



ACV – Applied Computer Vision

Bachelor Medientechnik & Creative Computing

Matthias Zeppelzauer
matthias.zeppelzauer@fhstp.ac.at

Djordje Slijepcevic
djordje.slijepcevic@fhstp.ac.at

Applied Computer Vision - Goals

General goal:

Learning important basics and applications of Computer Vision

and apply them in practice

But more importantly:

Having fun with Computer Vision!

Organisation

- Lecturers:
 - Matthias Zeppelzauer (lecture)
 - Djordje Slijepcevic (lecture)
 - Eric Stumpe (labs)
 - Adrian Böck (labs)
- Course page (Ecampus): <https://ecampus.fhstp.ac.at/course/view.php?id=32459>
- Lecture planned in presence. In case this is not possible -> [MS Teams](#)
- **Use eCampus forum for questions** – additional benefit for colleagues
- All course materials slides, codes, etc. on eCampus
- All assignments on eCampus



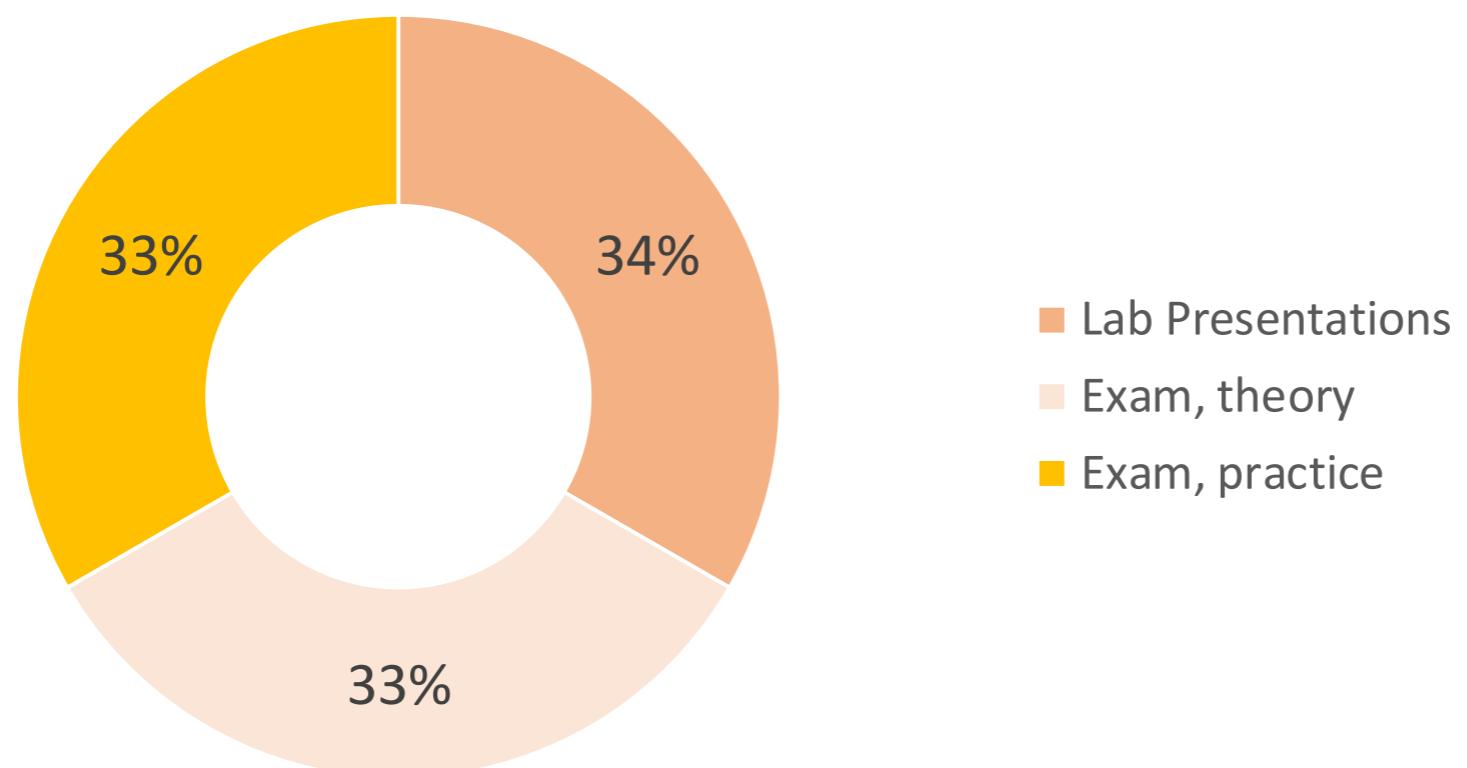
Organisation (cont.)

- 5 ECTS = $5 \times 25\text{h} = 125\text{h}$
 - Presence: 45 Lecture Units (LE)
 - Homework: 80h
- Contents:
 - Part I: Theory
 - Part II: Practice: small labs from week to week
- Ask questions – anytime!



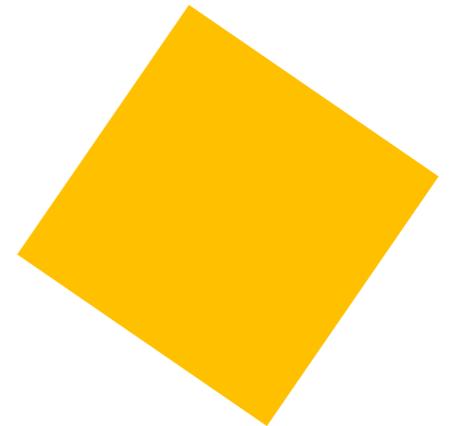
Organisation (cont.)

- Grading:
 - Exam: theory + practice, each 33% (end of semester)
 - Lab presentations, 33% (weekly)



Temporal Schedule

- Blocks of 3 lecture units per week
 - almost every week
 - sometimes 2 lectures per week (will be considered for lab deadlines)
- 2 units for theory
- 1 unit for practice (labs)
 - present results of labs
 - practice what you have learned in theory
 - Discuss next lab
 - room for questions / problem solving

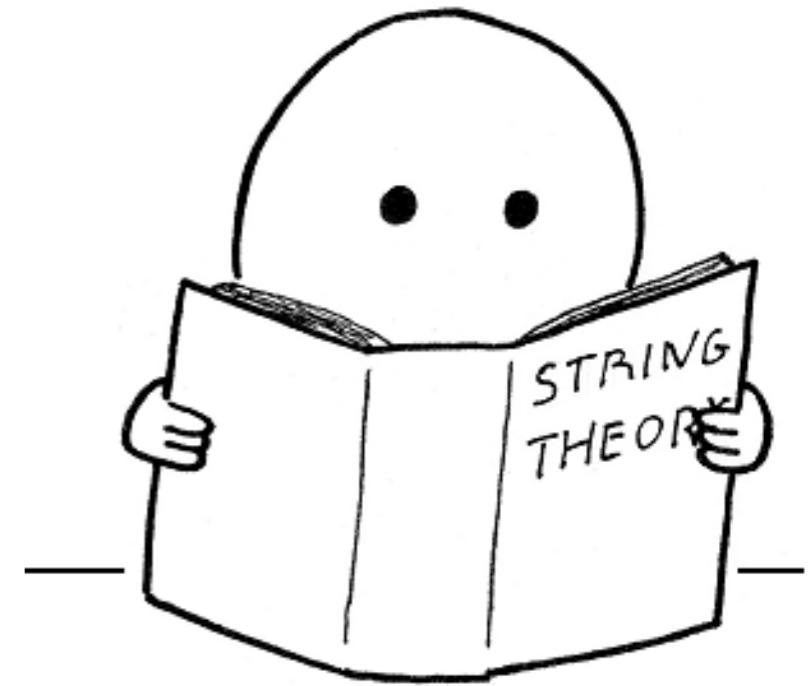


Lecture Organization

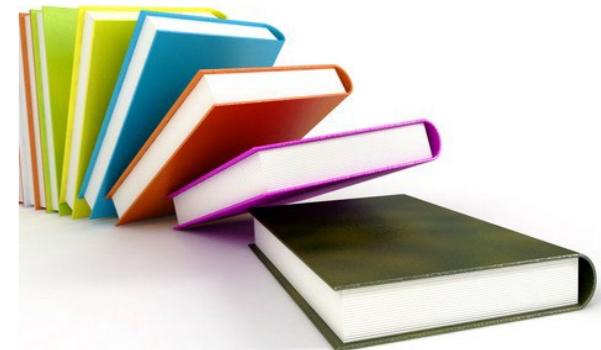
- Target platforms: eCampus + Microsoft Teams
- **Attendance is compulsory**
- Lecture sessions will not be recorded (recording of a previous year exists, only in German)
- **Basic Rules:**
 - if you are not sure about asking, then ask! → keep course interactive
 - If you loose the “red ribbon” -> interrupt and tell me (usually others have lost it then too)

Prerequisites

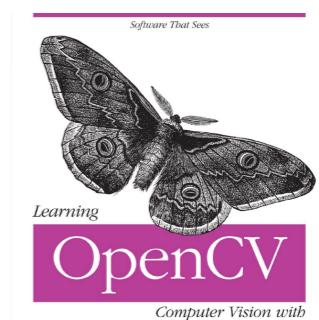
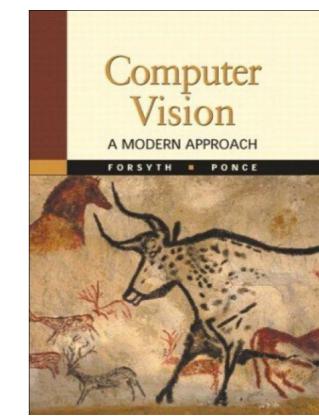
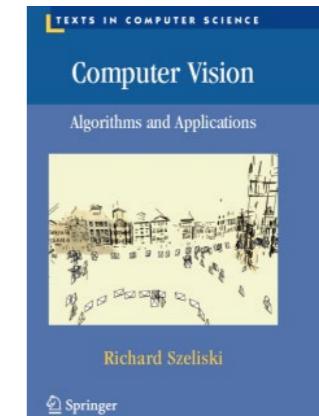
- Basic programming skills
- Reading English literature
- Interest in Image Processing and Computer Vision
- Willingness to understand what's going on there
- **No** programming experience in image processing and computer vision necessary → we will learn this!



Course Materials

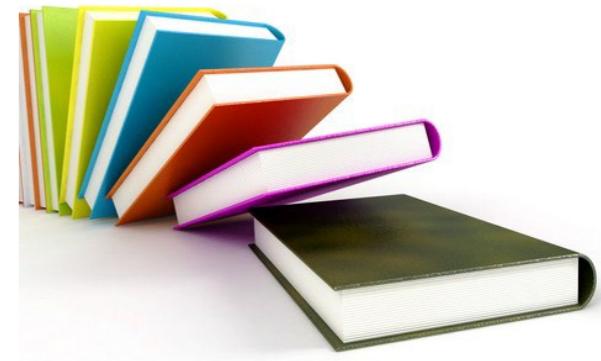


- Slides (eCampus)
- Textbooks
 - Richard Szeliski: "Computer Vision: Algorithms and Applications", available online: <http://szeliski.org/Book/>
 - Forsyth & Ponce: "Computer Vision – A Modern Approach", available online:
<https://github.com/avinashk94/CVIP/blob/master/David%20A.%20Forsyth%2C%20Jean%20Ponce%20Computer%20vision%20A%20modern%20approach.pdf>
 - Bradski & Kaehler: "Learning OpenCV": online:
<http://www.bogotobogo.com/cplusplus/files/OReilly%20Learning%20OpenCV.pdf> (most interesting are the theory parts on algorithms)
 - LinkedIn Learning: <https://www.linkedin.com/learning> (Login = FH email address)



O'REILLY® Gary Bradski & Adrian Kaehler

Course Materials



- Online resources
 - Deep Learning Book (HTML Version is free):
<https://www.deeplearningbook.org/>
 - Intro to neural networks:
<http://neuralnetworksanddeeplearning.com/chap1.html>,
 - Intro to CNNs (entire recorded Stanford course):
https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv&ab_channel=StanfordUniversitySchoolofEngineering.

Overall Goal of Lecture

Understand:

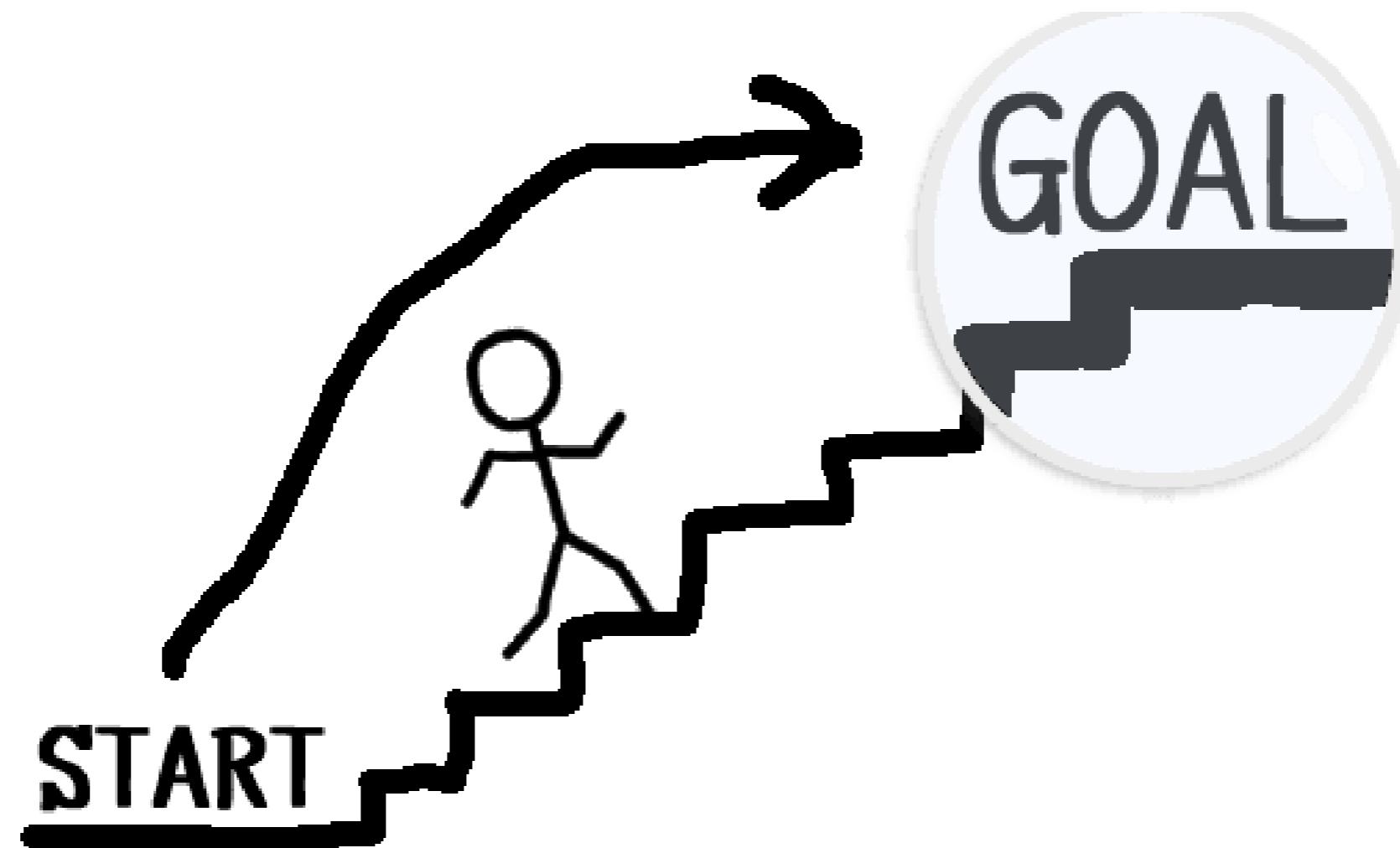
basics of image processing and computer vision

and

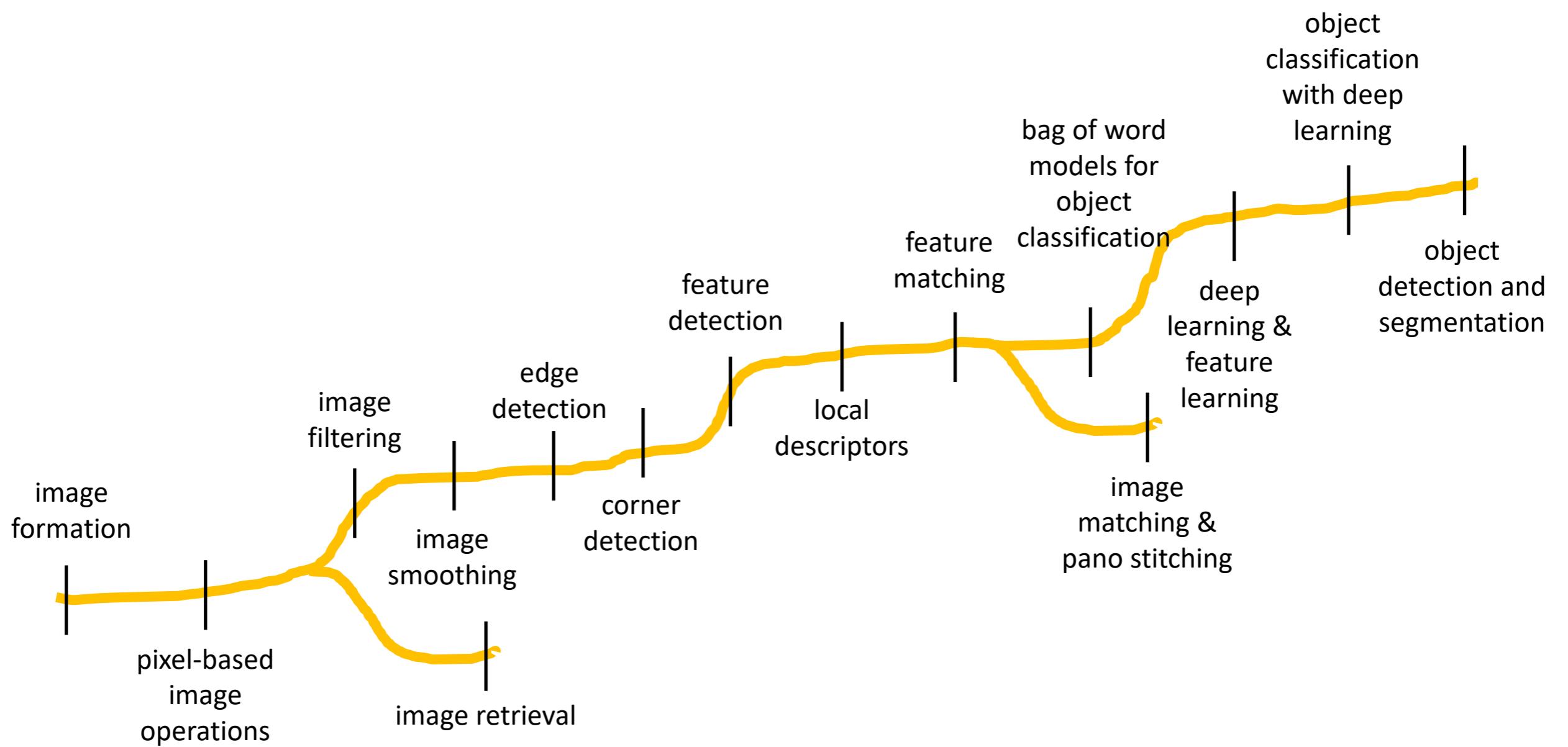
how object classification & recognition in images works

→ in theory and in practice!

