Tsinghua University

THU×SENSETIME - 80231202

Advanced Computer Vision

Friday, February 25, 2022

Course Introduction

Overview

This course involves **computer vision**, **deep learning** and other fields of knowledge. It elaborates with the latest academic achievements and practical cases of industrial scenes and explain the classic and state-of-the-art methods in computer vision.

What we have

- Focus on Both Classics and Frontiers
- Combination of Academia and Industry
- Teaching from the shallower to the deeper
- GPU clusters for experiments

What you will learn

- Basic theories and advanced methods in Computer Vision
- Understand and explore practical problems in the industry
- Improve your research ability and innovative ability

What you need

- **Mathematics**
  - Calculus
  - Linear Algebra
  - Basic Probability and Statistics
- **Coding ability**
  - **Python** is recommended
- Machine Learning

Syllabus

Tsinghua University sensetime

Chapter 1 - Computer Vision Overview and Deep Learing Basics

- Basics of computer vision & image processing
- Introduction of the neural network and deep learning framework

1. Computer Vision Basics
2. Feature Detection
3. CNN & High-level Feature Extraction
4. Training Framework and Model Optimization

Chapter 2 - Advanced Computer Vision Tasks

- Cutting-edge research directions in computer vision
- The algorithm model optimization and performance improvement methods in visual scenes.

5. Image Classification
6. Object Detection
7. Image Segmentation
8. Video Understanding and Sequence Analysis
9. 3D Vision
10. Low-Level Computer Vision Task
11. Neural network Model Acceleration and Compilation
12. Representation Learning in Vision Tasks

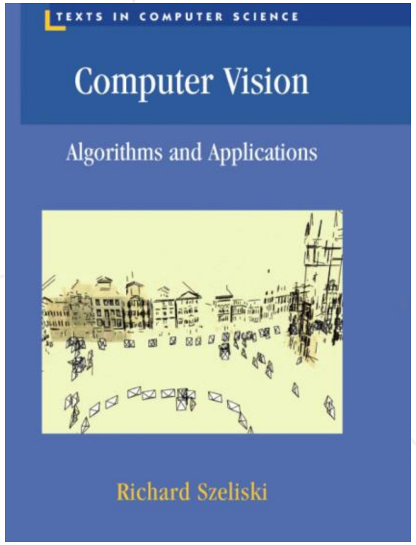
Chapter 3 - Lectures on industry applications

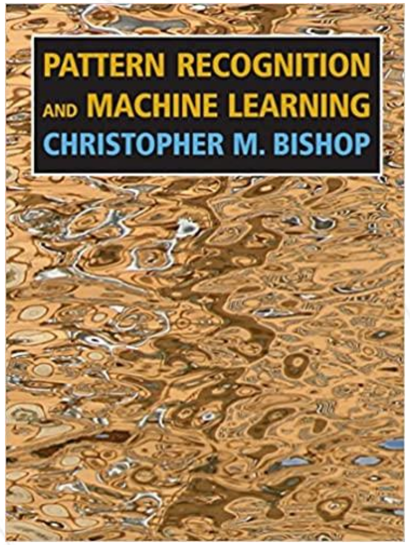
- The practical problems faced by computer vision and the solution ideas in combination with the specific scenes of industry.

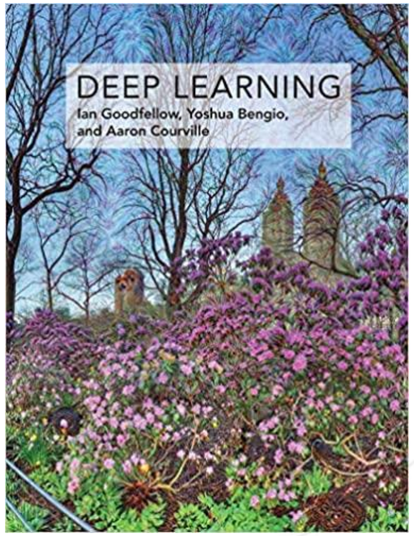
13. AutoPilot
14. 3D Vision and Augmented Reality

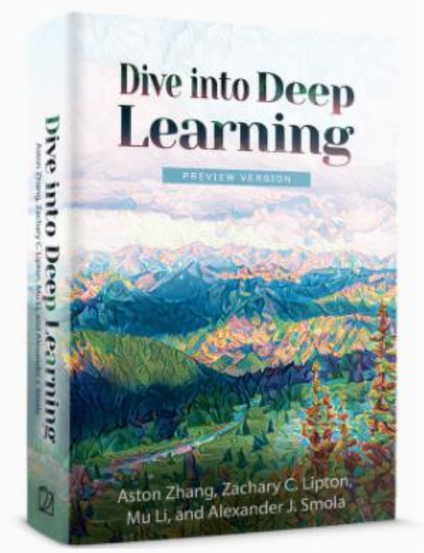
Course Introduction

- **Textbook**
  - ***Computer Vision Algorithms and Applications***
    - by Richard Szeliski
    - Preview version： [Link]
  - ***Pattern Recognition and Machine Learning***
    - by Christopher Bishop
    - Free online version： [Link]
  - ***Deep Learning***
    - by Goodfellow, Bengio, and Courville
    - Index： [Link]
  - ***Dive into deep learning***
    - An interactive deep learning book with code, math, and discussions, based on the NumPy interface
    - Free online version： [Link]









- **Assignment & Final Project**

Assignments （30%）

- 1 Assignments finish after class by one person
- You can finish assignment on your local machines or on clusters provided by SenseTime
- **Topic**
  - Advanced Computer Vision Task
- **Released Date - Due Date**
  - March. 25 - Apr. 8

Final Project （70%）

- Collaboration in groups of up to 3 people
- Choose one topic and finish the project
- **You should submit**

1.One page proposal and discuss it with TAs (topic, idea, method, experiments)

2.A term paper of 4 pages (excluding figures) in maximum

3.Code and sample data

4.Project presentation

- # Final Project

Mar. 11

Final Project release

Apr. 20

Submit topics of the final project

Apr. 21

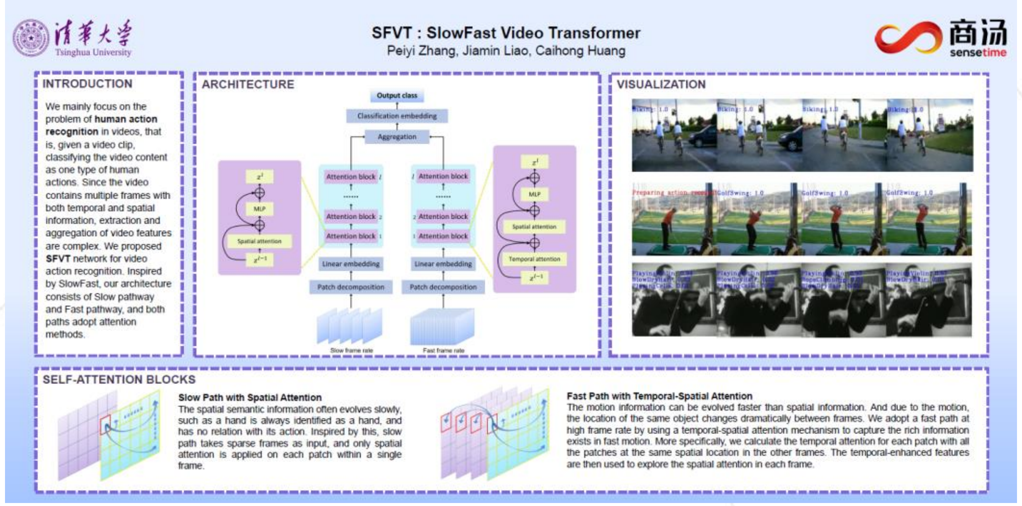
Tutorial (optional to attend): Discuss with TA-in-charge

May 9

Submit proposal (1-2 pages)

May 28

Final project Seminar (10 minutes presentation and 3 minutes Q&A)



Course Introduction

- ## Instructors

Dr. Li Yali

- Tsinghua EE Assistant Reseacher
- liyali13@mail.tsinghua.edu.cn

Dr. Dai Jifeng

- SenseTime Executive Research Director
- daijifeng@sensetime.com

Dr. Li Hongyang

- SenseTime Senior Research Manager
- lihongyang@sensetime.com

- ## TAs

Dr. Wang Han

- i@hann.wang

- ## Coordinators

Chen Qingchen

chenqingchen@sensetime.com

Zhang Qifan

zhangqifan@sensetime.com

Course Introduction

- **Lecture Time & Venue**
  - **Friday**, 9:50am-11:25am
  - **1102**, No.3 Teaching Building
- **Optional Tutorials & QA Time**
  - **Thursday**, 19:00-20:00
  - Tencent Meeting Room： 785 271 5223
- **Course Homepage**
  - https://thu-acv.github.io
- **Discussions**
  - WeChat Group
  - Tencent Meeting Room： 785 271 5223

