.org/wiki/File:Super-resolution_example_closeup.png
https://openai.com/blog/dall-e-introducing-outpainting
Edge detection,JonMcLoone at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=44894482
A. Vedaldi, SIFT tutorial https://www.vlfeat.org/overview/sift.html
R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR (2014)
(Image source for pose) https://colab.research.google.com/drive/1yM-K03ZvSiEi8Pudj5Vj_pWiZ6GI__y#scrollTo=2ZfPn__3He6I
D. Fleet et al., "Optical flow estimation", *Handbook of mathematical models in computer vision* (2006)

Does it end there?

Computer Vision - A Whirlwind Tour Of Further Themes (Cont.)

3D vision/scene analysis

Depth estimation 

3D reconstruction 

High-level vision tasks

Visual question answering/captioning 

Emotion recognition 

Image synthesis

Text-to-image 

Style transfer 

Video synthesis

TODO for research community (active)

References:
(Depth estimation samples) https://paperswithcode.com/task/monocular-depth-estimation
(3D reconstruction) https://github.com/autonomousvision/convolutional_occupancy_networks
(Visual question answering) J. Li et al., "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", arxiv (2023)
(Emotion recognition) P. Burkert et al., "Dexpression: Deep convolutional neural network for expression recognition", *arxiv* (2015)
(Annoyed cat) Generated by Samuel Albanie with DALL-E
(Style transfer) https://godatadriven.com/blog/how-to-style-transfer-your-own-images/