Step-by-Step



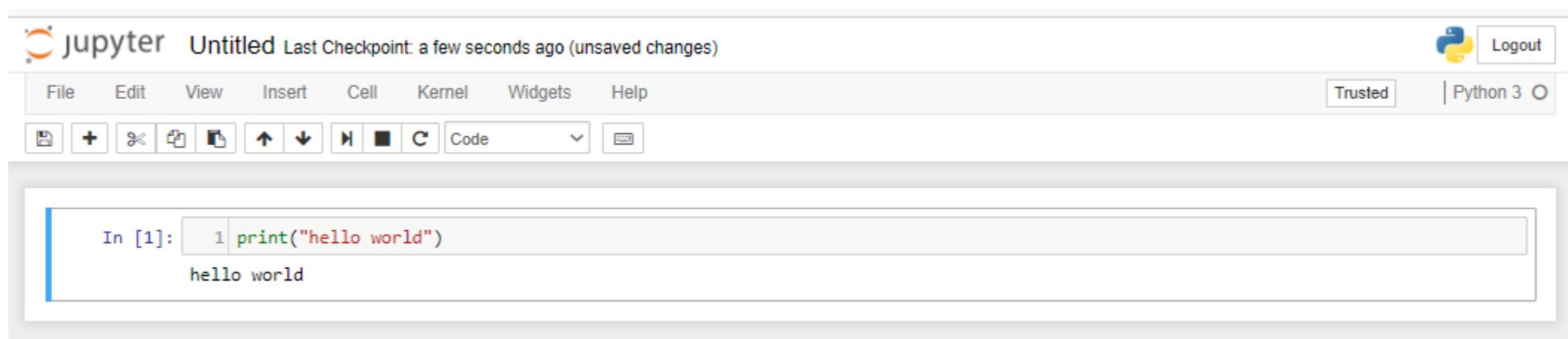
Roadmap



How to solve it in practice?



- Computer Vision == Python
- Python is a powerful script language
- Real-time execution, no compiling necessary
- Perfect for experimentation
- Python provides libraries for almost everything we need including Deep Learning!
- We learn it step-by-step and will keep it at a simple level
- Tool for labs: Jupyter Notebook



And now: who are you?

- And what do you expect to get out of this class?
- Previous experience in programming, vision, learning, graphics?
- Special Interest areas? “What I always wanted to know...”



Questions before we start?

What is Computer Vision?

Credits

- Antonio Torralba, CSAIL, MIT
- Trevor Darrel, Berkeley
- James Hays, Brown University
- Kristen Grauman, University of Texas
- Steve Seitz, University of Washington
- Richard Szeliski, University of Washington
- and many more...

The importance of images

Some images are more important than others, e.g.:



“Dora Maar au Chat”
Pablo Picasso, 1941

100 million \$



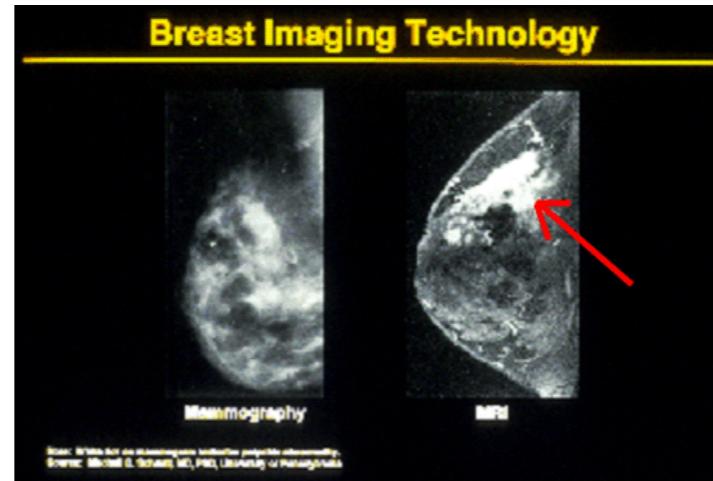
**Visual information is
everywhere...**

**...and of high
importance to us all**

Application areas



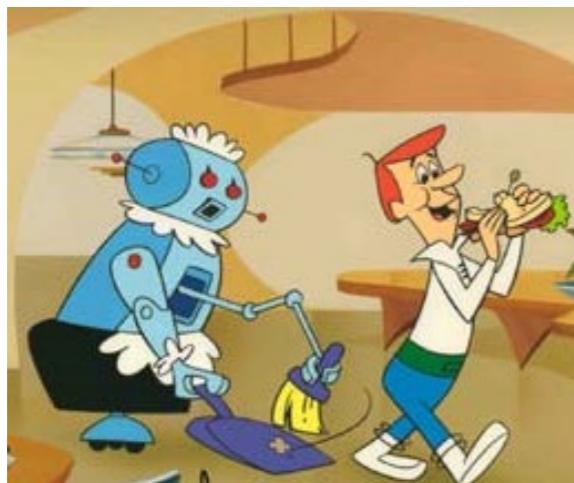
Safety



Health



Security



Comfort



Fun

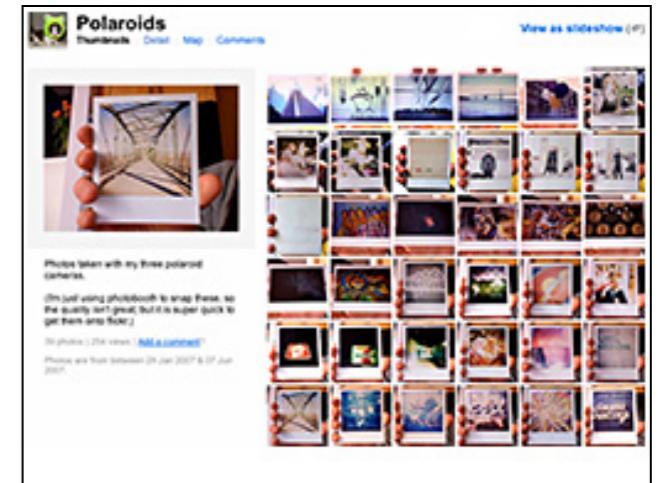


Photo Collections

Computer Vision

- Teaching computer to see!
- What does it mean, to see? “to know what is where by looking”.
- “How to discover from images what is present in the world, where things are, what actions are taking place” [Marr 1982].
- Develop algorithms and representations to allow a machine to recognize objects, people, scenes, and activities.
- “Deals with the development of the **theoretical** and **algorithmic** basis by which useful information about the 3D world can be automatically extracted and analyzed from a **single** or **multiple** 2D images of the world.” [unknown]



What kind of scene?

Where are the cars?

How far is the building?

...

Every picture tells a story

- Goal of computer vision is to write computer programs that can interpret images
- “Ein Bild sagt mehr als Tausend Worte”
- And that’s exactly the challenge!



Can computers match (or beat) human vision?

- Yes and no..
 - humans are much better at “hard” things
 - computers can be better at “easy” things
- But this is currently changing. Computers start to outperform humans even for “hard” tasks (e.g. face recognition)



La Gare Montparnasse, 1895

What does ChatGPT say..



Image Caption Generator

When Monday hits you like a runaway train. #MondayMood #HistoryInPhotos #TrainWreck
#VintageChaos #EpicFail



You

Explain what you see in this image



Image Caption Generator

This image depicts the historical train derailment that occurred at the Gare Montparnasse in Paris, France, in 1895. A steam locomotive has crashed through the station's exterior wall and is precariously hanging over the street below, with its nose plunged to the ground. Debris from the damaged wall is scattered around the accident site. This photograph captures a dramatic moment of transportation history, showcasing the aftermath of a significant train accident from the late 19th century.

Montparnasse derailment

[25 languages](#)

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#)

From Wikipedia, the free encyclopedia

The **Montparnasse derailment** occurred at 16:00 on 22 October 1895 when the Granville–Paris Express overran the buffer stop at its *Gare Montparnasse* terminus. With the train several minutes late and the driver trying to make up for lost time, it approached the station too fast and the driver's application of the *train air brake* was ineffective. After running through the buffer stop, the train crossed the station concourse and crashed through the station wall; the locomotive fell onto the Place de Rennes below, where it stood on its nose. Although the passengers survived, a woman in the street below was killed by falling masonry.

Derailment [\[edit\]](#)

On 22 October 1895, the Granville to Paris and Montparnasse express, operated by *Chemins de fer de l'Ouest*, was made up of steam locomotive No. 721 (a type 2-4-0, French notation 120) hauling three luggage vans, a post van, and six passenger coaches.^[1] The train had left *Granville* on time at 08:45, but was several minutes late as it approached its Paris Montparnasse terminus with 131 passengers on board. In an effort to make up lost time,^{[1][2]} the train approached the station faster than usual, at a speed of 40–60 km/h (25–37 mph), and when the driver attempted to apply the *Westinghouse air brake*, it was faulty or ineffective.^{[1][3][4]} The locomotive brakes alone were insufficient to stop the train, the momentum carried it into the buffers, and the locomotive crossed the almost 30-metre (98 ft) wide station concourse, crashing through a 60-centimetre (24 in) thick wall before falling onto the Place de Rennes 10 metres (33 ft) below, where it stood on its nose.



The wreckage of the station, photographed by Studio 51

Human perception has its shortcomings...



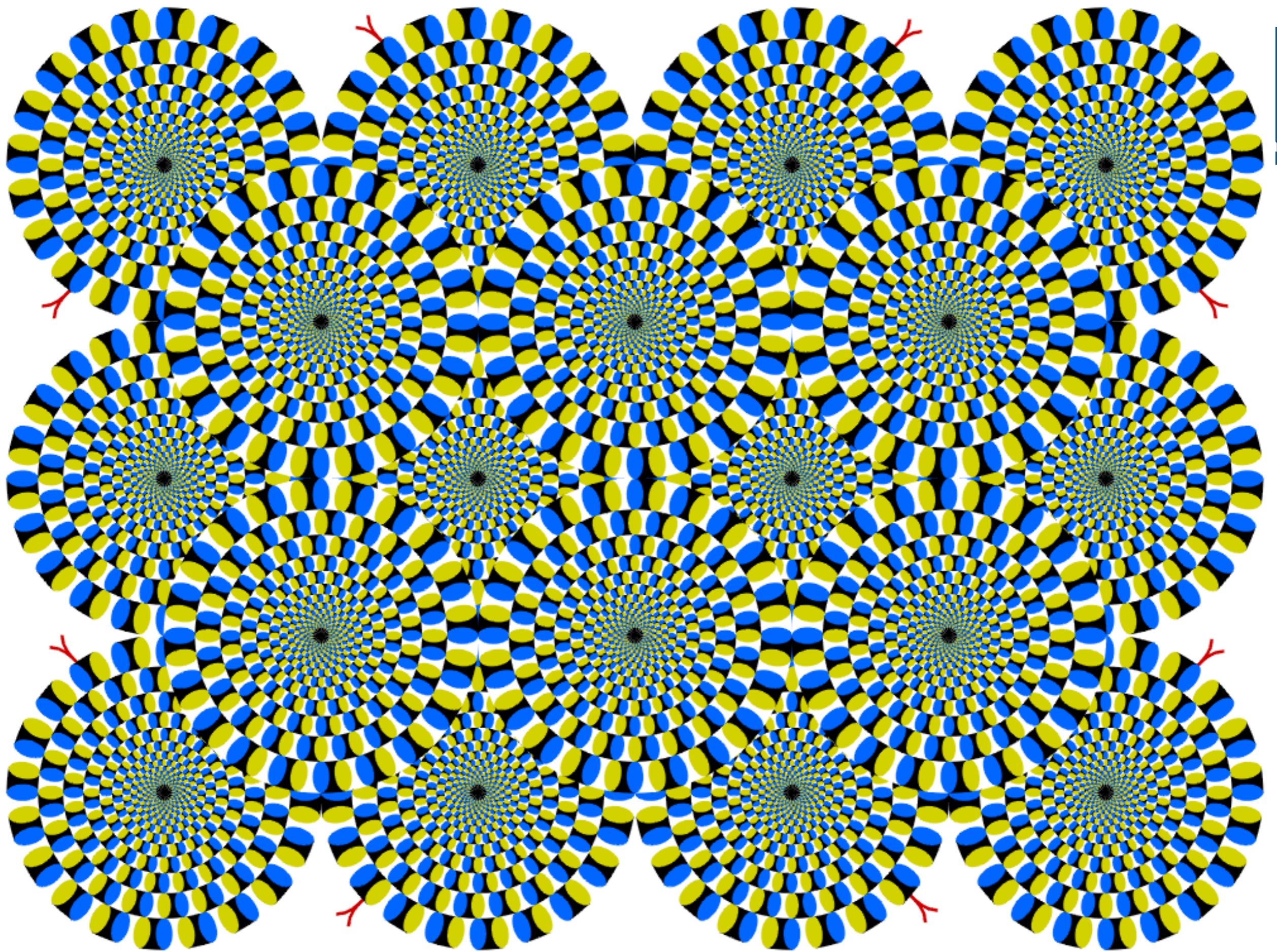
You

Does this image show two different people or is it the same person shown two times?



Image Caption Generator

The image shows two different people, one standing in the foreground, actively speaking into microphones, and another person in the background, seemingly listening or waiting for their turn to speak.

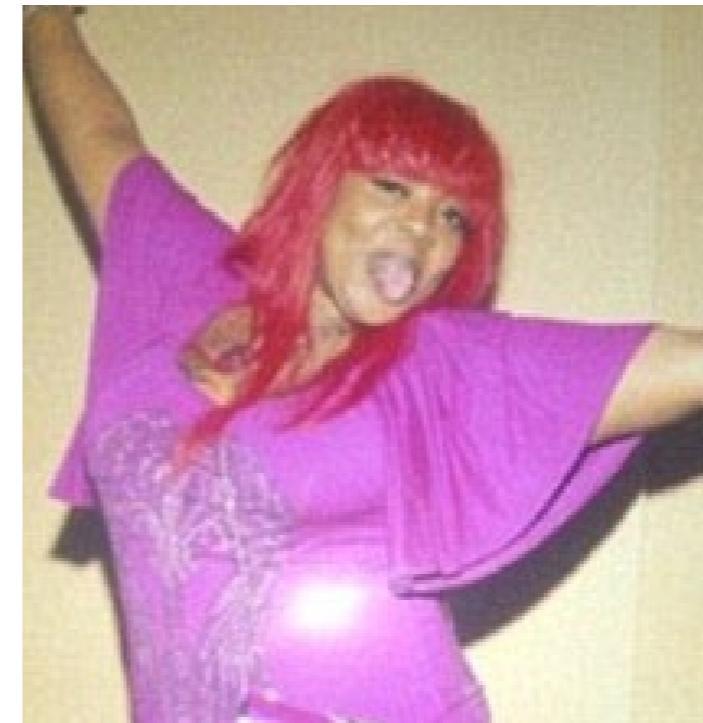


Face Recognition

- How many faces can you remember?