2022春-THU高等计算机视觉

该二维码7天内(3月2日前)有效，重新进入将更新

商汤学术小助手

中国大陆

Tsinghua University

Advanced Computer Vision

THU×SENSETIME – 80231202

sensetime

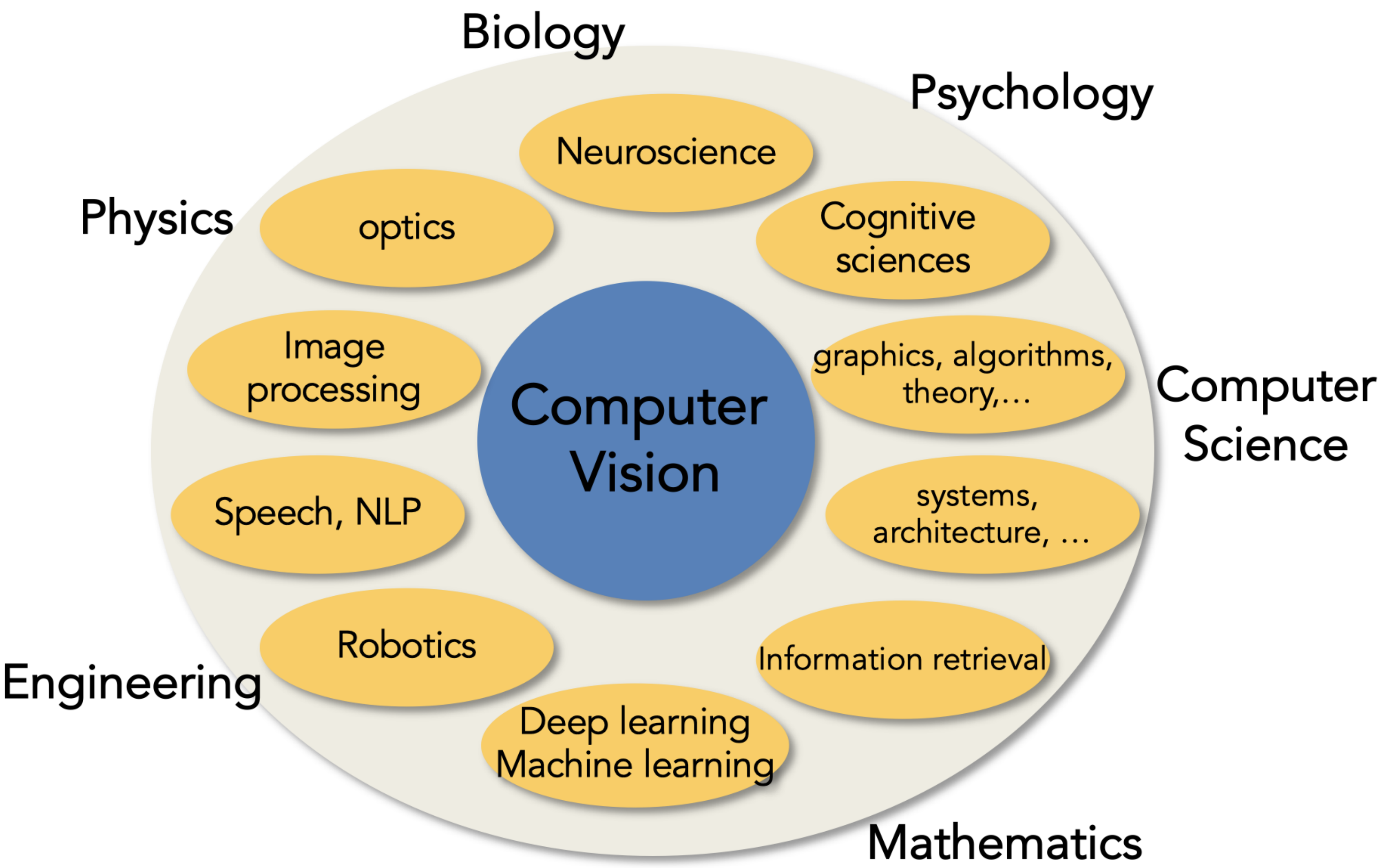
Chapter1 - Section 1 Part 1

Computer Vision Basic

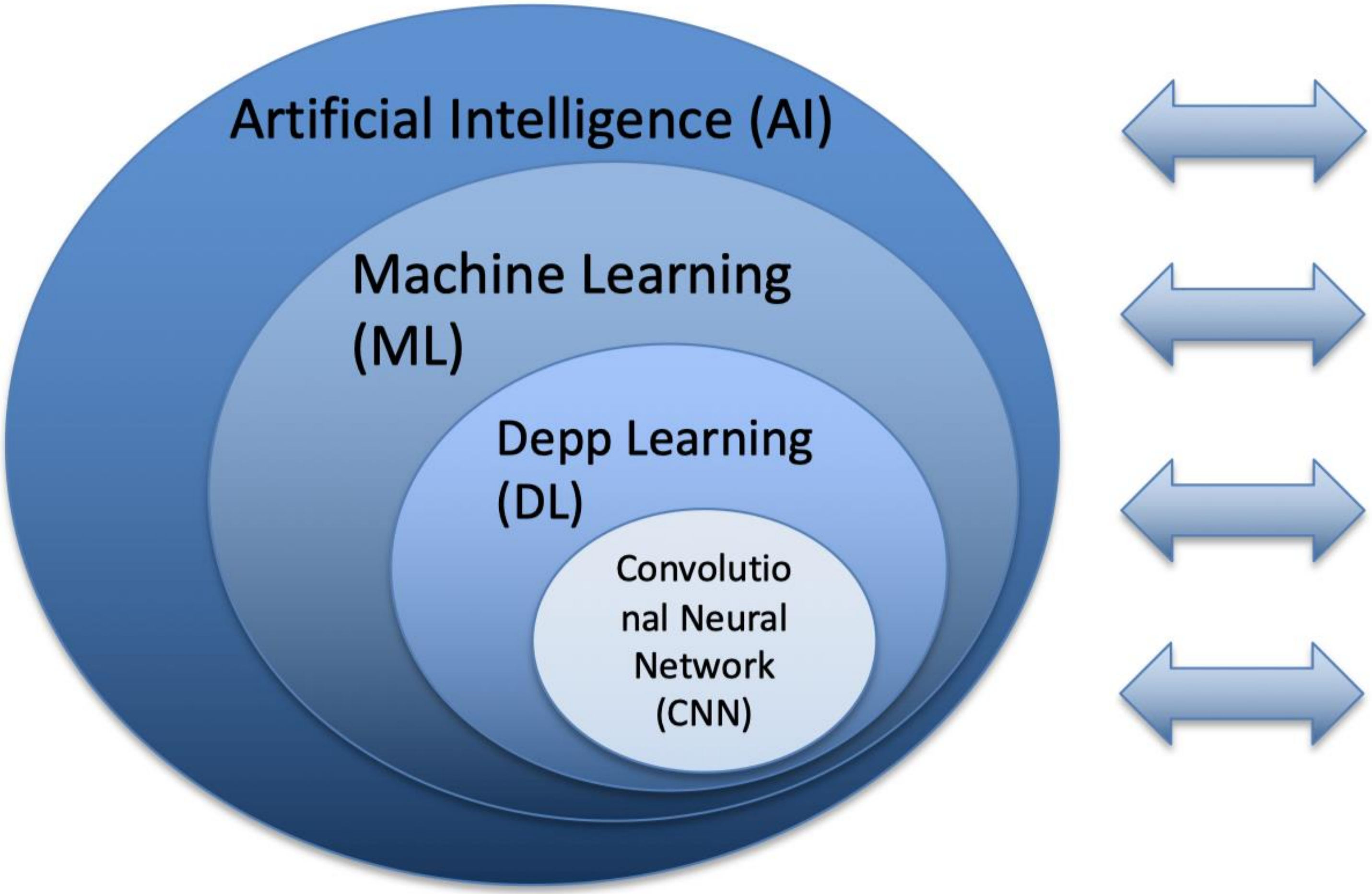
Dr. Dai Jifeng

Friday, February 25, 2022

What's Computer Vision



What's Computer Vision

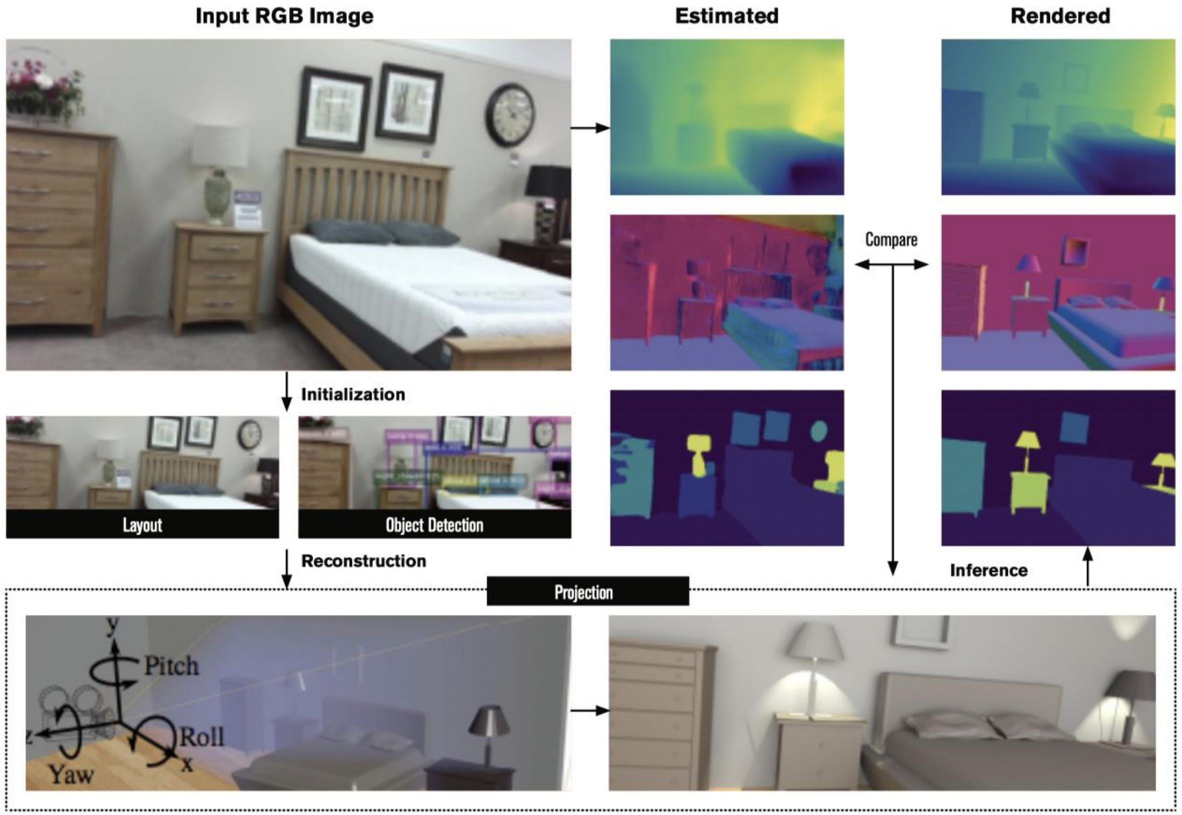


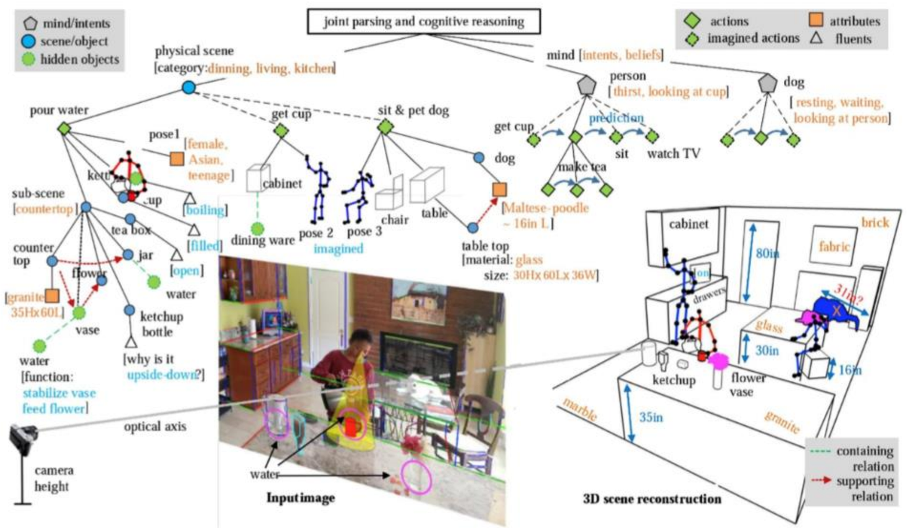
Computer Vision

- Object detection
- Object classification
- Scene understanding
- Semantic scene segmentation
- 3D reconstruction
- Object tracking
- Human pose estimation
- Activity recognition
- VQA
- ....

What's Computer Vision

Vision is the most important source of information for the human brain and is the “**entrance hall**” of AI.