<https://www.dailymail.co.uk/sciencetech/article-3125173/Are-super-recogniser-test-one-elite-group-people-rarely-forgets-face.html>

- Try it out yourself!



same person?

Timeline before Alexnet 2012

The first digital image scanner was invented by transforming images into grids of numbers.

Basic image operations, such as edge detection, filtering, histograms etc.

Further advances in image processing, stereo vision, 3D reconstruction

“Energy based” Image segmentation, poisson blending

Broad range of approaches emerge for object recognition, tracking, and (BOW) classification

1950

1960

1970

1980

1990

2000

2010

Larry Roberts, the “father” of CV, described the process of deriving 3D info about solid objects from 2D photographs.

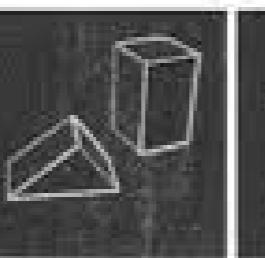
1960's: interpretation of synthetic worlds



Larry Roberts
"Father of Computer Vision"



Input image



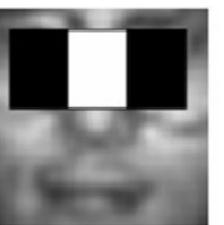
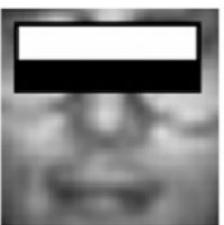
∇^3 gradient operator



computed 3D model
rendered from new viewpoint

ANNs proposed for Computer Vision, but mostly abandoned due to lack of computer power.
(AI Winter)

First Face Detection by Viola and Jones, by finding the nose and eyes

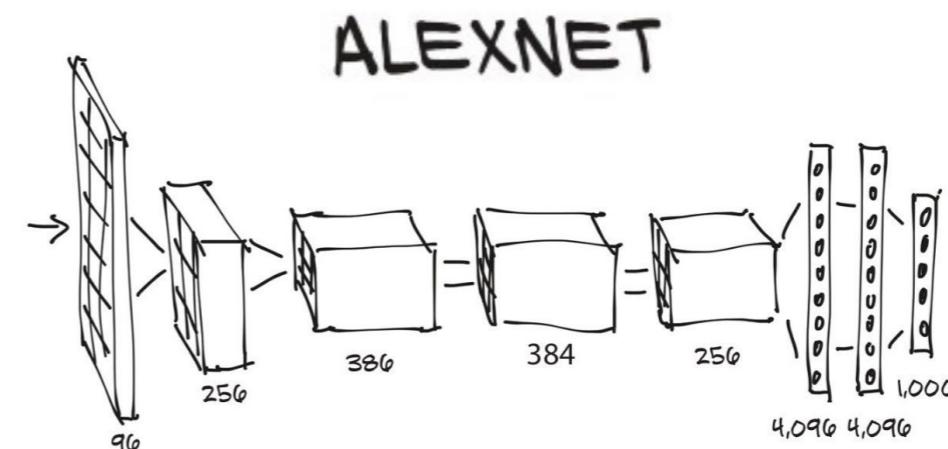
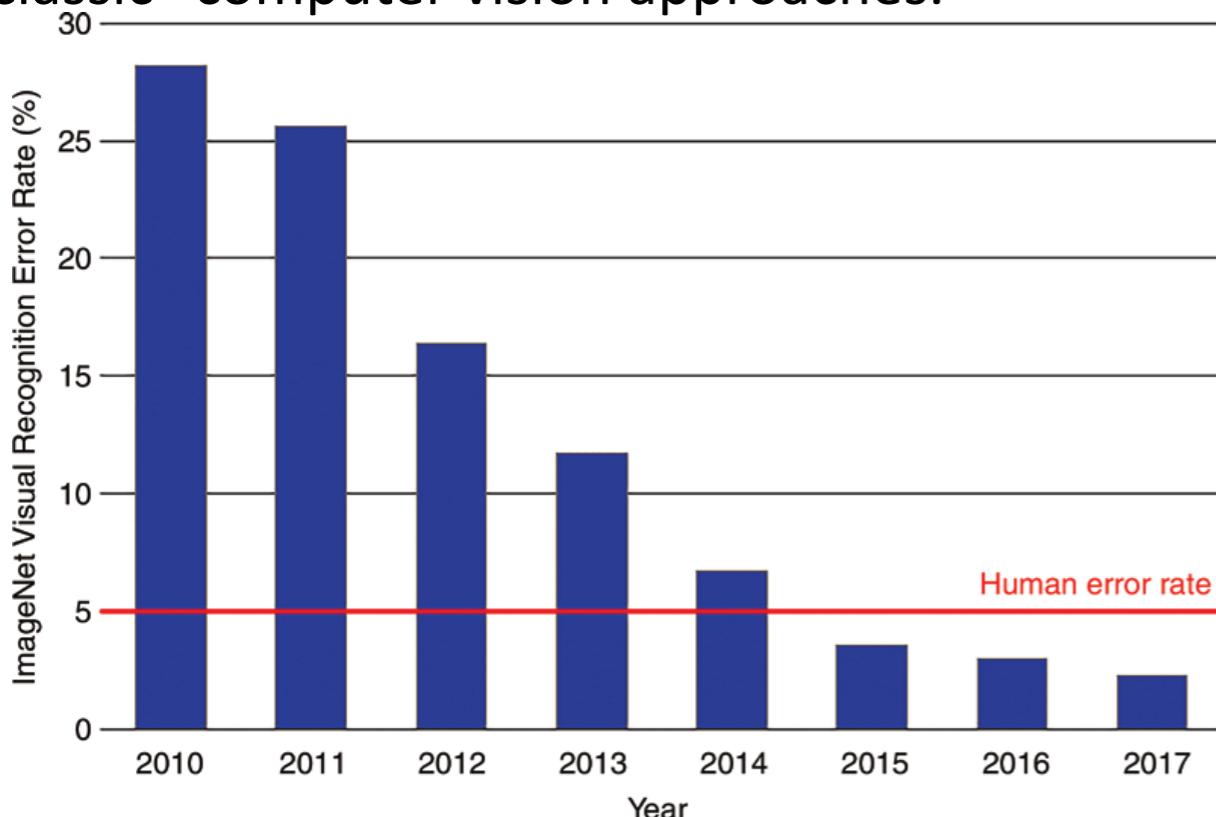


ImageNet challenge

The ImageNet project contains a very large amount of hand-labeled images (14 million+ images, 1000 classes)

The **ImageNet Challenge** is a yearly competition where computer vision experts try to build better **classification** models.

Before 2012, the winning models all used “classic” computer vision approaches.



Alex Krizhevsky

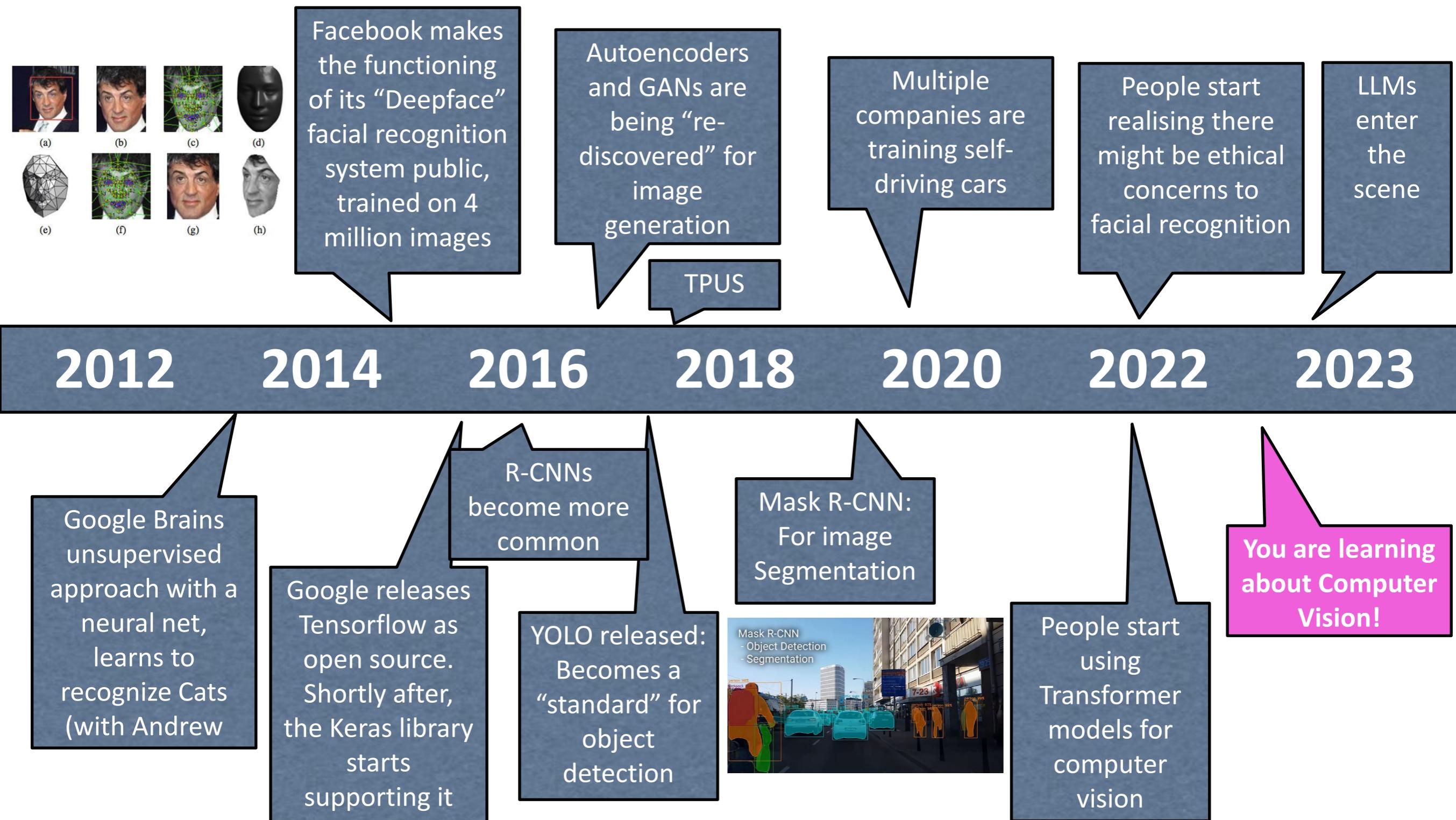
AlexNet was the winning approach of 2012, and is considered to be the trigger for the massive hype of Artificial Intelligence that followed

It is considered a „turning point“ in Computer Vision, by many

It was a (convolutional) neural net based approach that proved significantly better in recognizing images than any approach tried before