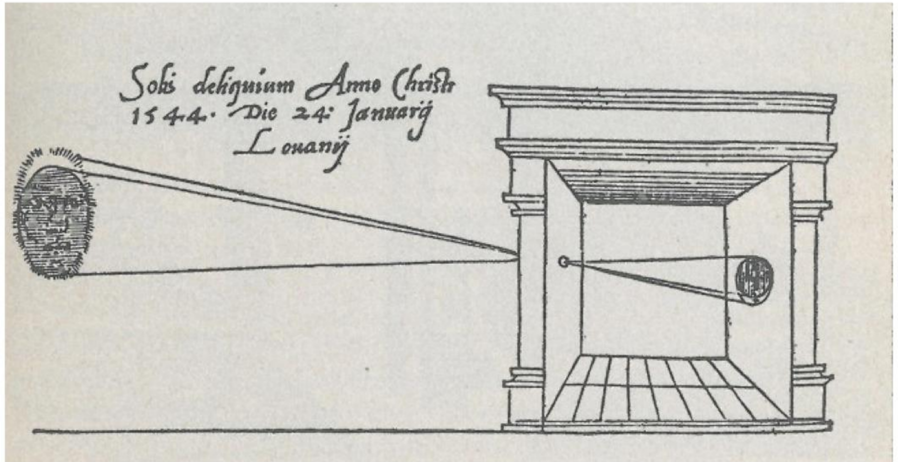


- **Biological Vision**

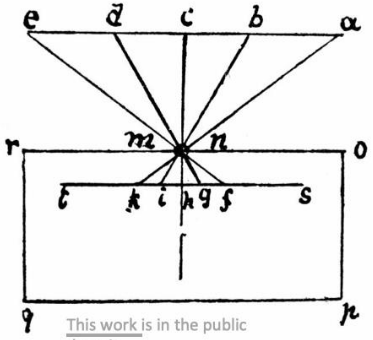
- **Ancient Human Vision**

Camera Obscura

Gemma Frisius, 1545



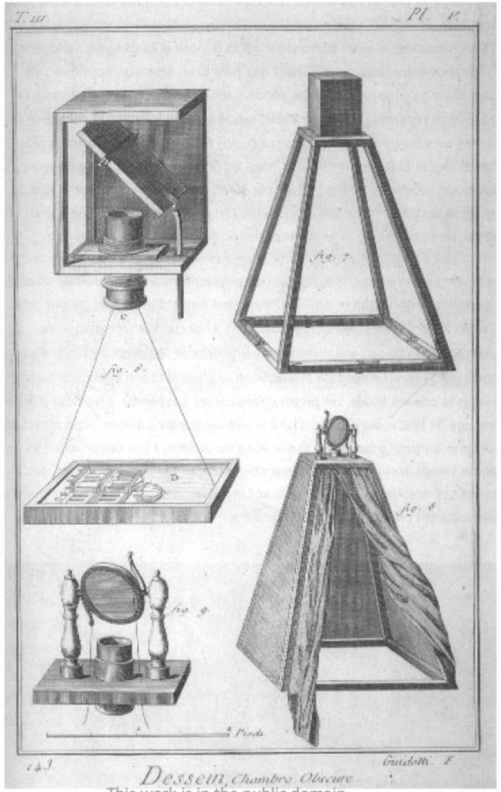
This work is in the public domain



This work is in the public domain

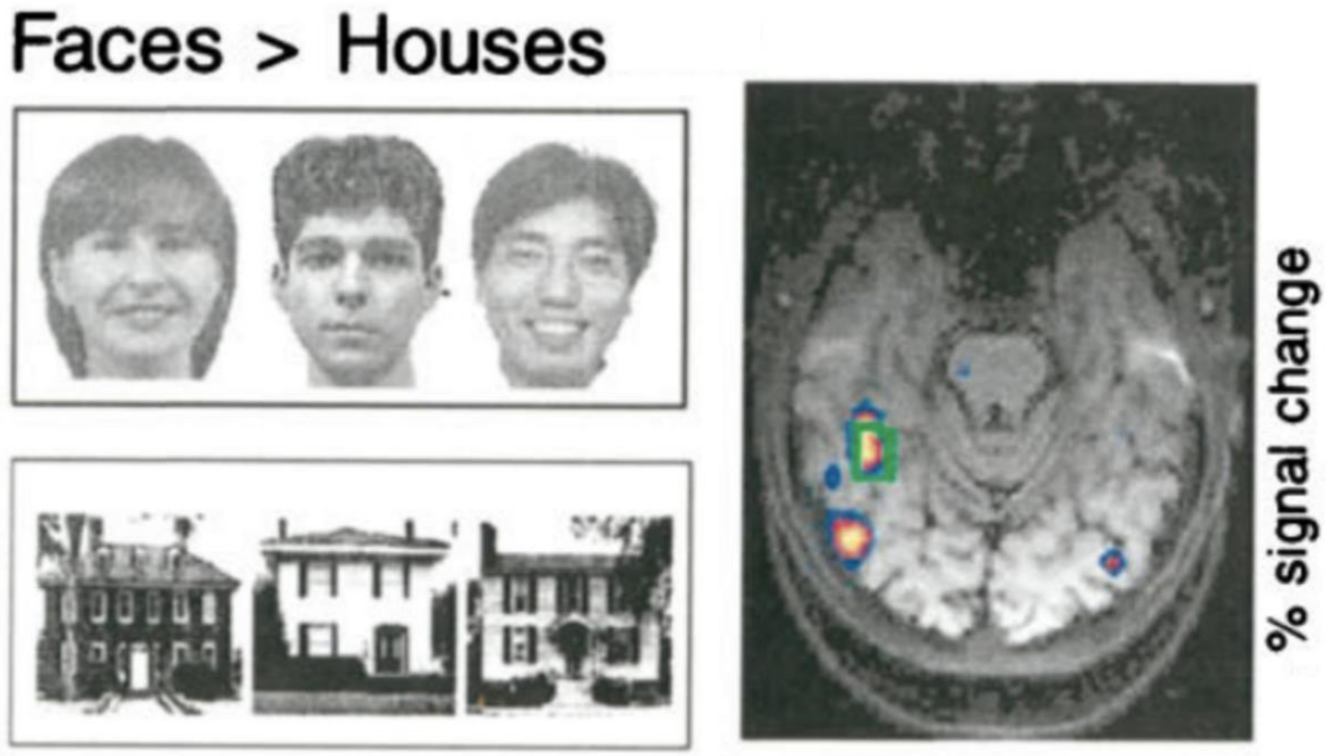
Leonardo da Vinci, 16th Century AD

Encyclopedia, 18th Century

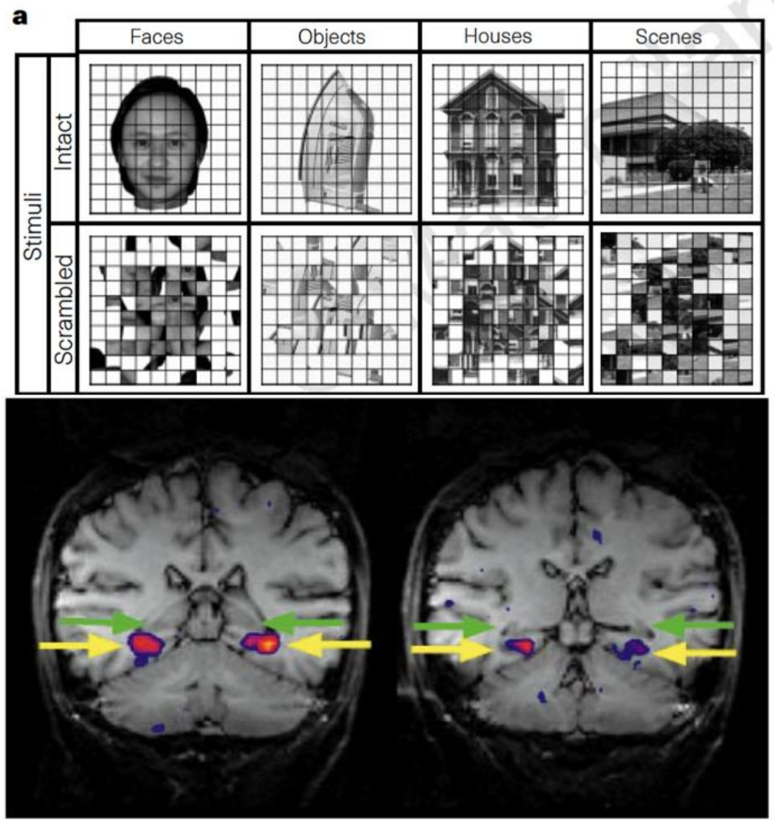


This work is in the public domain

- **Neuroscience and Vision**

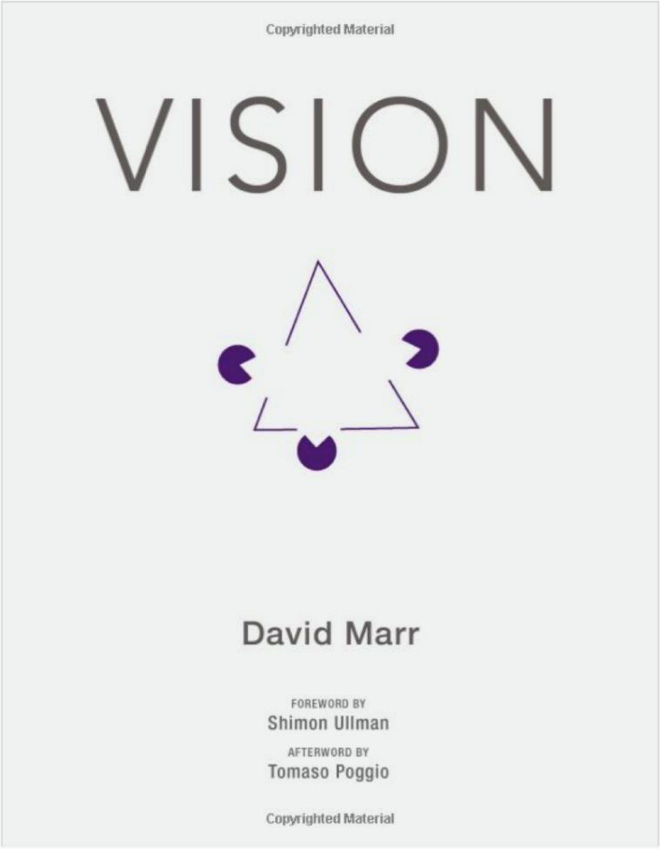


Kanwisher et al. J. Neuro. 1997



Epstein & Kanwisher, Nature, 1998

- **Marr Computational Vision**

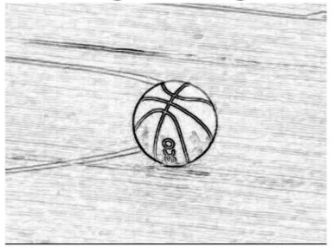


3D Reconstruction
Not talent, but
computation

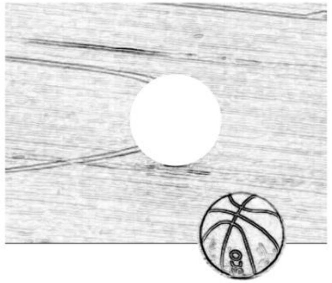