All future winners of the ImageNet challenge were also Neural Net based

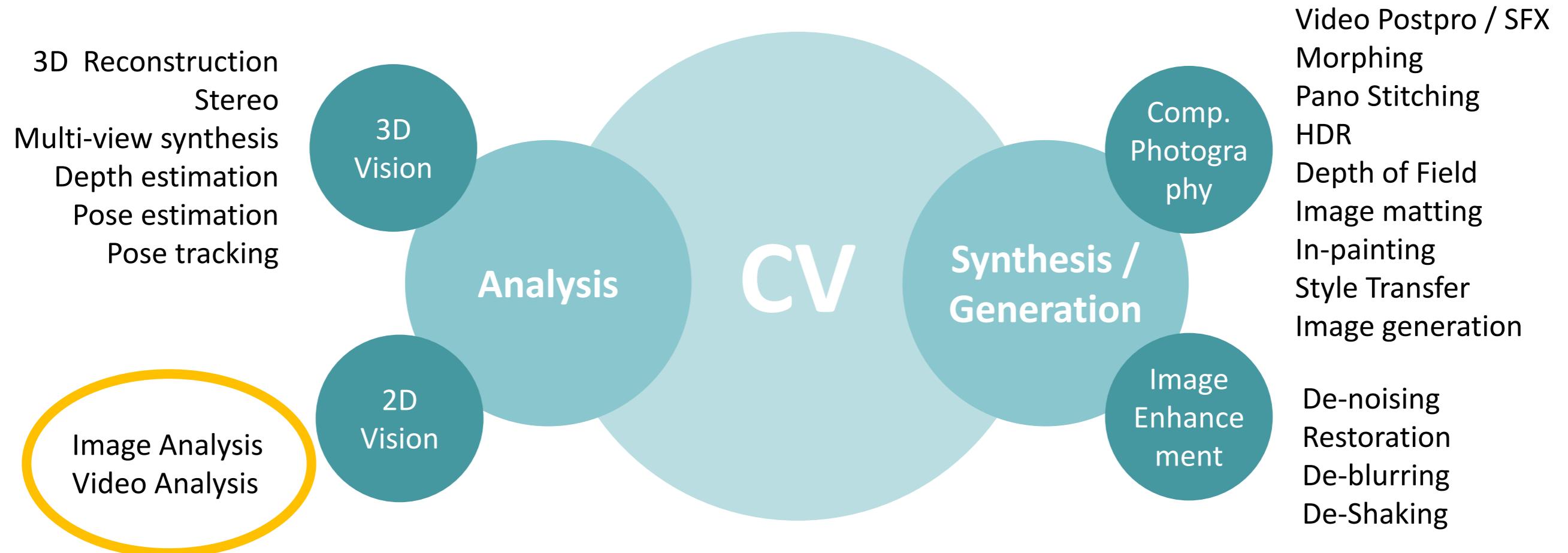
Timeline after Alexnet 2012





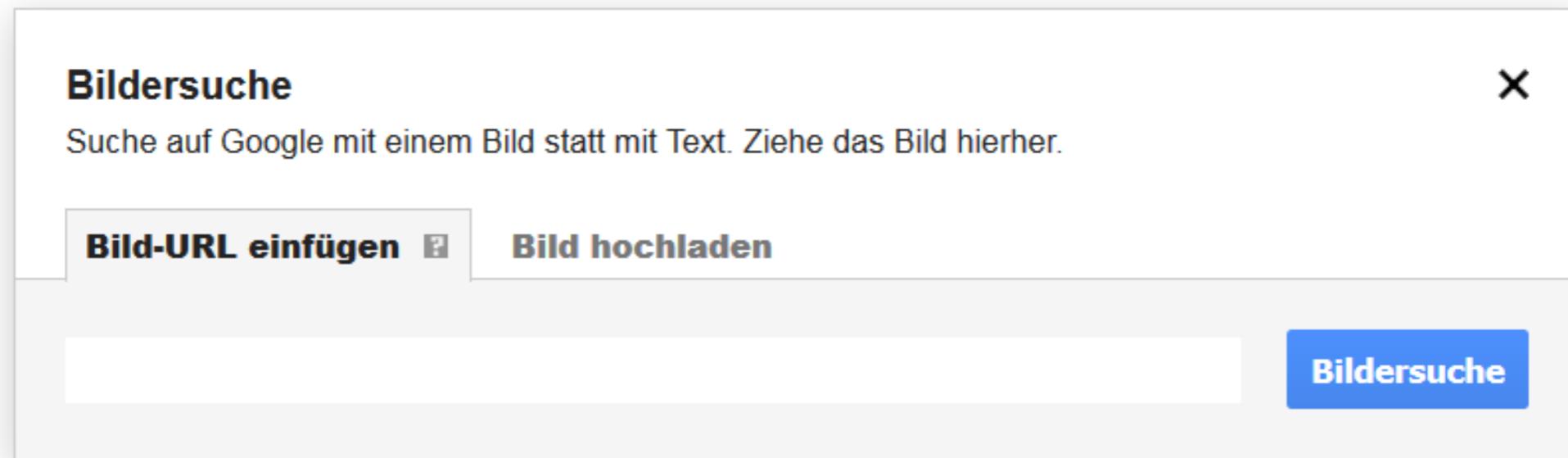
Applications of Computer Vision

Computer Vision Applications



Google Reverse Image Search

- <http://images.google.com>



Sketch-based image retrieval

- Query by sketch: <http://sketchy.eye.gatech.edu>

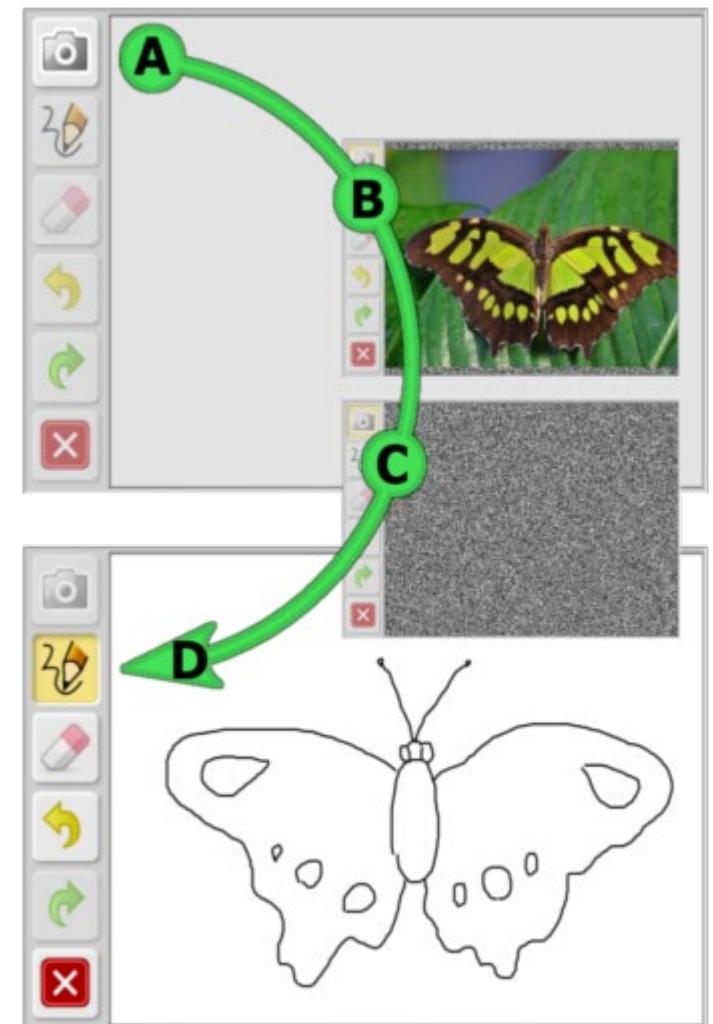
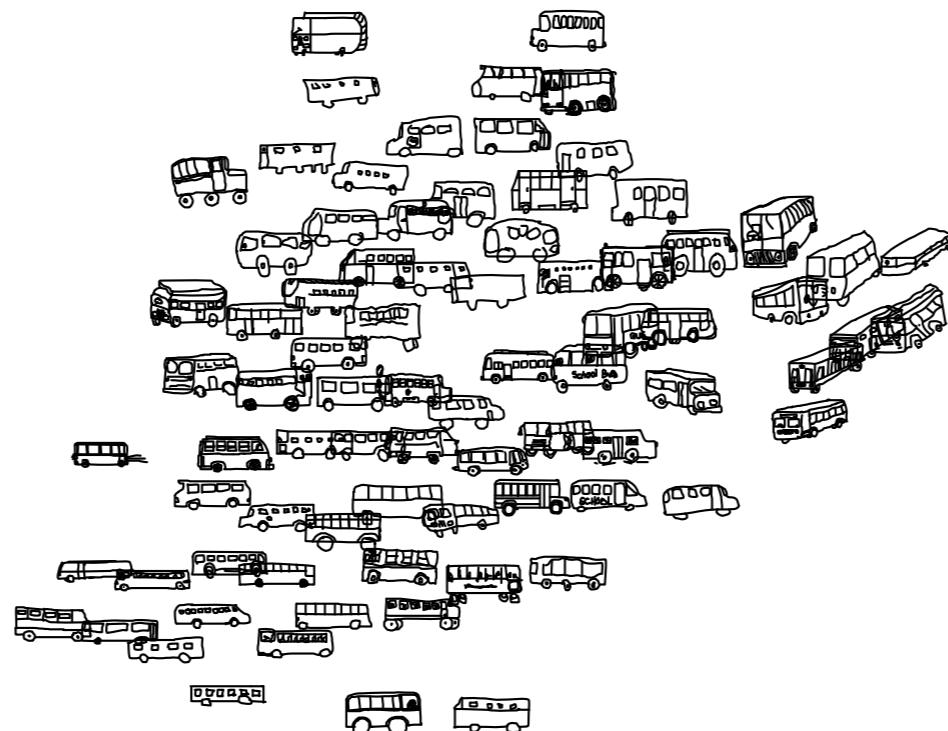
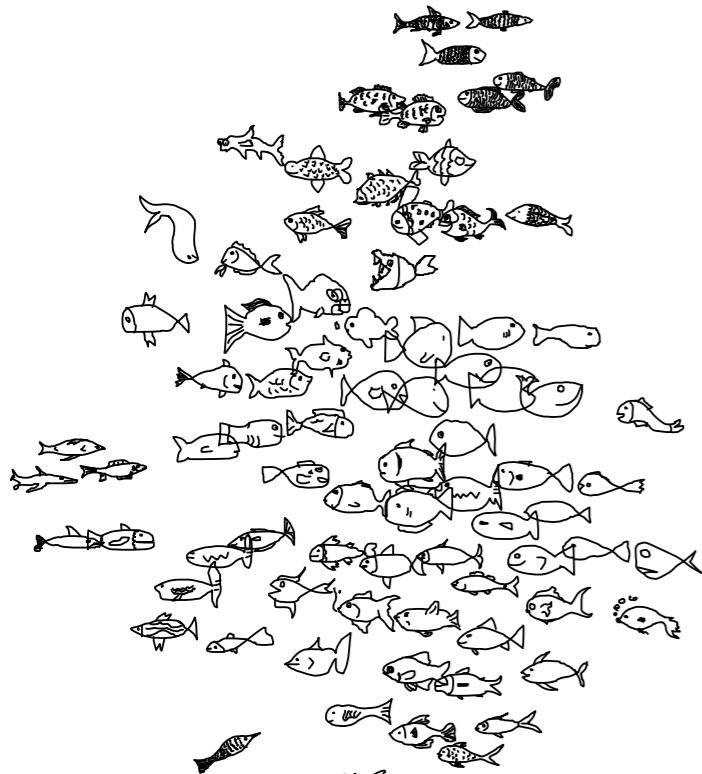


Image Tagging

The screenshot shows a web browser window with the Clarifai demo page open at <https://www.clarifai.com/demo>. The page displays a photograph of a smiling couple and lists several predicted concepts with their corresponding probabilities.

Predicted Concepts:

Predicted Concept	Probability
togetherness	0.967
love	0.967
outdoors	0.966
woman	0.961
nature	0.950
people	0.938
affection	0.938

Language: English (en)

General [VIEW DOCS](#)

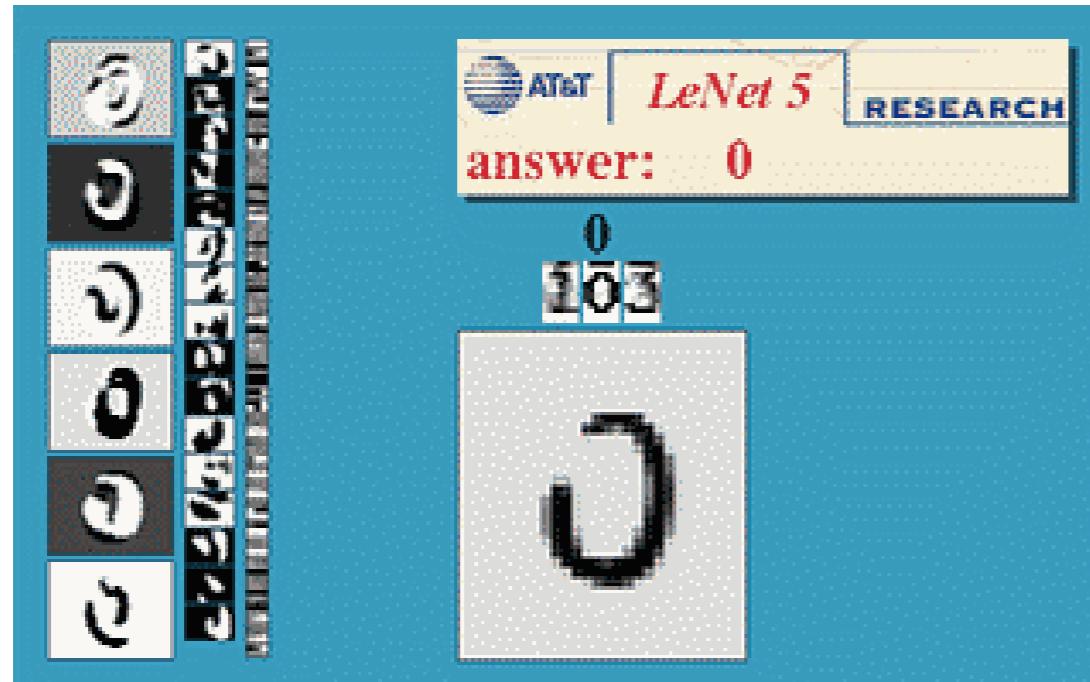
TRY YOUR OWN IMAGE OR VIDEO

<https://clarifai.com/explore>

Optical character recognition (OCR)

Technology to convert scanned docs to text

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

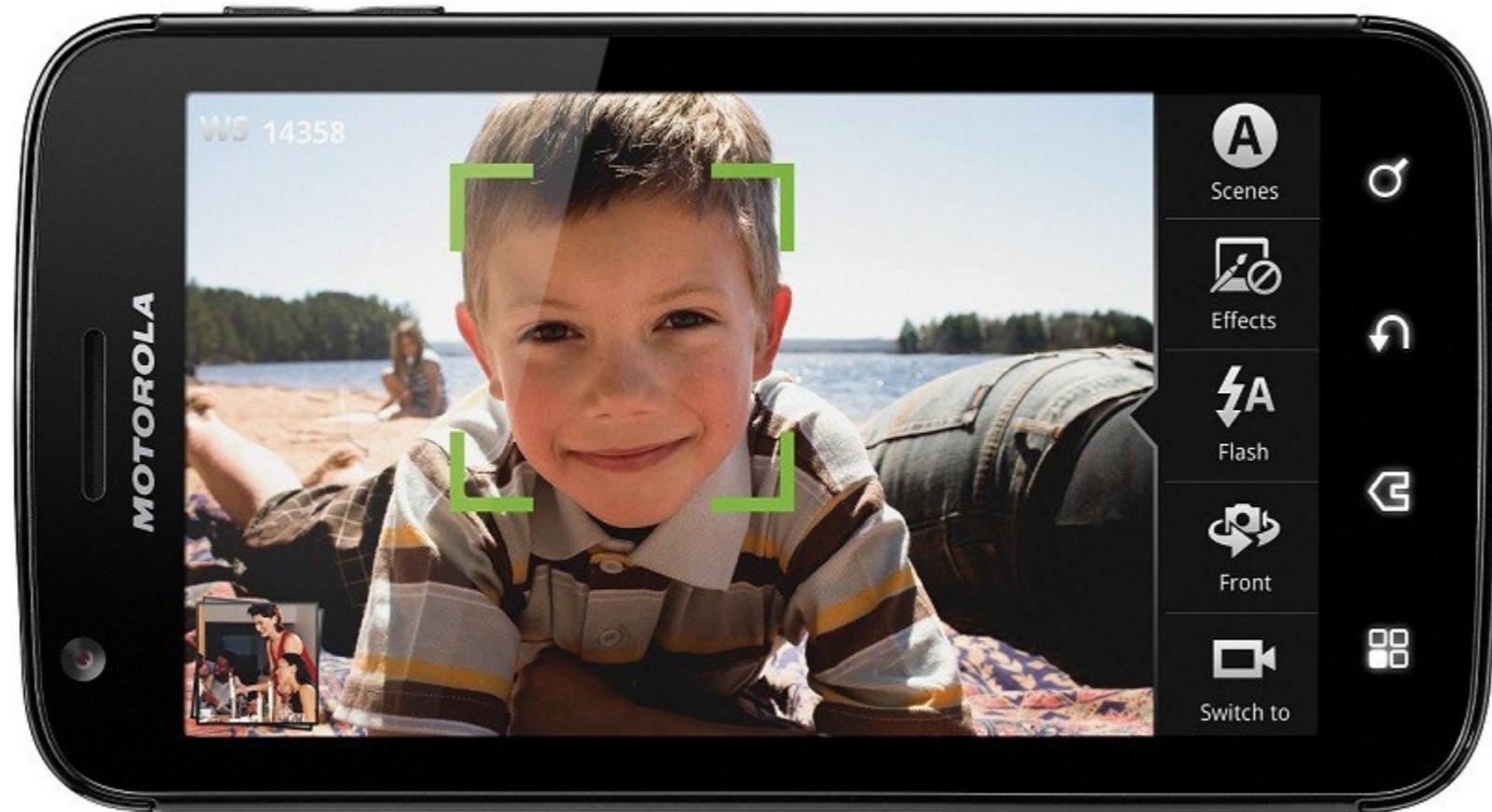


Digit recognition, AT&T labs
<http://yann.lecun.com/exdb/lenet/>



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

Face detection



Almost all cameras today detect faces!



[Face priority AE] When a bright part of the face is too bright

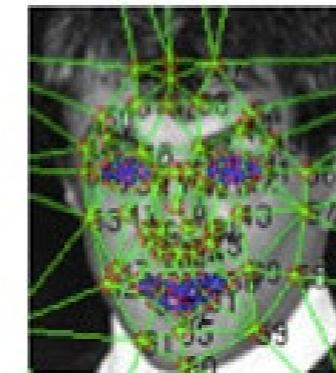
Face Recognition - Facebook DeepFace



(a)



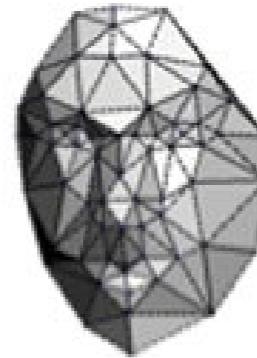
(b)



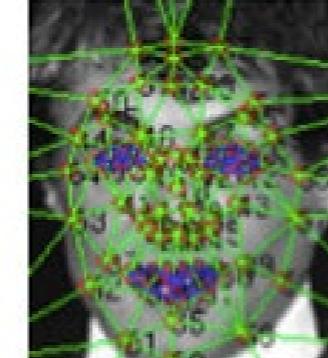
(c)



(d)



(e)



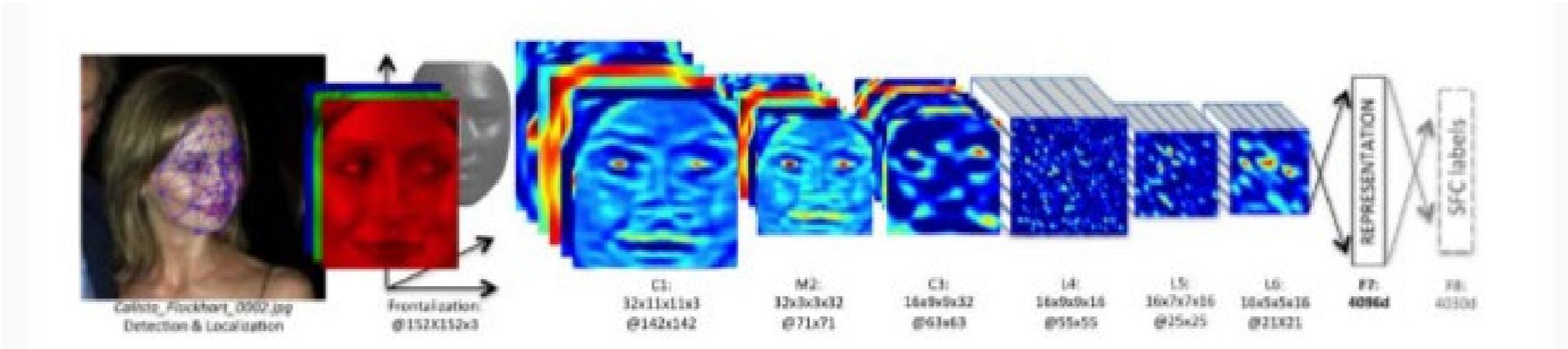
(f)



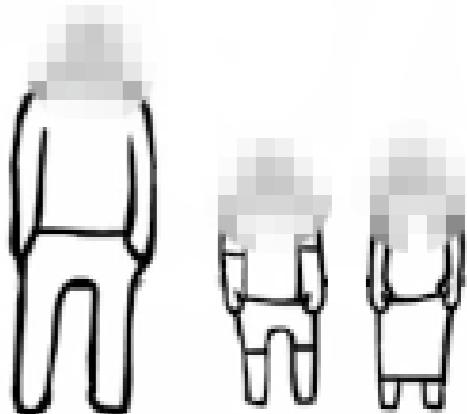
(g)



(h)



Google Street View – Blurring Faces



[Web](#) [Images](#) **Maps** [News](#) [Shopping](#) [Mail](#) [more ▾](#)

philipp.lenssen@gmail.com | [My Profile](#) | [Help](#) | [Web History](#) | [My Account](#) | [Sign out](#)

Google Maps e.g., "10 market st, san francisco" or "hotels near lax"
1504 broadway, nyc

[Search the map](#) [Find businesses](#) [Get directions](#) [Search Maps](#) [Show search options](#)