- **Marr Computational Vision**

Input image

Edge image



2 ½-D sketch



3-D model

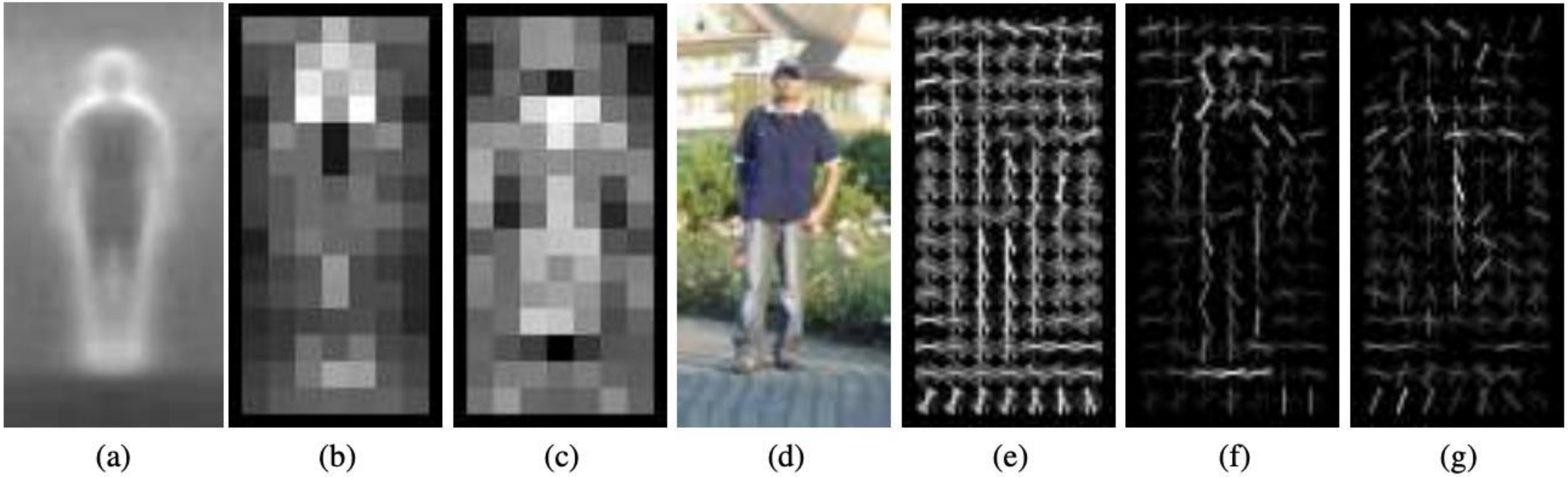
Input Image	→	Primal Sketch	→	2 ½-D Sketch	→	3-D Model Representation
Perceived intensities		Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries		Local surface orientation and discontinuities in depth and in surface orientation		3-D models hierarchically organized in terms of surface and volumetric primitives

Stages of Visual Representation, David Marr, 1970s

- **Feature Detection——SIFT**

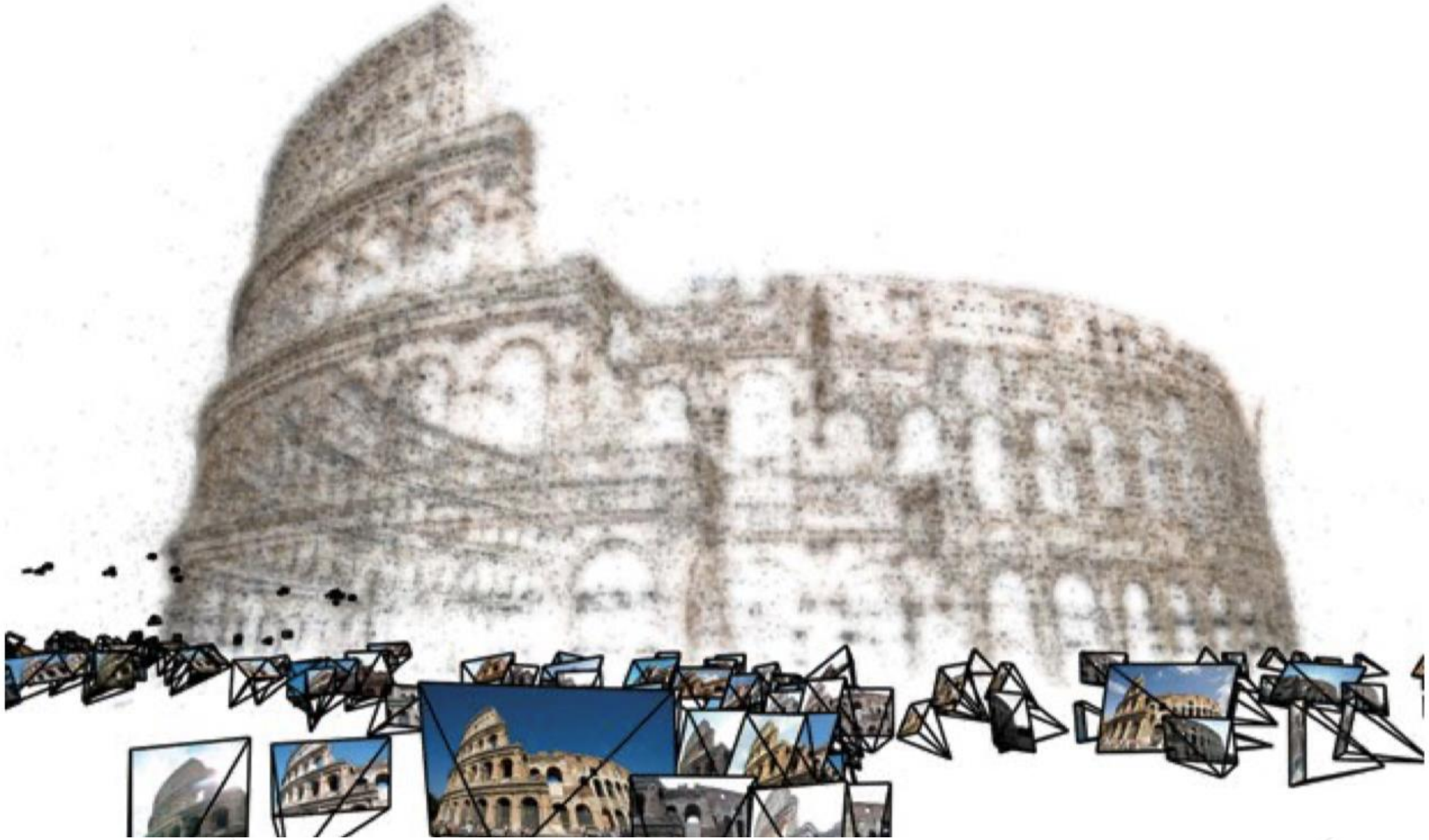
https://www.cs.ubc.ca/~lowe/papers/iccv99.pdf

- **Feature Detection——HOG**



https://web.archive.org/web/20110408220331/
http://www.acemedia.org/aceMedia/files/document/wp7/2005/cvpr05-inria.pdf