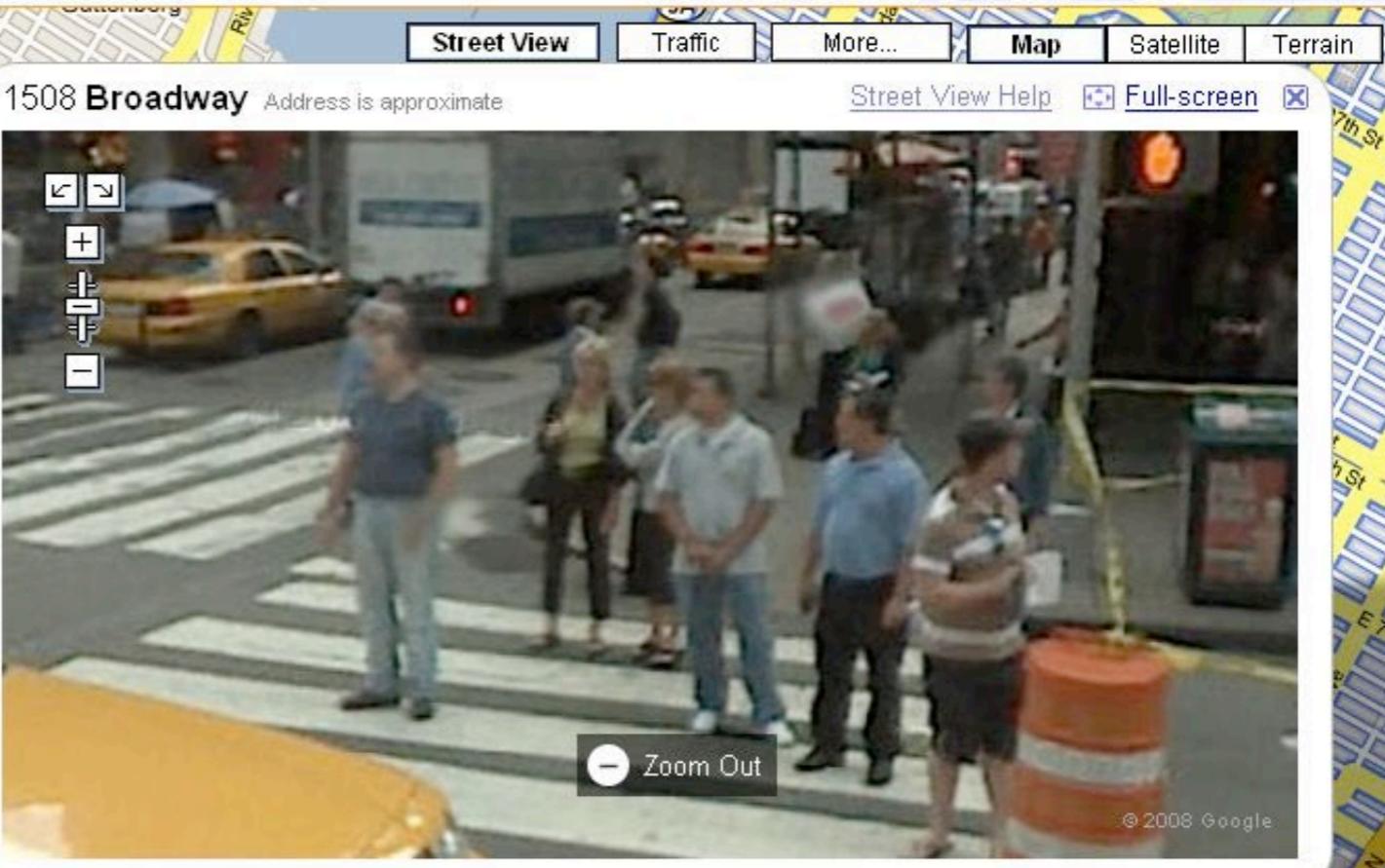
[Search Results](#) [My Maps](#) [Print](#) [Send](#) [Link to this page](#)

1504 Broadway
New York, NY 10036
[Make this my default location](#)

New!
[Explore this area »](#)

Photos

User-Created Maps
[Cheap Eats in NYC](#)
[North American New Humanist Forum](#)
[Barrel to Bottles](#)
[More photos, videos, and user-created maps »](#)

Street View [Traffic](#) [More...](#) [Map](#) [Satellite](#) [Terrain](#)
[Street View Help](#) [Full-screen](#)

© 2008 Google

Google street view



Object Detection & Segmentation

Semantic Segmentation



GRASS, CAT,
TREE, SKY

No objects, just pixels

Classification + Localization



CAT

Single Object

Object Detection



DOG, DOG, CAT

Multiple Object

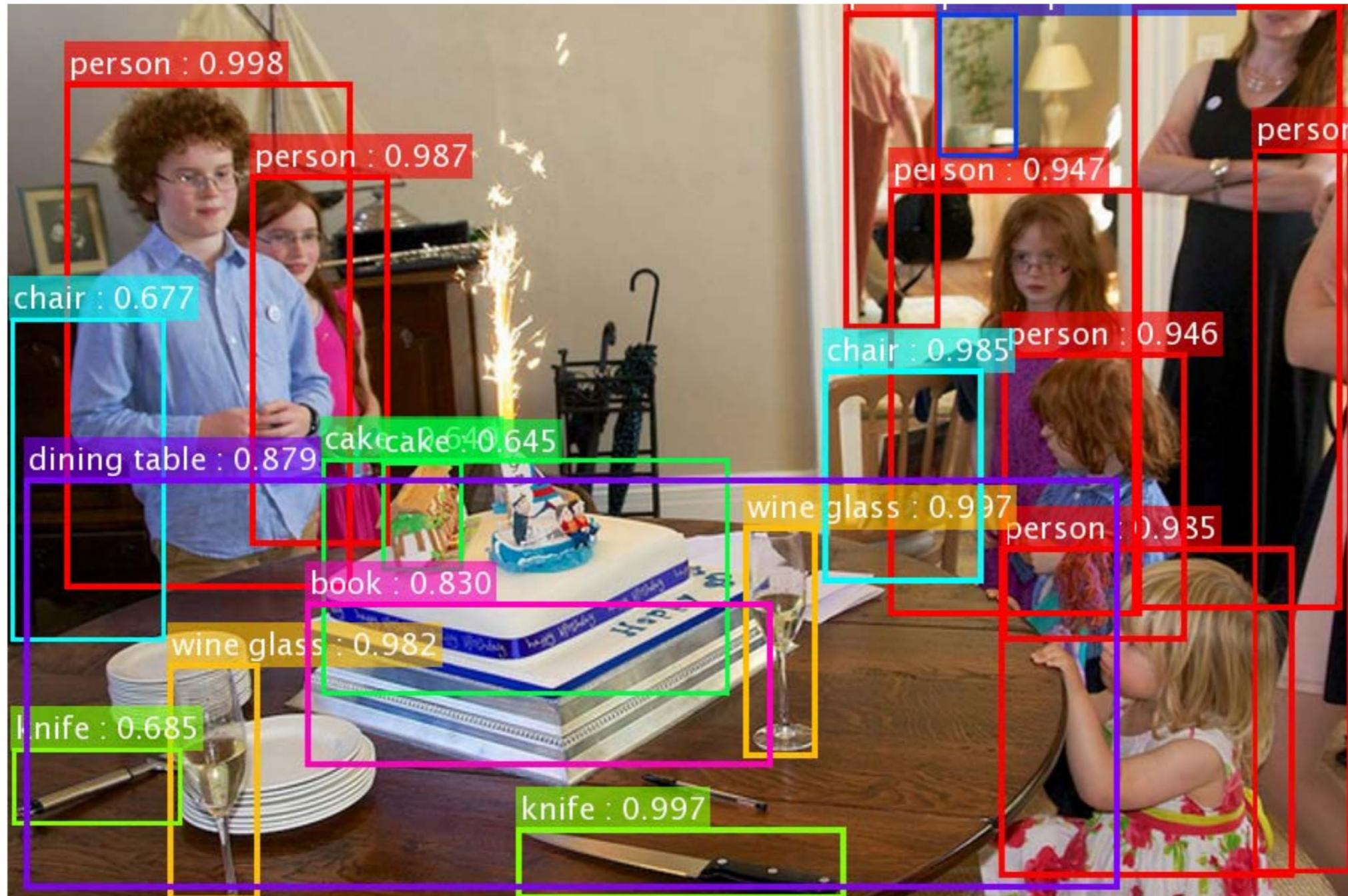
Instance Segmentation



DOG, DOG, CAT

This image is CC0 public domain

Object Detection & Recognition



Object recognition (in mobile phones)

<https://lens.google/>

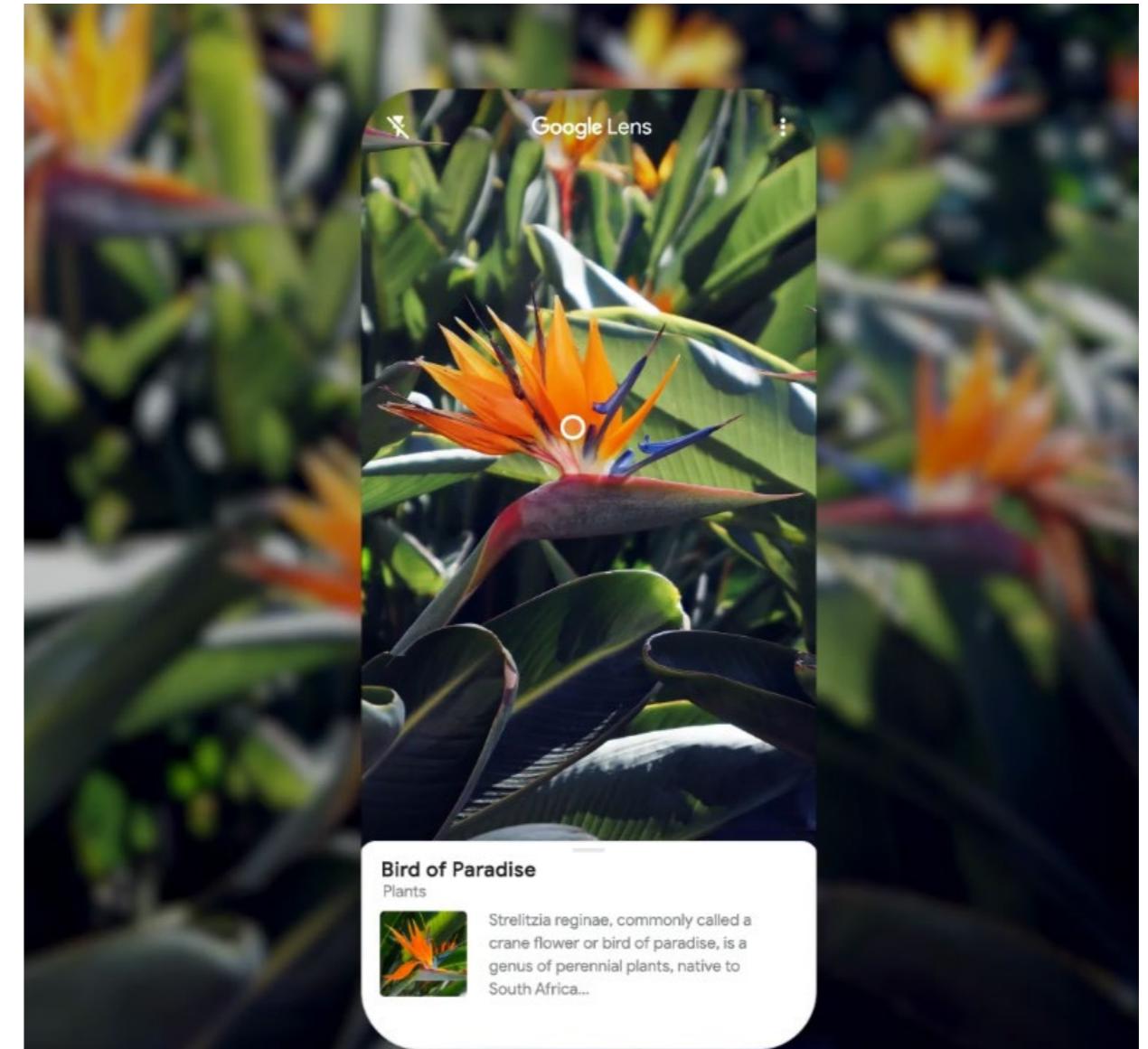


Image Segmentation

- E.g. Powerpoint function “Freistellen” 
- Basis: 10 years of research in segmentation and energy minimization
- Lets try:



Segment Anything (Meta)

- Demo: <https://segment-anything.com/demo>



Automatic Image Captioning

- Automated image caption generation
- Vinyals et al. [2014] <https://arxiv.org/pdf/1411.4555.pdf>

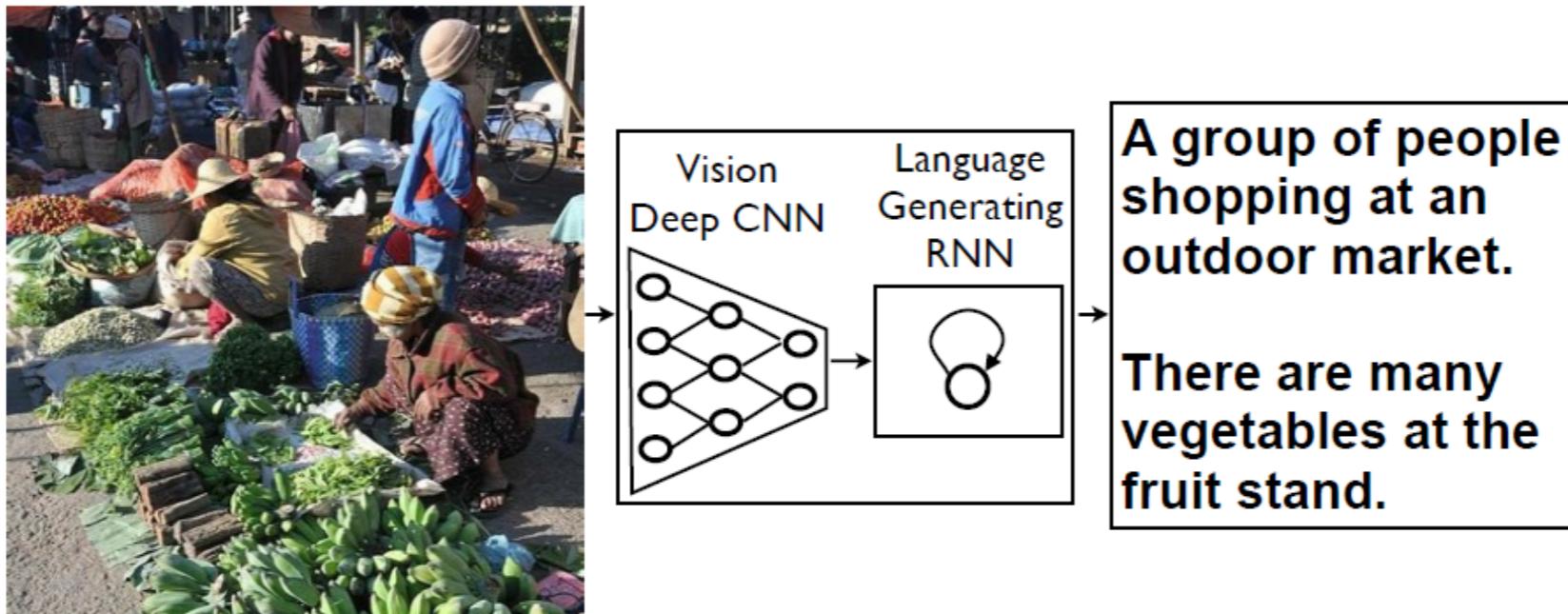
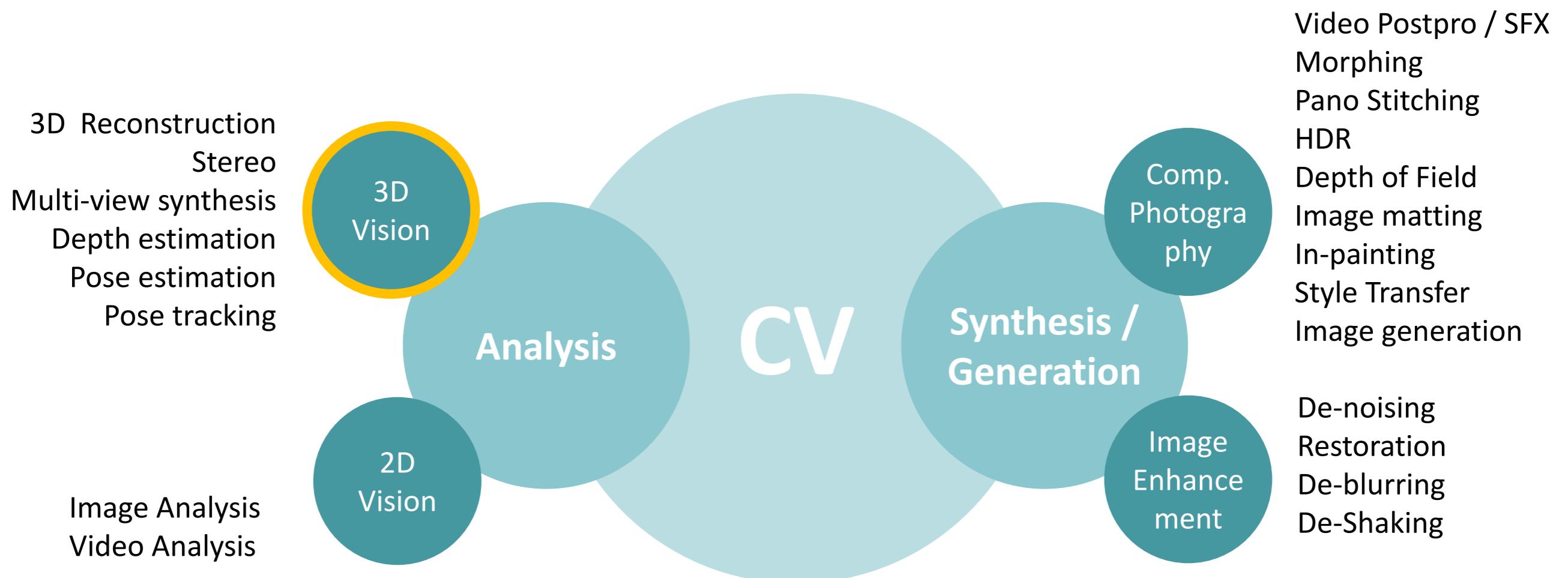
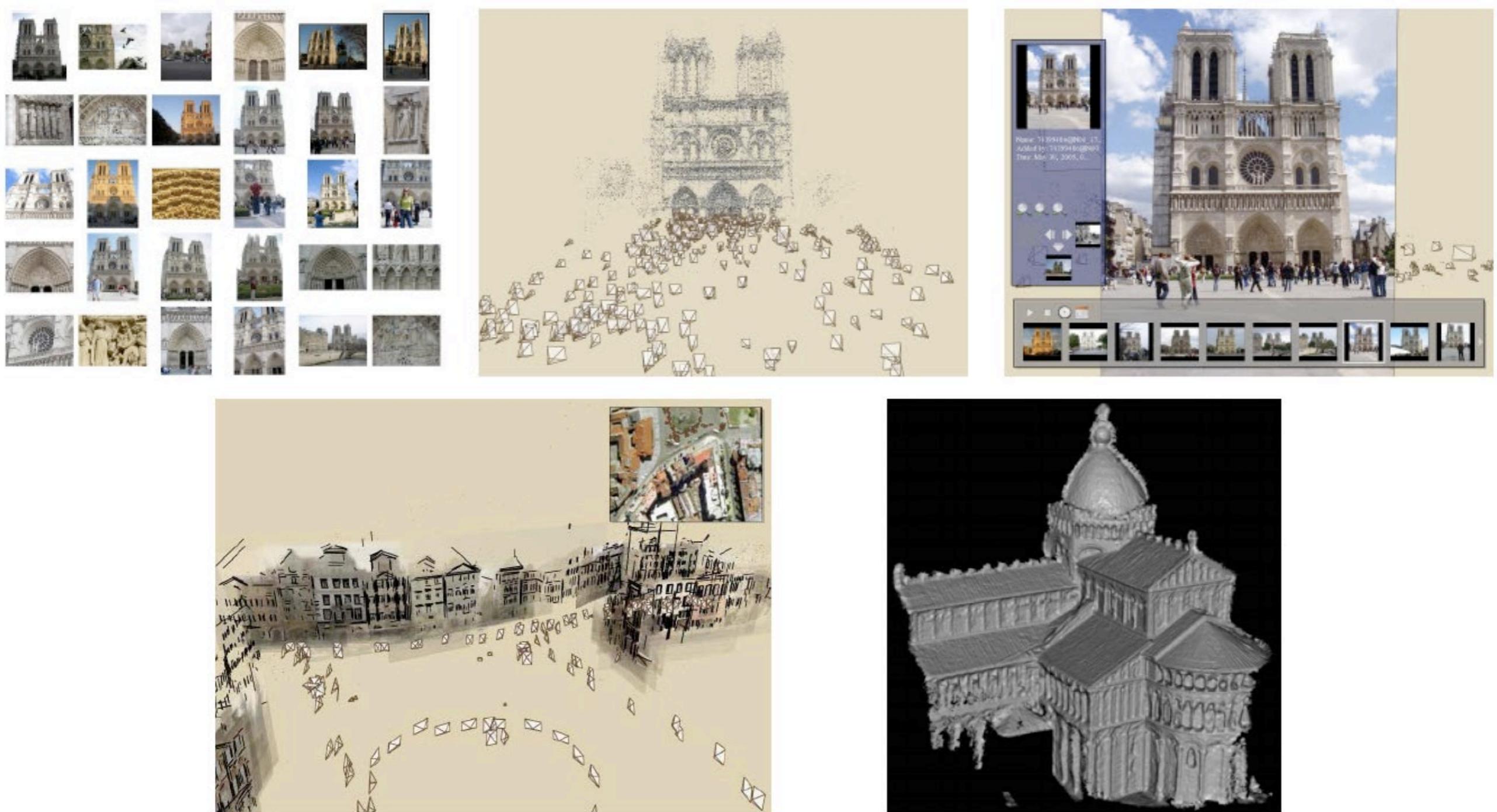