- **3D reconstruction**



Agarwal et al.
ICCV, 2009

- **Image Classification**

Caltech 101 images

Fei-Fei et al. 2004

Visual Object Classes Challenge 2009 (VOC2009)

PASCAL2

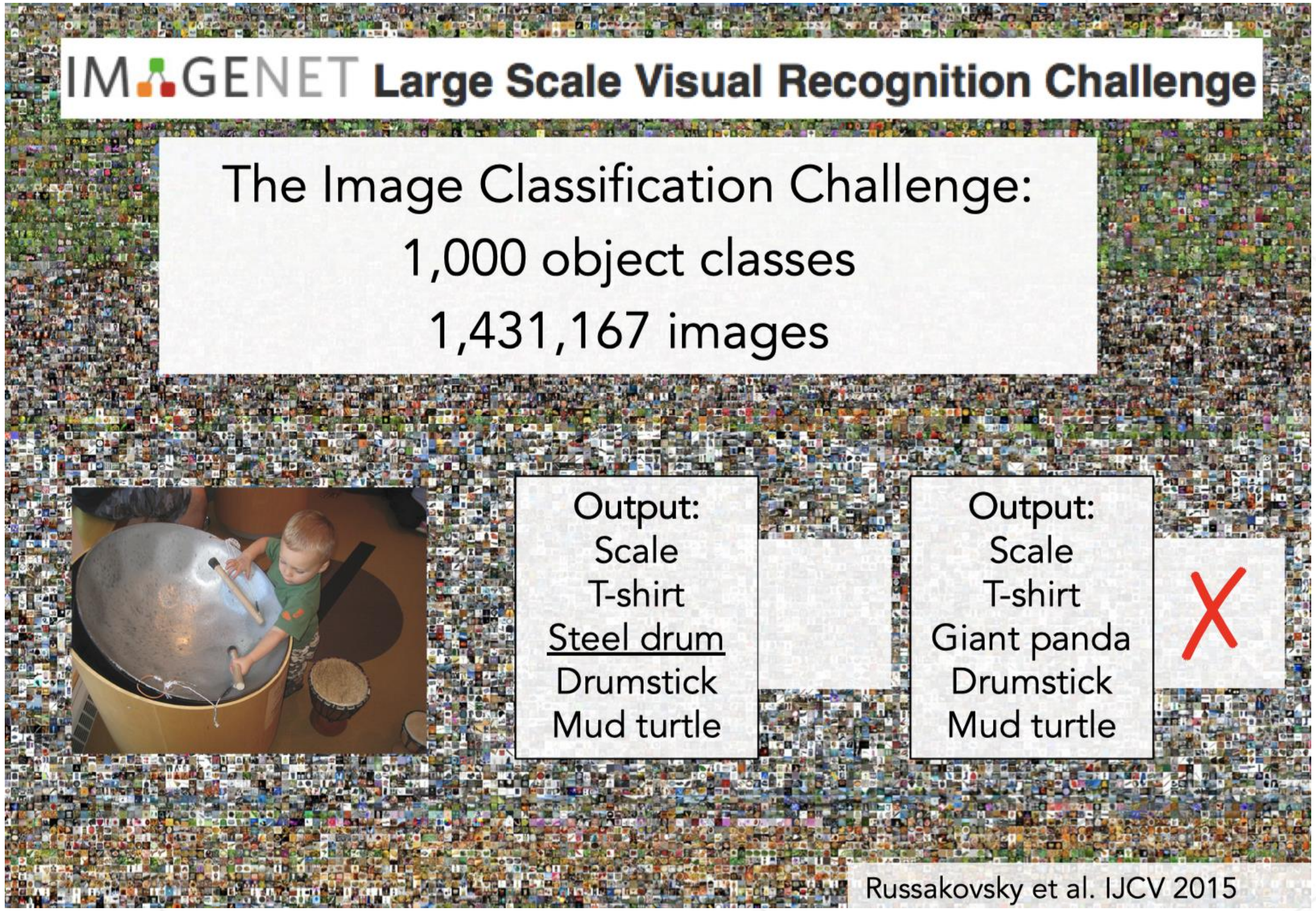
Pattern Analysis, Statistical Modelling and Computational Learning

[click on an image to see the annotation]

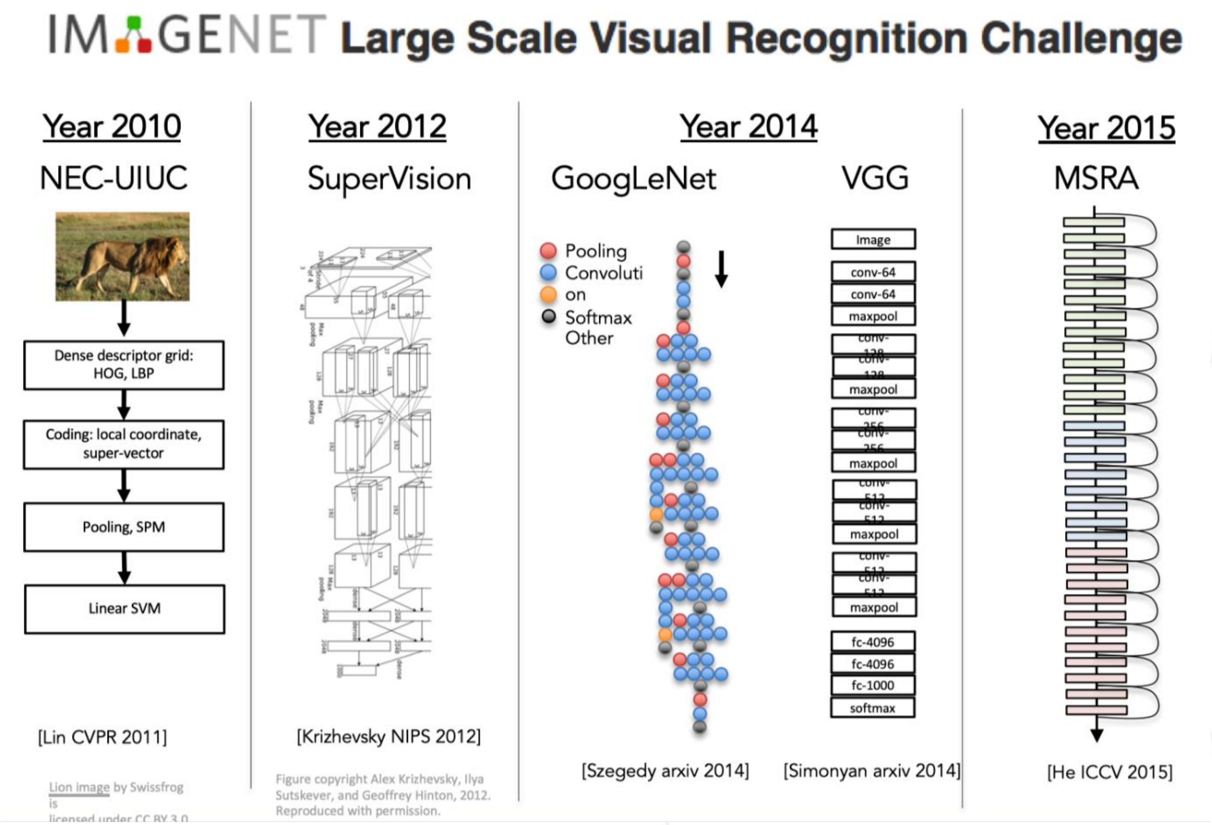
Everingham et al. 2006-2012

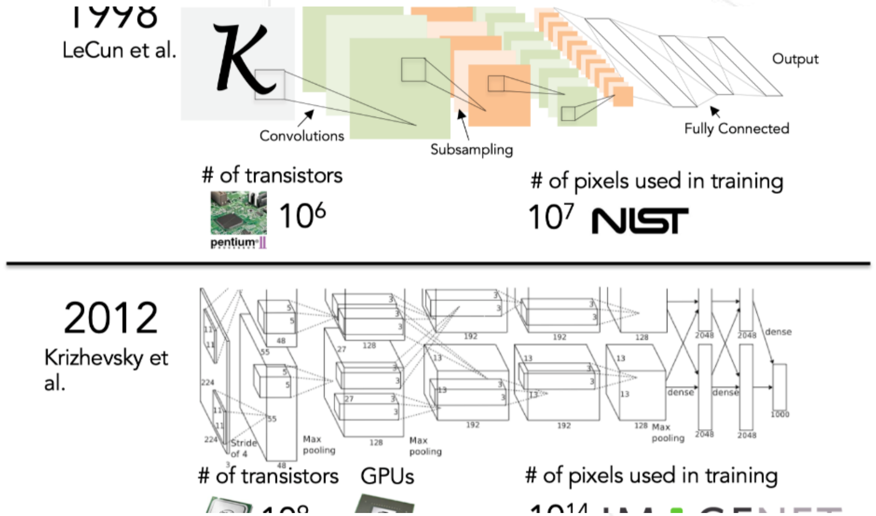
- **IMAGENET Challenge**

- **IMAGENET Challenge**



- **IMAGENET Challenge**





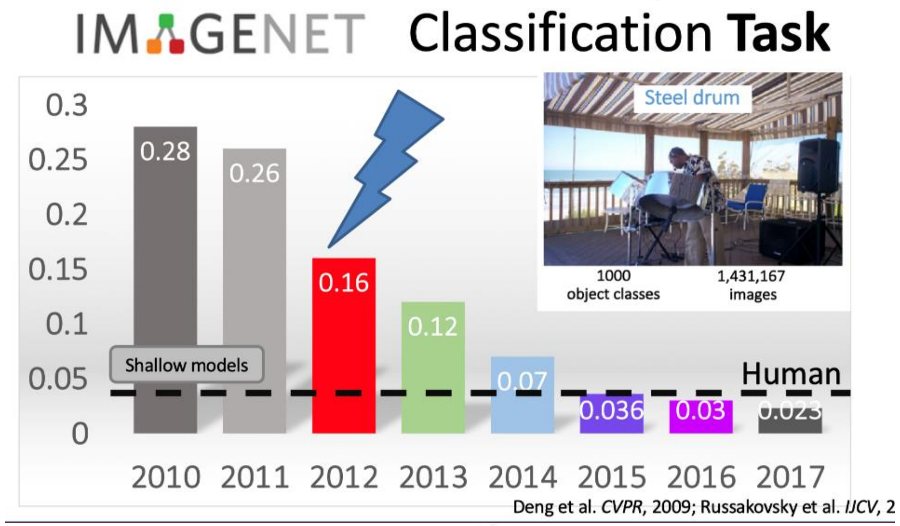
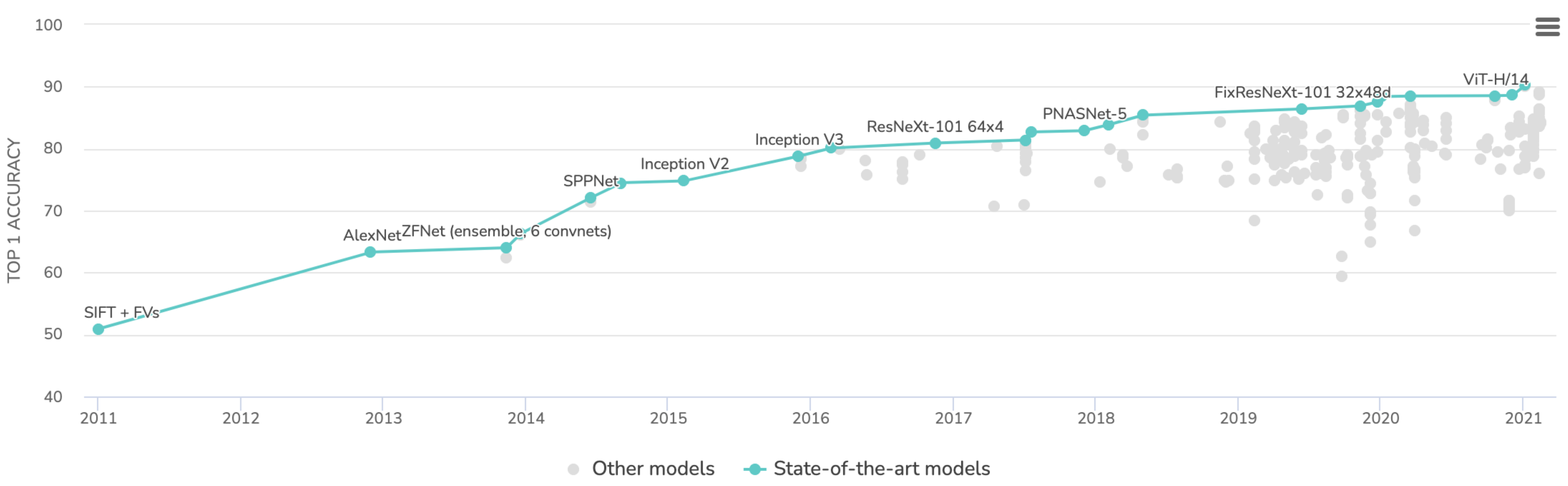