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

Computer Vision Applications



3D Reconstruction



Snavely et al. 2006

Goesele et al. 2007

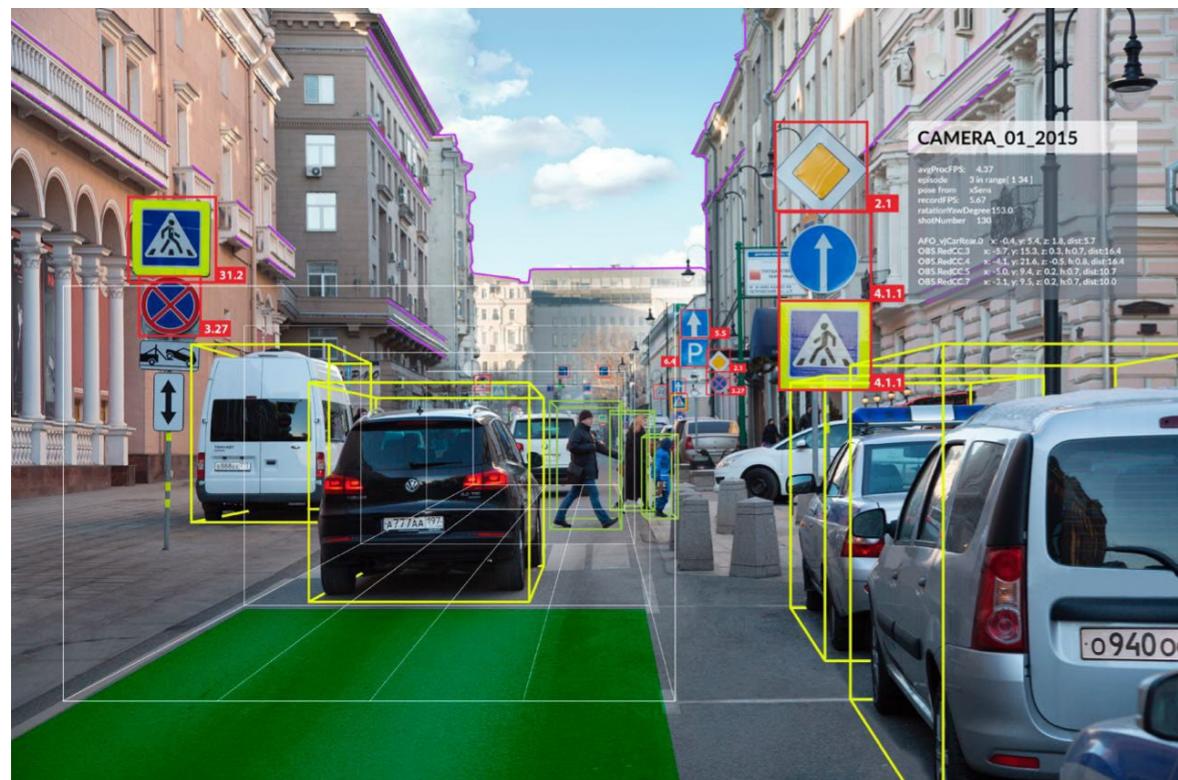
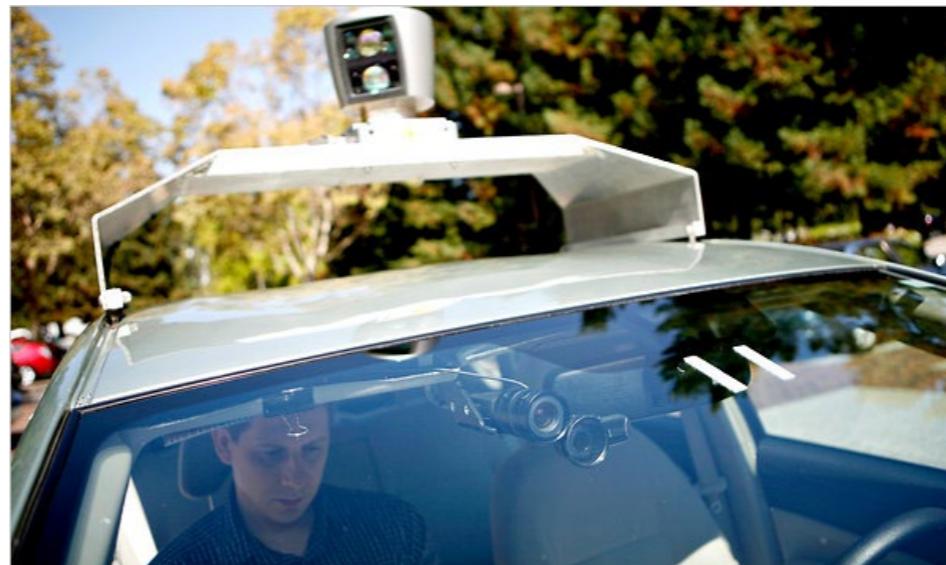
3D from thousands of images



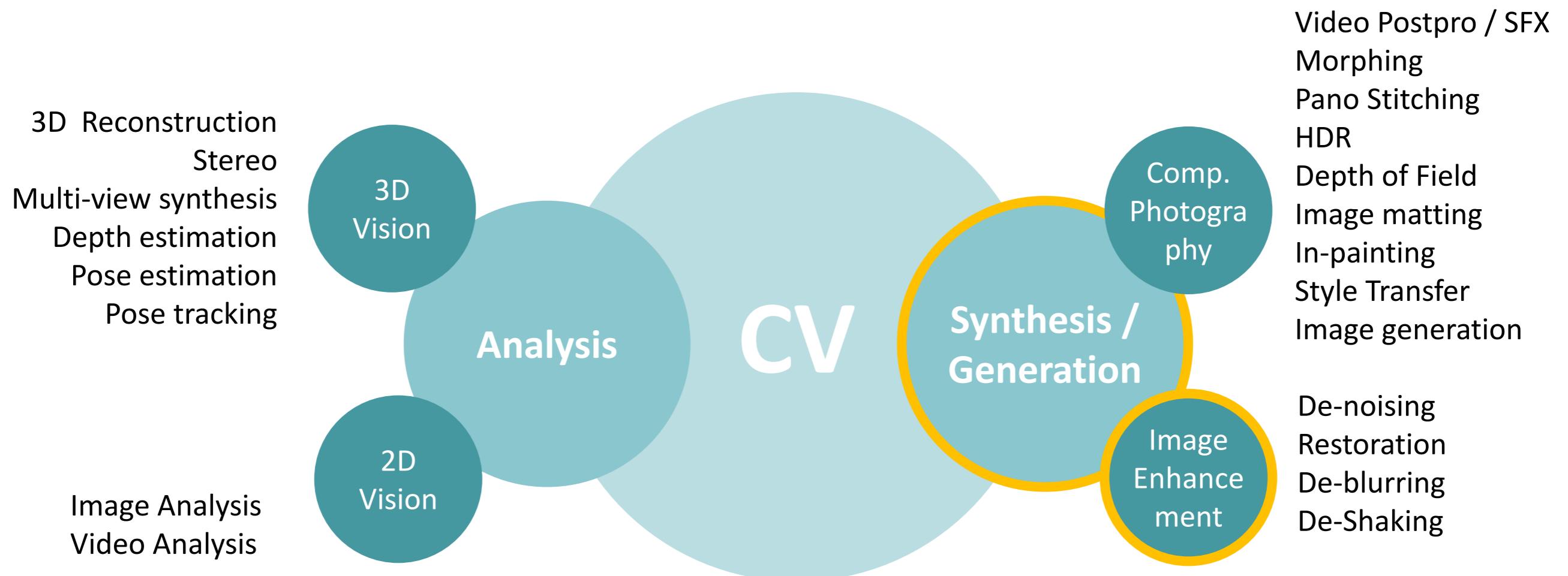
<https://grail.cs.washington.edu/rome/>

Building Rome in a Day: Agarwal et al.