Image Classification on ImageNet

Leaderboard Dataset



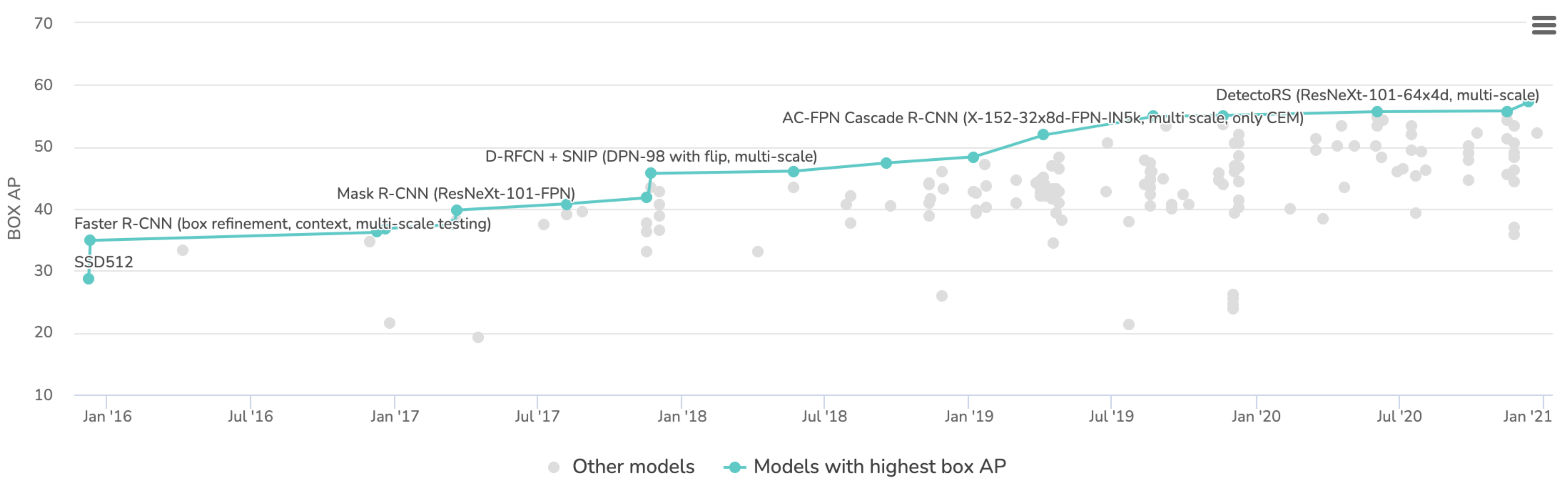
- **Object Detection**

https://cocodataset.org/

Tsinghua University sensetime

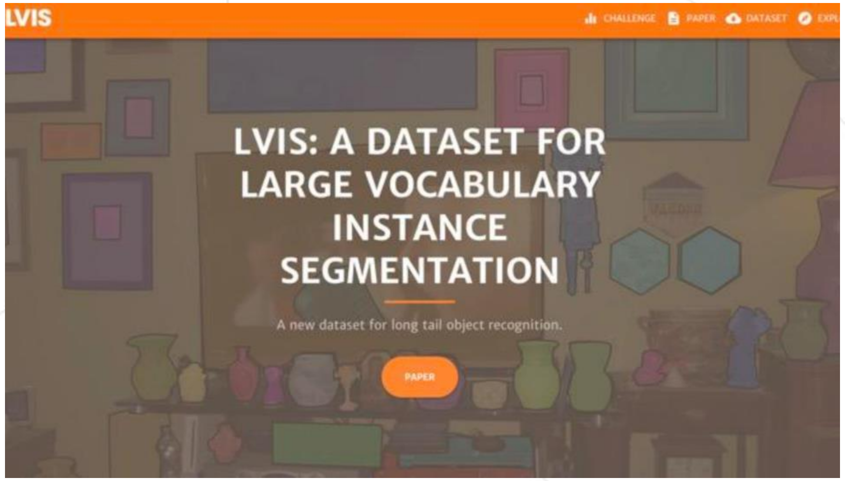
Object Detection on COCO test-dev

Leaderboard Dataset



- **Instance Segmentation**

https://www.lvisdataset.org/explore

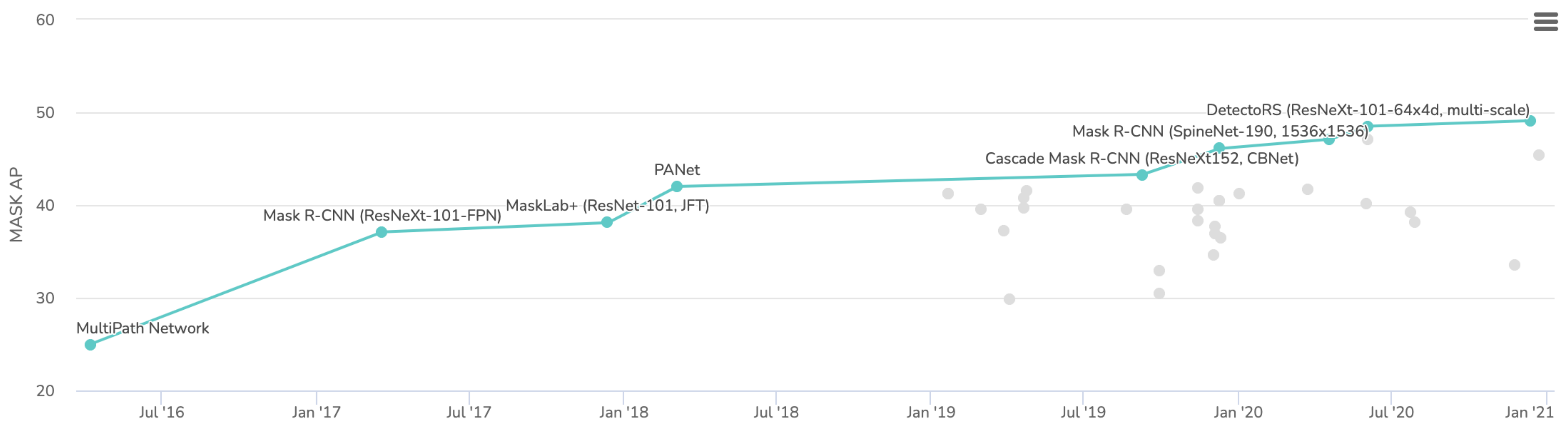


- **Semantic Segmentation and Instance Segmentation**

Input Image | Semantic Segmentation | Instance Segmentation

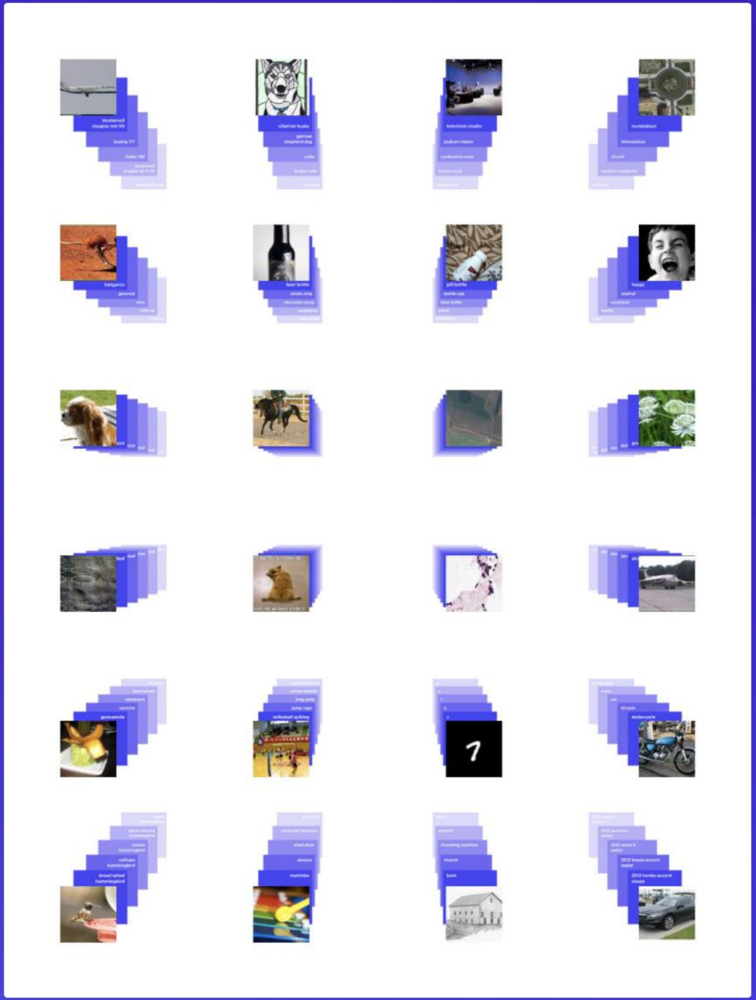
Instance Segmentation on COCO test-dev

Leaderboard Dataset



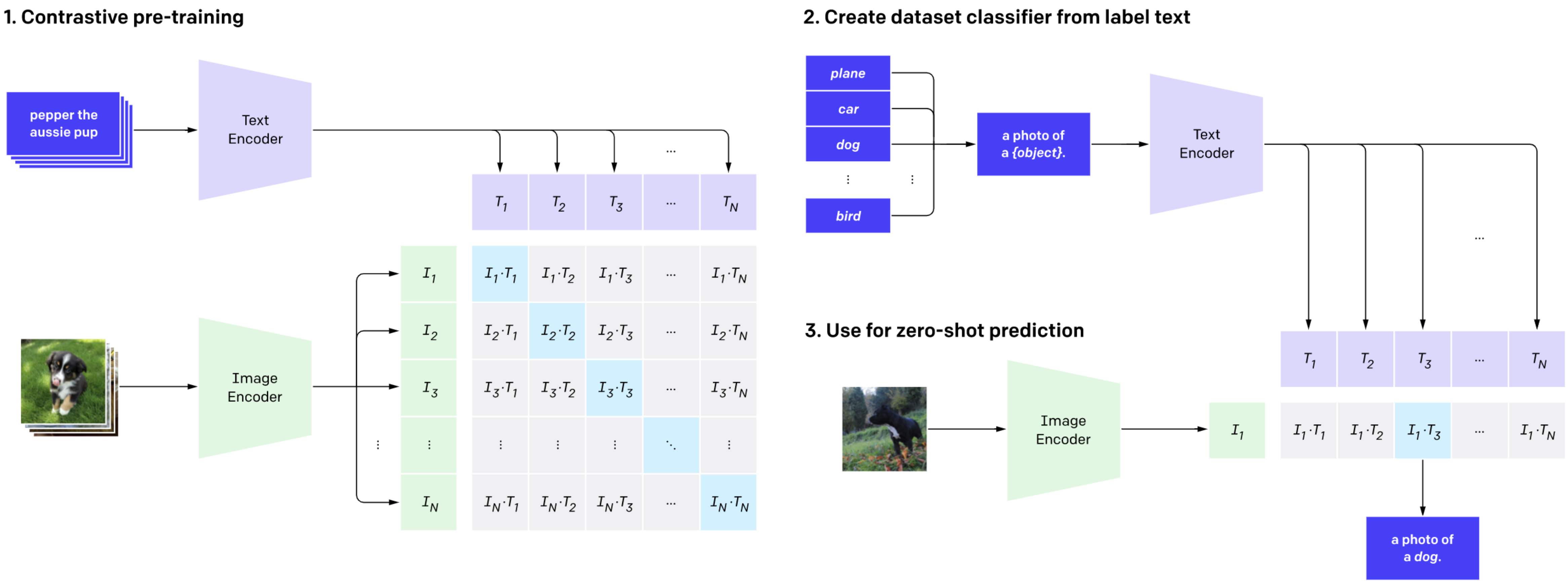
CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

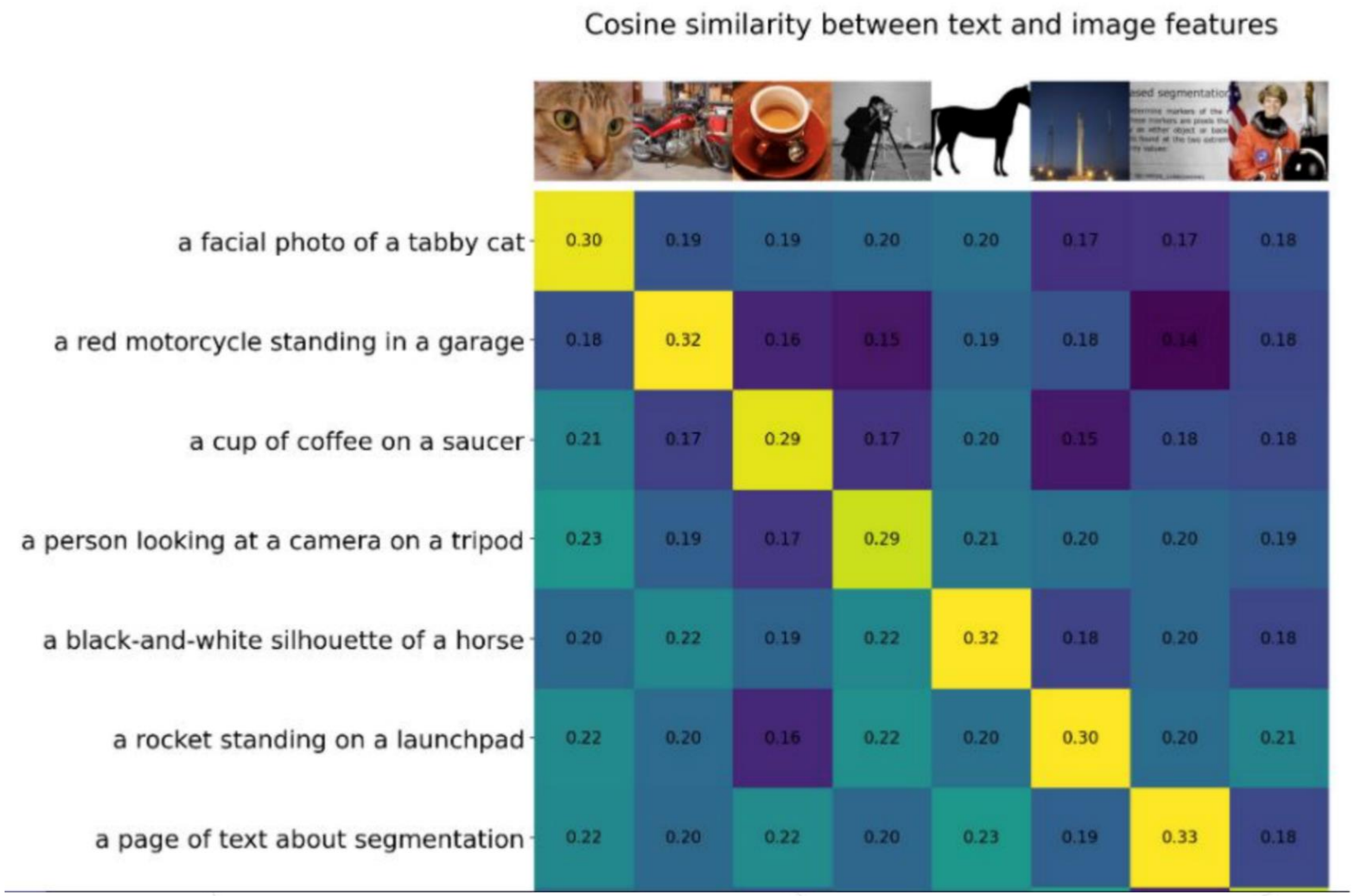


January 5, 2021
15 minute read

- **CLIP: Connecting Text and Images**



- **CLIP: Image-Text Match**



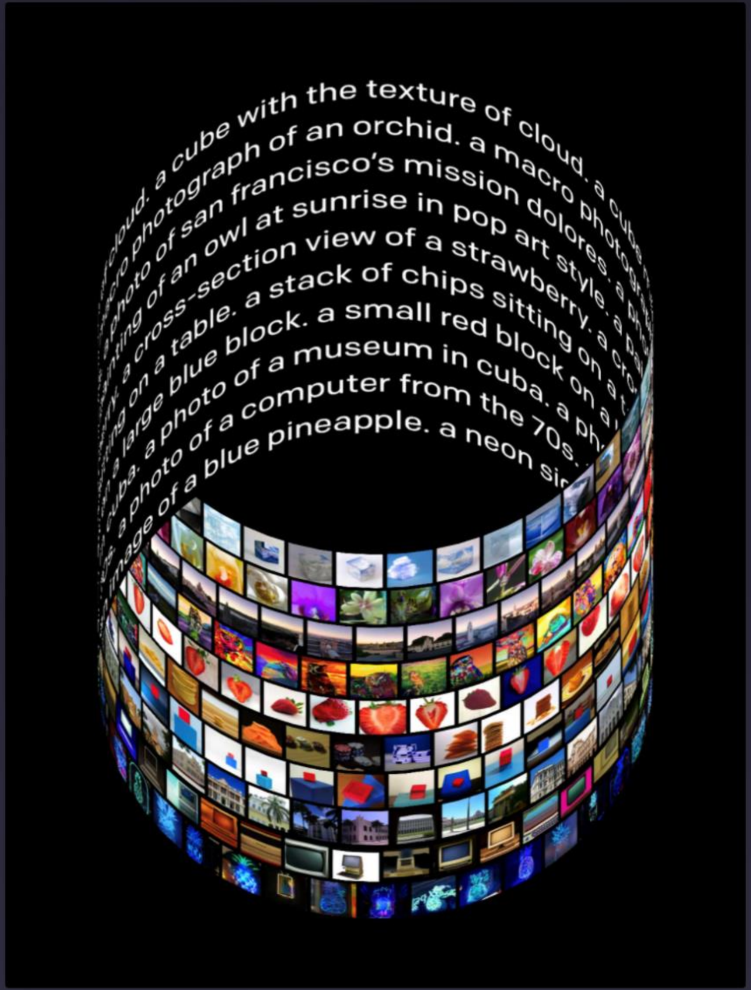
Tsinghua University

sensetime

DALL·E: Creating Images from Text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.