Self-driving cars

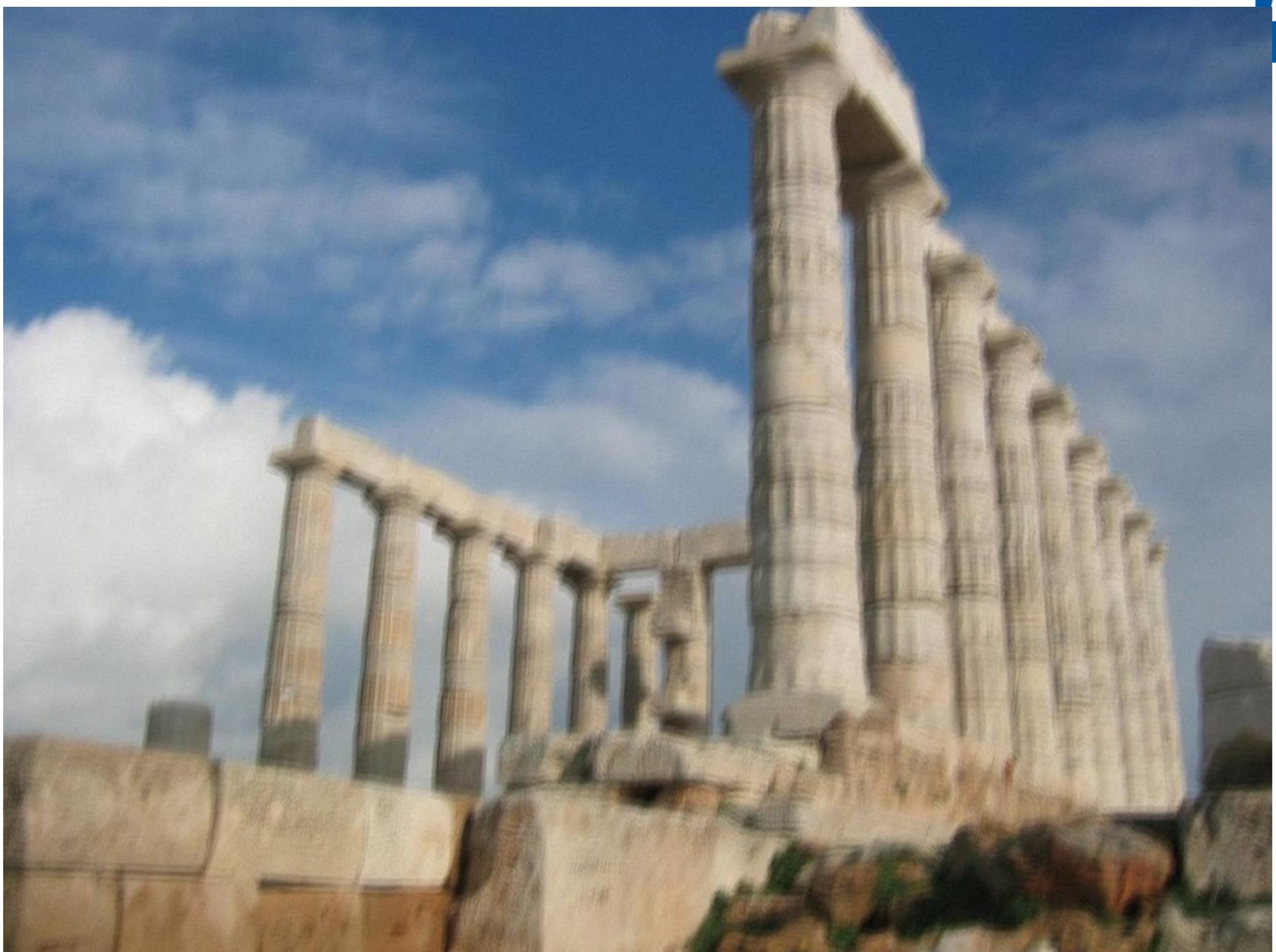


Computer Vision Applications



Repairing Camera Shake





Reconstruction



Original



Computer Vision Applications

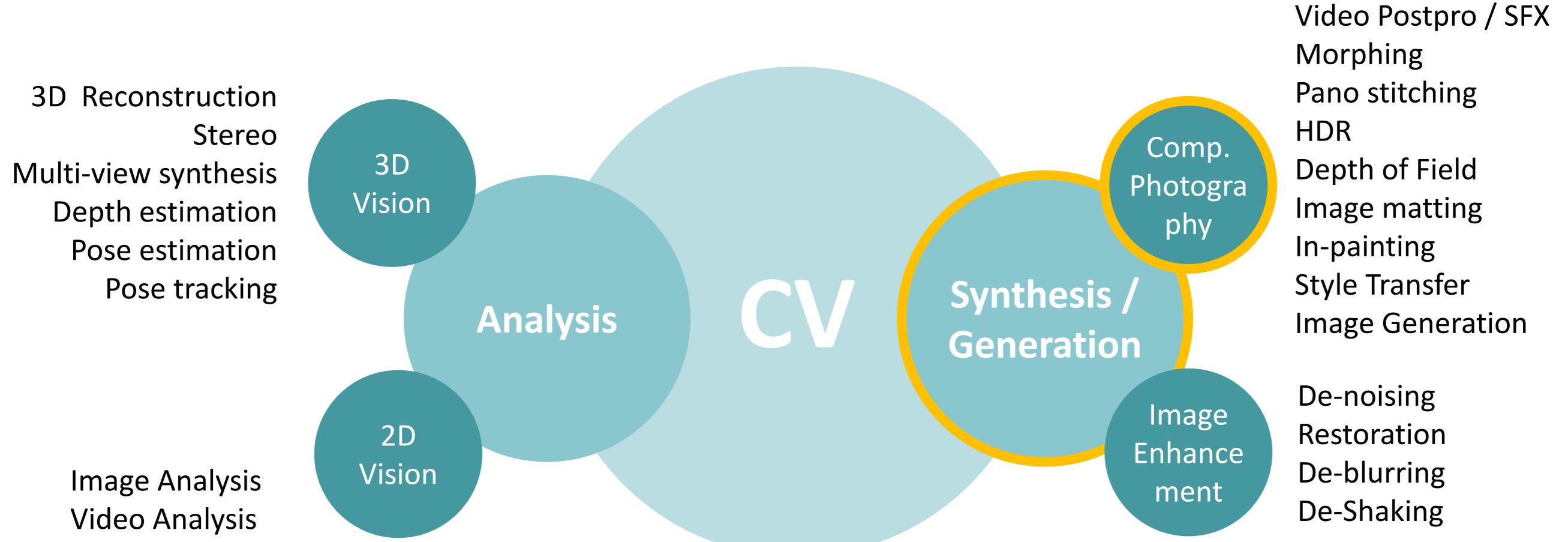


Image Matting

User draws a rectangle or lasso around an object

Object edges are detected and feathered

Approach: binary graph cut w/ color statistics



User Input



Segmentation

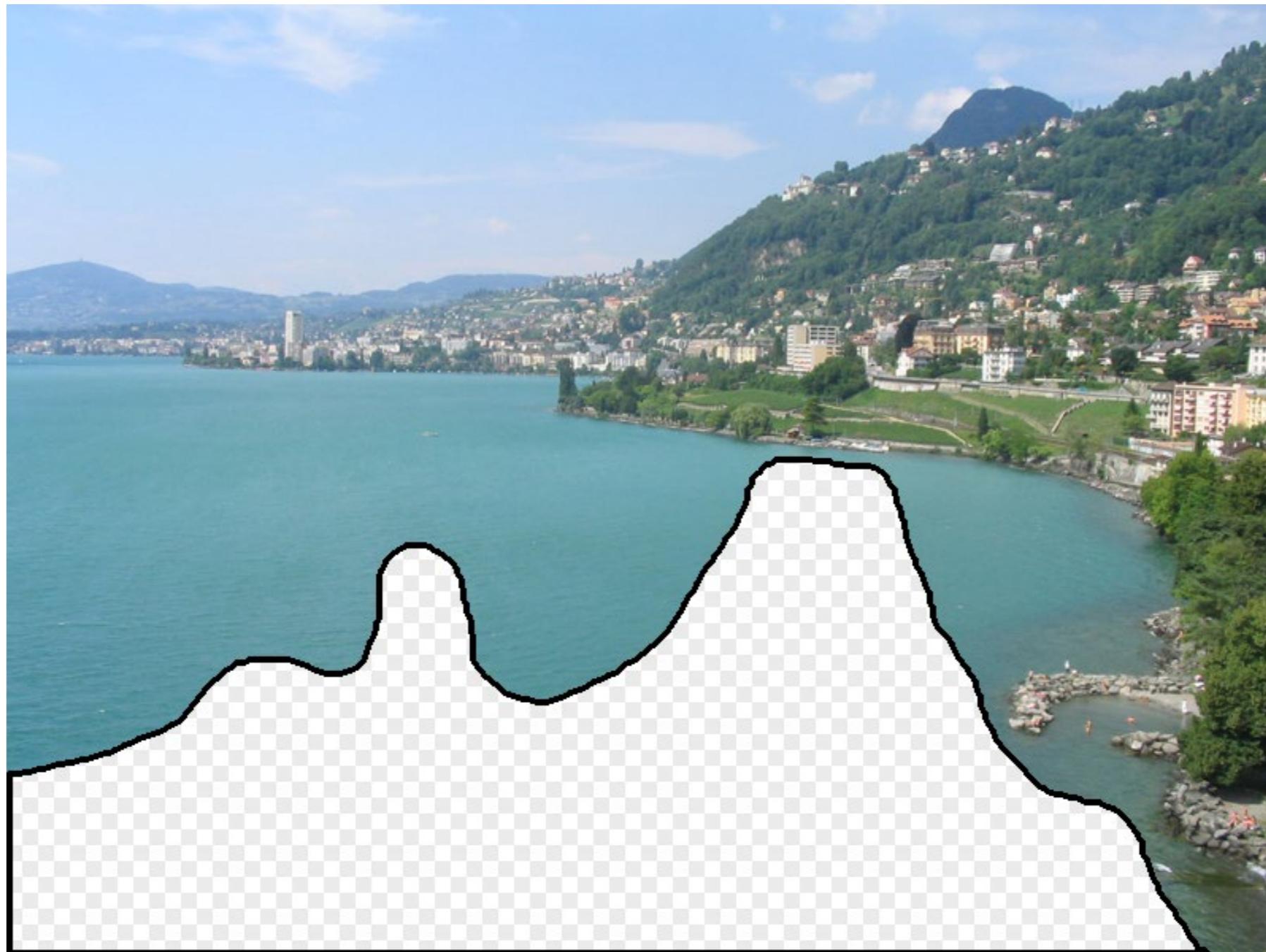


New composed Image

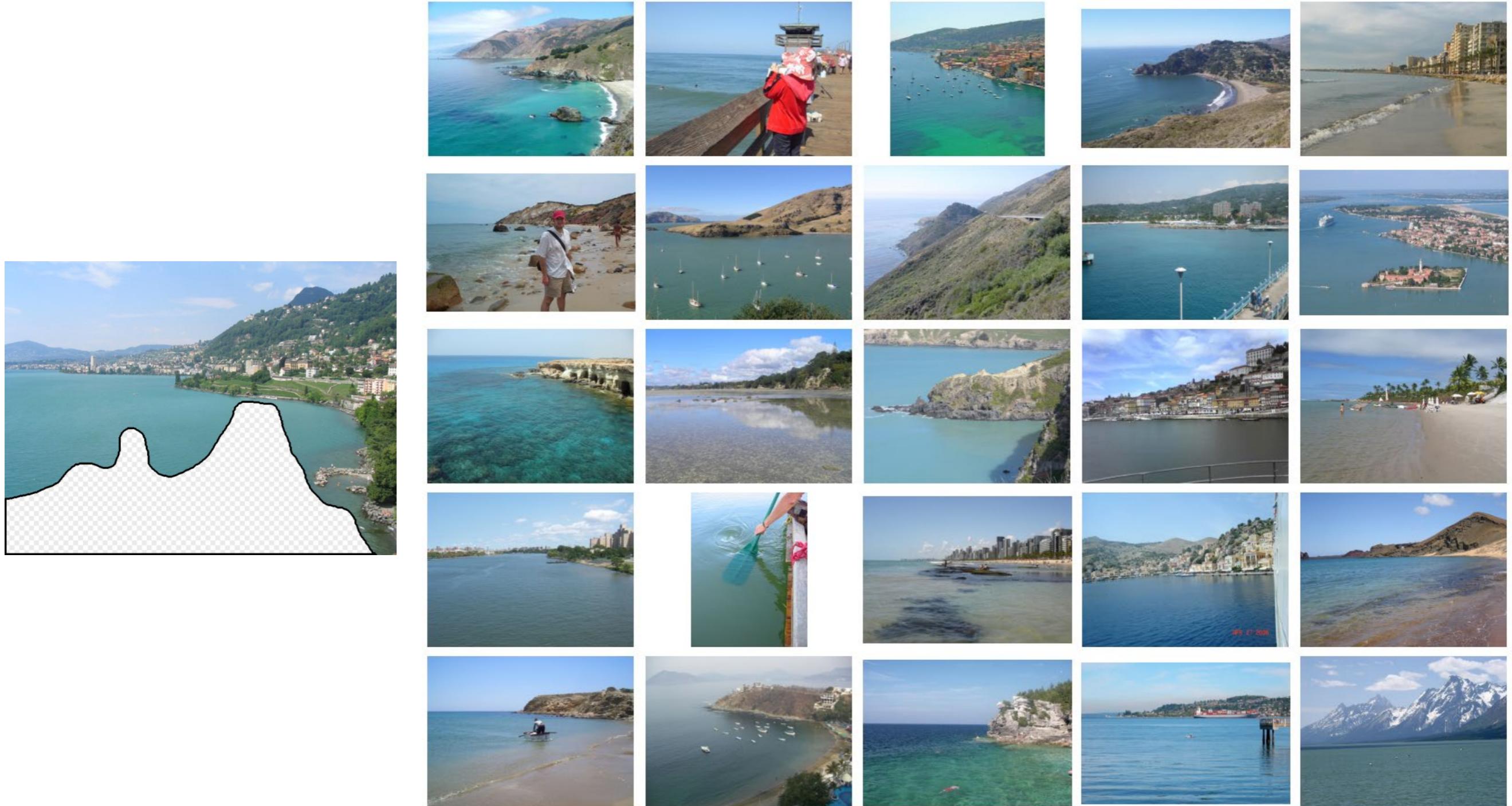
Removing people



Scene Completion



[Hays and Efros. Scene Completion Using Millions of Photographs.
SIGGRAPH 2007 and CACM October 2008.]



Nearest neighbor scenes from database of
2.3 million photos



Graph cut + Poisson blending

Image stitching / Panorama stitching



(a) Image 1



(b) Image 2



(c) SIFT matches 1



(d) SIFT matches 2



(e) RANSAC inliers 1



(f) RANSAC inliers 2



(g) Images aligned according to a homography

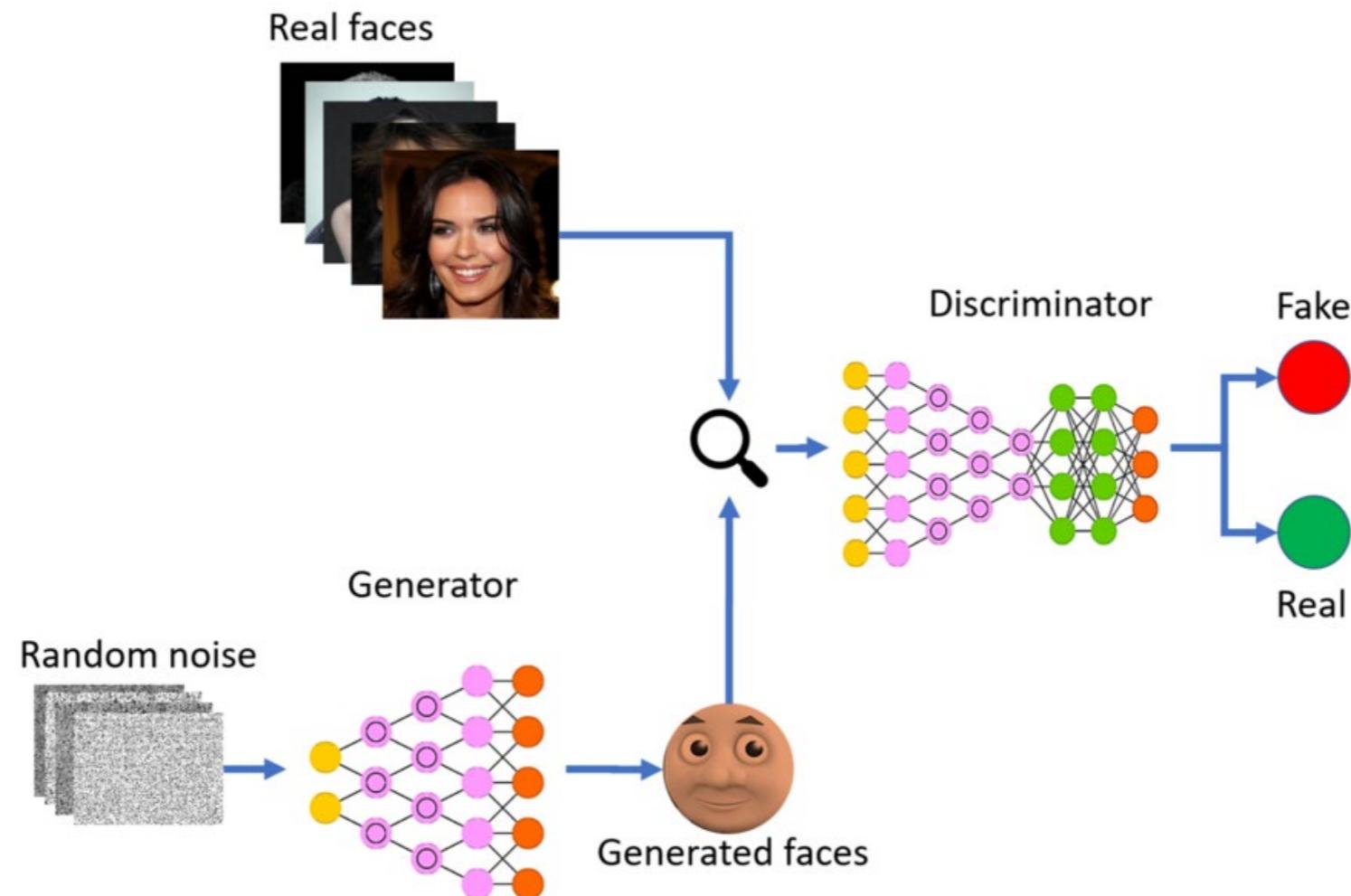
Style Transfer

- Mapping an image to a new style that has been learned before
- <https://algorithmia.com/algorithms/deeplearning/DeepFilter>



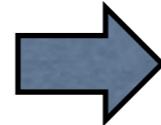
Image Generation by GANs

- GAN = Generative Adversarial Network
- Introduction: <https://www.analyticsvidhya.com/blog/2017/06/introductory-generative-adversarial-networks-gans/>



GAN-based Inpainting

- Example: GanPaint (<https://ganpaint-demo.vizhub.ai/>)



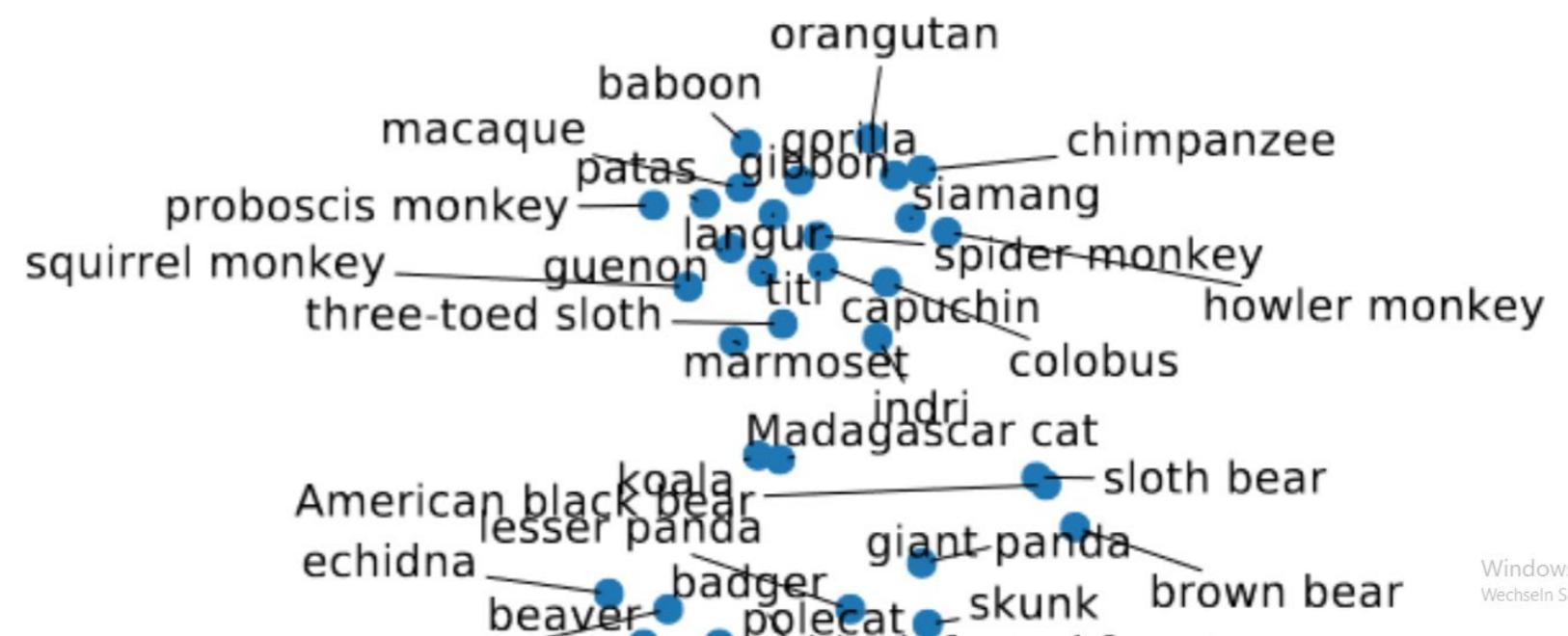
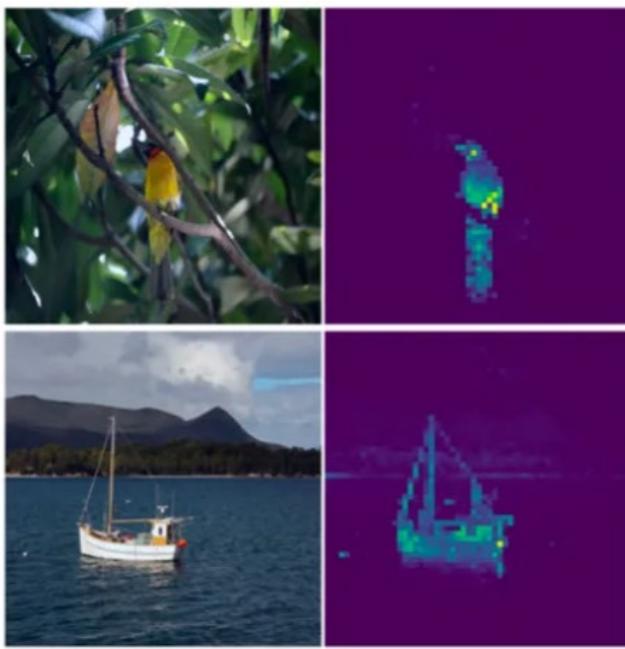
Generate Realistic Faces

- Demo: This person does not exist:
<https://this-person-does-not-exist.com/en>



Foundation Models in Computer Vision

- Foundation model: a (machine learning model) that has been trained to gain a general understanding of an image
- Not trained for a specific task, rather to understand images in general
- Example: DINO (Meta): <https://towardsdatascience.com/dino-emerging-properties-in-self-supervised-vision-transformers-summary-ab91df82cc3c>
- Nicely explained here:
https://www.youtube.com/watch?v=csEgtSh7jV4&ab_channel=AI%20Papers%20Academy



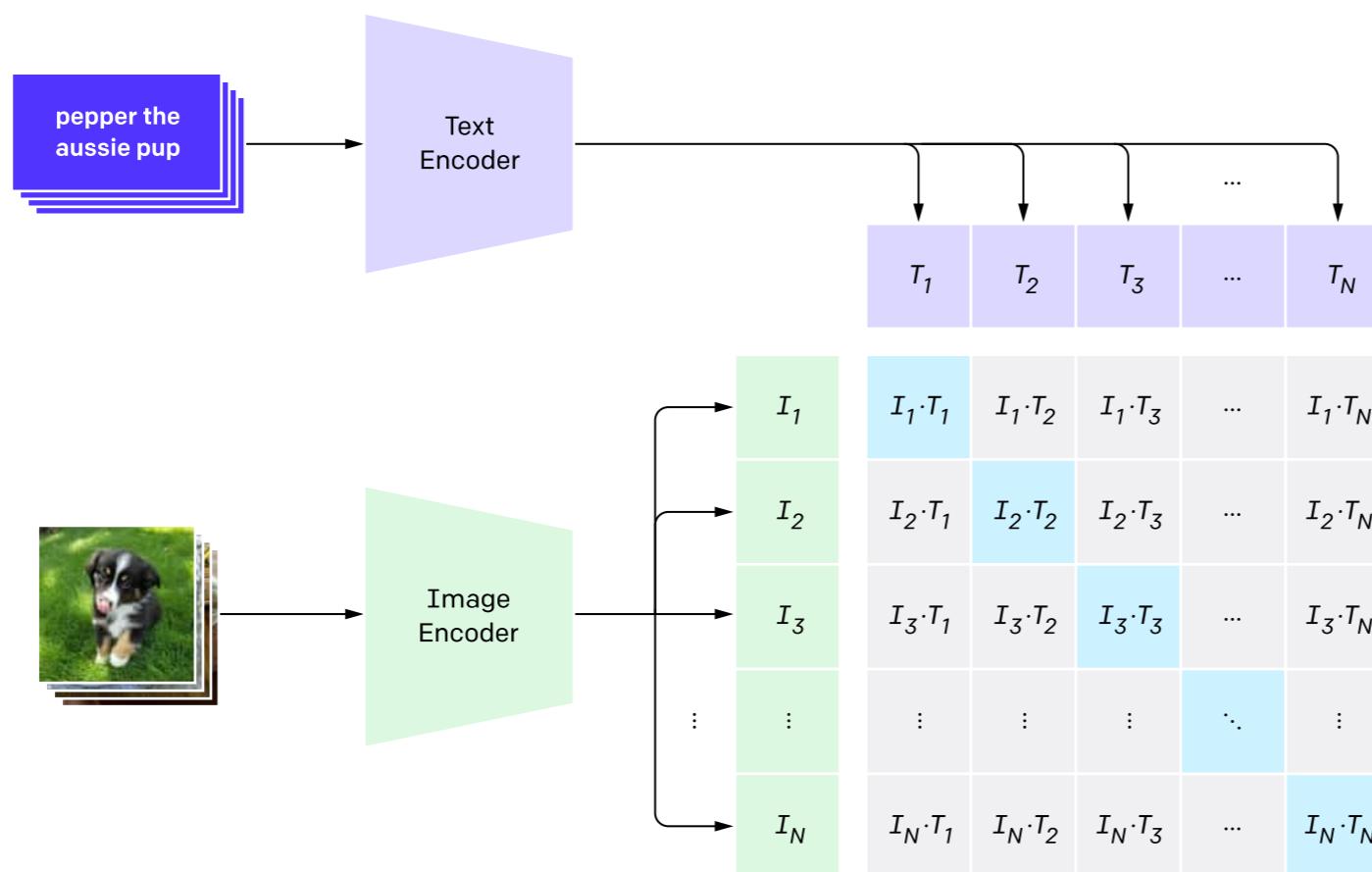


**Large Language Models open a new Chapter in
Computer Vision..**

Large Language Models Enter the Scene..

- CLIP: Multimodal (text – image) representations:
<https://openai.com/research/clip>

1. Contrastive pre-training



Generate Images from Text

- Demo DALL-E 2 (OpenAI) <https://openai.com/dall-e-2/#demos>

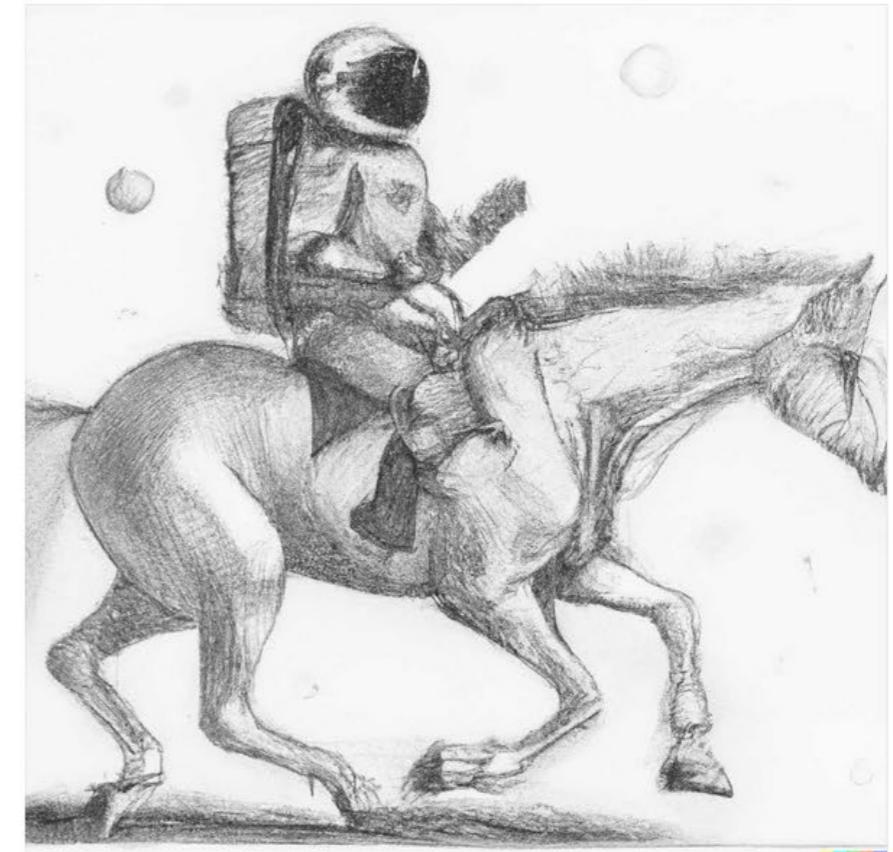
TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a photorealistic style in the style of Andy Warhol as a pencil drawing

DALL-E 2



Open Vocabulary Object Recognition

- Yolo World: <https://github.com/AILab-CVC/YOLO-World>
- Combine conventional object detector (YOLO) with multimodal CLIP representation



{men, women, boy, girl} {elephant, ear, leg, trunk, ivory} {golden dog, black dog, spotted dog} {grass, sky, zebra, trunk, tree}



LISA: Reasoning Segmentation via Large Language Model

- LISA: <https://github.com/dvlab-research/LISA>

- Segmentation + explanatory answers

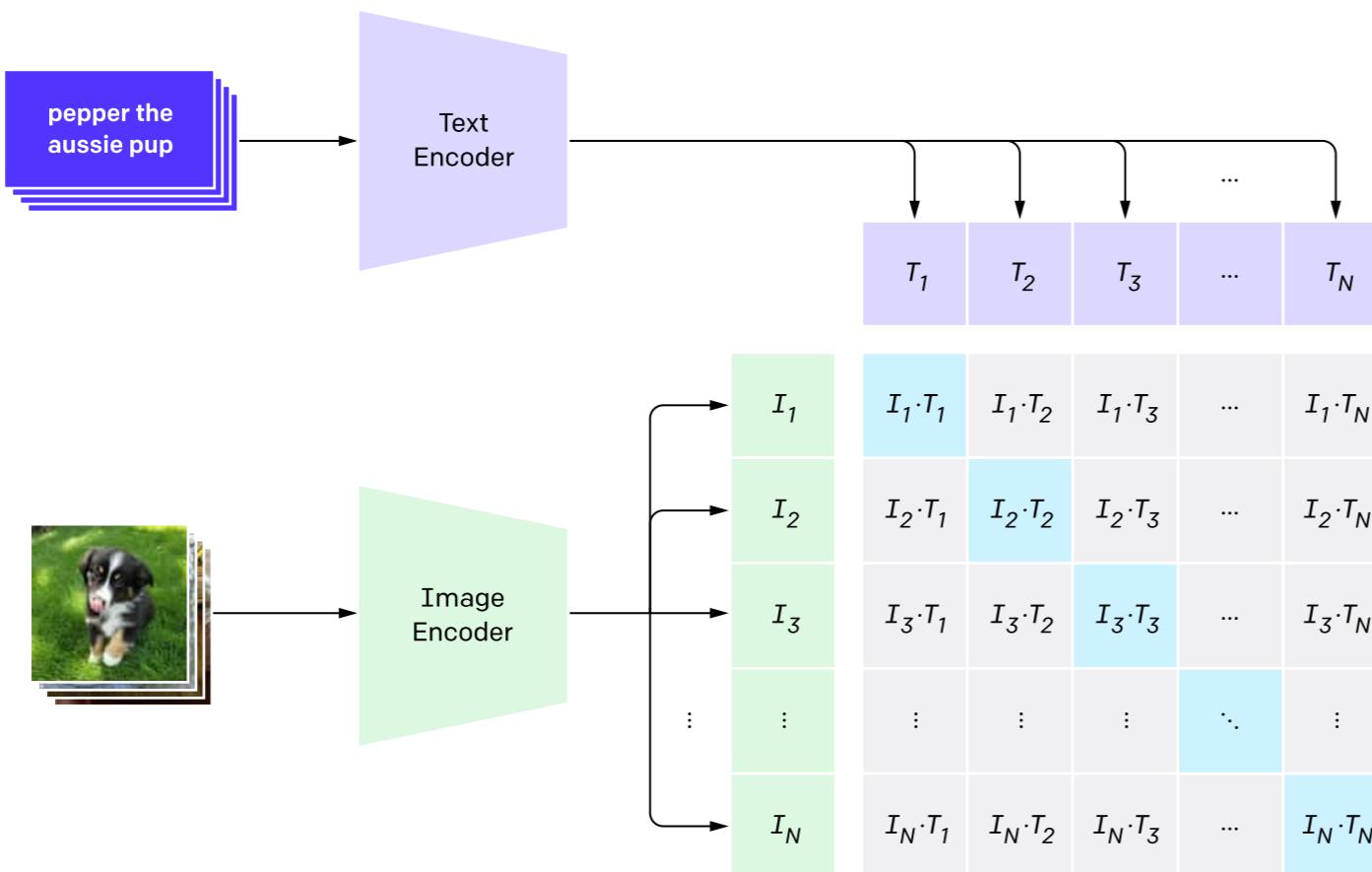
- Combines multimodal model (LLaVA) with segmentation model SAM (from Meta) for advanced image understanding

Input	Output	Input	Output
			
"Who was <u>the president of the US</u> in this image? Please output segmentation mask and explain the reason."	"Sure, the segmentation result is [SEG]. The President of the United States in the image is <u>President Obama</u> ."	"Who was <u>the president of the US</u> in this image? Please output segmentation mask and explain why."	"Sure, [SEG]. In the image, the President of the United States is <u>President Trump</u> ."
			
"What can <u>make the woman stand higher</u> ? Please output segmentation mask and explain why."	"Sure, [SEG]. The woman is <u>standing higher by using a ladder</u> ..."	"Can you segment <u>the camera lens that is more suitable for photographing nearby objects</u> in this image?"	"Sure, [SEG]."

Large Language Models Enter the Scene..

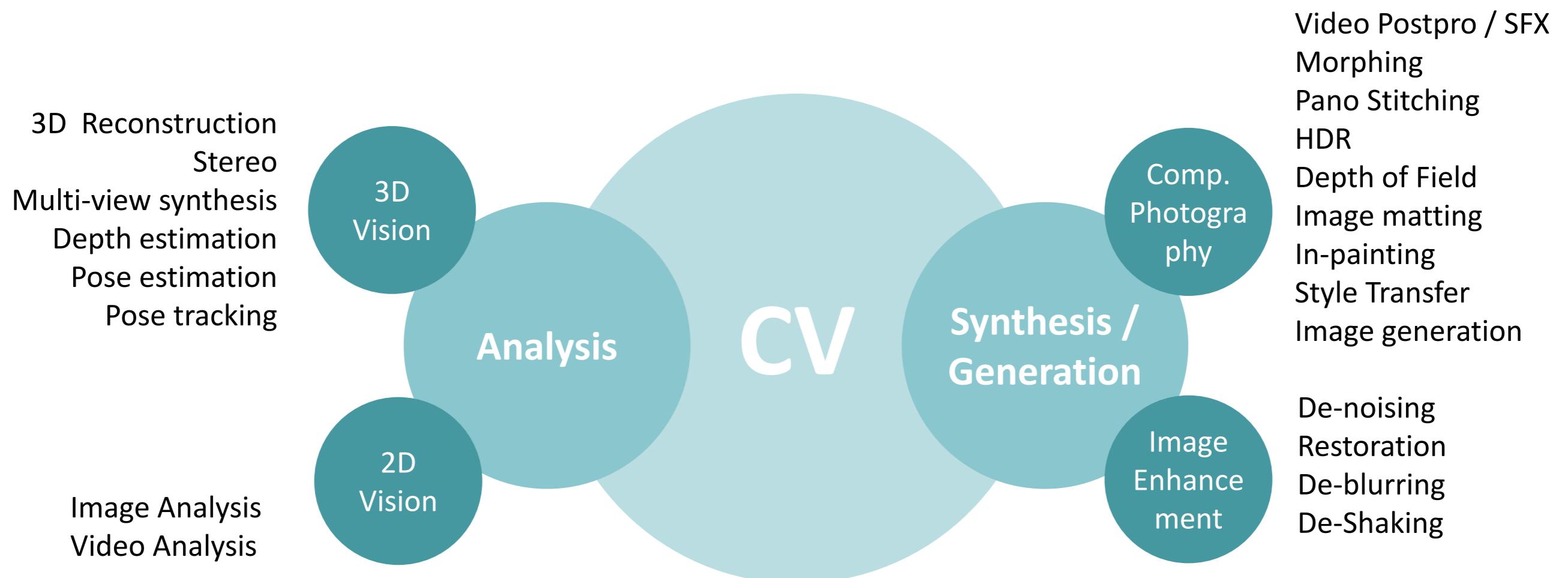
- CLIP: Multimodal (text – image) representations:
<https://openai.com/research/clip>

1. Contrastive pre-training



- More: <https://towardsdatascience.com/clip-model-and-the-importance-of-multimodal-embeddings-1c8f6b13bf72>

Computer Vision Applications



Next time

- Challenges of Computer Vision, i.e. why is computer vision hard?
- Image Formation – how is an image generated?
- Basic Image Operations, e.g. adding / subtracting images

Now:

- First Lab & first hands-on experience