January 5, 2021
27 minute read



TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT

a store front that has the word ‘openai’ written on it [...]

AI-GENERATED IMAGES

Edit prompt or view more images↓

DALL-E
Creating Images from Text

TEXT PROMPT a stained glass window with an image of a blue strawberry

AI-GENERATED IMAGES

- **Low-level Vision**

SenseTime – Pioneer in Deep Learning and Computer Vision

Tsinghua University · sensetime

Founding of CUHK MMLab by Prof. Tang Xiao'ou

CUHK MMLab is the **earliest Chinese** research team to work on **deep learning**

MMLab's CV Papers published **almost equal to** the total amount published by all Chinese universities

MMLab's deep learning papers published in CV conferences are **almost half of** the total amount papers published **globally**

MMLab's DeepID algorithm **exceeds human eye accuracy** for the first time in history

SenseTime incorporated

SenseTime became 1st Chinese company to **win** ImageNet competition, and was ranked top in video analysis

IMAGENET

SenseTime **won three** world champions in five key ImageNet competitions

IMAGENET

SenseTime **surpassed Google and Facebook** in CVPR & ICCV (1) Paper Submission

MIT- SenseTime Alliance was formed to support AI Research

SenseTime and its joint labs published 62 papers and 57 papers in 2019 CVPR and ICCV respectively, ranking 1st in the world (2)

SenseTime and its joint labs published 62 papers in CVPR 2020

2001 | 2004-2008 | 2011 | 2011-2013 | 2014.6 | 2014.10 | 2015.11 | 2016.9 | 2017.10 | 2018.2 | 2019.7 | 2020.6

2011.8 | 2012.12 | 2013.3 | 2013.12 | 2014.11 | 2016.3 | 2016.7 | 2017.3 | 2018 | 2019.3 | 2020

2011.8 – Microsoft: Microsoft significantly improved the accuracy of Deep Learning-driven voice recognition

2012.12 – IMAGENET: Hinton, the originator of Deep Learning, won the ImageNet visual recognition competition

2013.3 – Google: Google employed Prof. Hinton for US$50 million

2013.12 – facebook: Facebook established an AI Lab in New York and employed Yann LeCun

Google DeepMind: Google acquired DeepMind, a Deep Learning company, for US$660 million

2014.11 – AlphaGo: DeepMind's AlphaGo, the Go AI, beats Lee Sedol in Go

cruise, GM: GM acquired Cruise Automation, a startup in autonomous vehicle technology for US$1 billion

2016.3 / 2016.7 – SoftBank, ARM: Softbank launched a US$100 billion investment fund that focuses on AI and completed US$32 billion acquisition of ARM

2017.3 – intel, MOBILEYE: Intel acquired Mobileye, a leader in computer vision for autonomous driving technology

2018 – TESLA, WAYMO: Tesla launched self-driving vehicle. Waymo launched self-driving taxi service

NVIDIA: NVIDIA launches its own GPU cloud

2019.3 – 

Hinton, LeCun & Bengio received the Turing Award, the "Nobel Prize" of Computing

2020 – Global tech companies are exploring the use of AI to combat the outbreak of COVID-19.

(1) CVPR、ICCV、ECCV are the top 3 computer vision conferences worldwide with highest impact factor They accept the best work on computer vision and deep learning

(2) Based on statistics released by different companies and organizations to date

How to Generate the Best AI

清華大學 Tsinghua University 商汤 sensetime

Expertise	Data	Computing Power	Positive Feedback Loop
Fundamental research & technological capabilities determine rate of innovation	Large amount of high quality data fuels the algorithm iteration	Super fast computing power ensures speed of training	Vertical partnerships ensure technology and data feedback for adaptive improvement

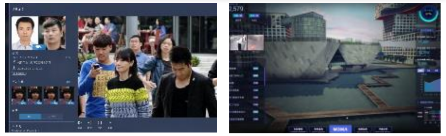
商汤 sensetime

SenseTime Excels at All of These Core Capabilities

SenseTime – World Leading AI Innovation Platform

Tsinghua University | sensetime

Smart City


Smart Surveillance | Smart City Management System

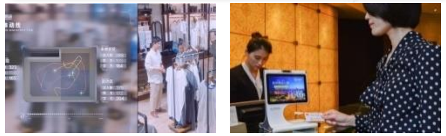

Smart Traffic Management | Fire Detection


Smart Crowd Management | Abnormal Behavior Detection


Garbage Detection | Illegal Parking Detection


Illegal Occupation Detection | Abnormal Objects Detection on Road

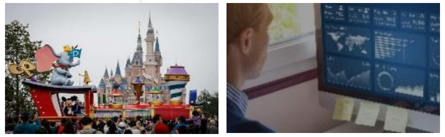
Business Intelligence


Retail Analytics Solutions | Intelligent Hotel Check in System


Smart Airport Solution | Smart Metro Solution


Smart Office Management System | Smart Tourism Area Management


Smart Entertainment Solution | Smart Campus Solution

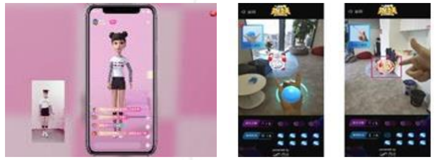

Smart Amusement Park Solution | Real Estate Sales Management

Mobile Solution


Face Unlock | Photo Processing

Image Super Resolution | 3D Face Beautification

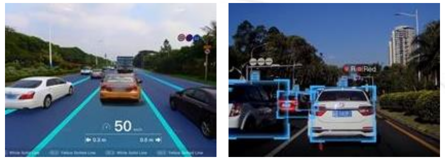
AR Platform


AR Live Streaming | AR Game

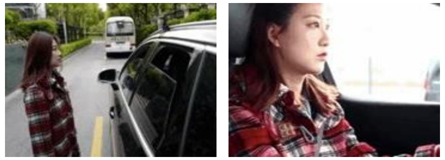

AR Classroom | AR Effect

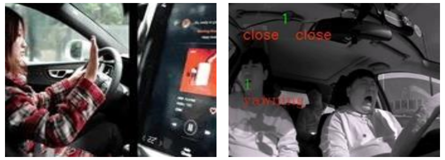
Autonomous Driving


Guide Line Prediction | Human Face Prediction

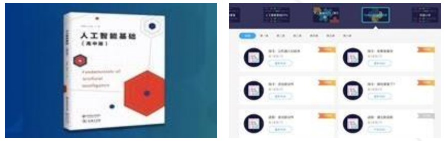

Lane Detection | Front Vehicle Detection

Intelligence Cabin Sensing


Face Unlock | Gaze Tracking


Gesture Tracking | Drowsiness detection

AI Education Package

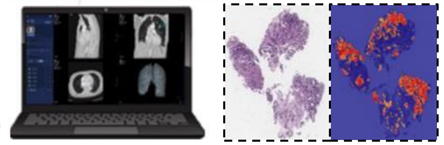

AI Textbook | AI Experiment Platform


AI RobotCar | AI Lab

Remote Sensing


Road Network Extraction | Cloud and Snow Detection

AI-Enabled Diagnosis, Treatment and Rehabilitation


Lung AI Application | Pathology Application

IN THE MOOD FOR LOVE

WONG KAR-WAI'S

IN THE MOOD FOR LOVE

Tsinghua University

Advanced Computer Vision

THU×SENSETIME – 80231202

sensetime

Chapter1 - Section 1 Part 2

Image and Video Processing

Dr. Dai Jifeng

Friday, February 25, 2022

Outline

Part 1	Image and video representation
Part 2	Image processing
Part 3	Video processing

Highlights

Image & video representation in computer

Basic applications of image processing

Traditional video processing and feature extraction methods

Common algorithms for image and video compression

History of digital image processing

商汤 sensetime

清華大學 Tsinghua University

Outline

Part 1	Image and video representation
Part 2	Image processing
Part 3	Video processing

Image & video Representation

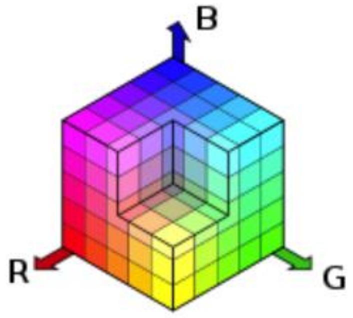
- Image -- A 2D discrete signal

111	115	113	111	112	111	112	111
135	138	137	139	145	146	149	147
163	168	188	196	206	202	206	207
180	184	206	219	202	200	195	193
189	193	214	216	104	79	83	77
191	201	217	220	103	59	60	68
195	205	216	222	113	68	69	83
199	203	223	228	108	68	71	77

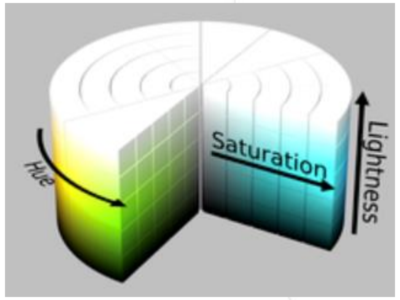
- Video -- Sequences of images

time

- **Color Model**
  - RGB



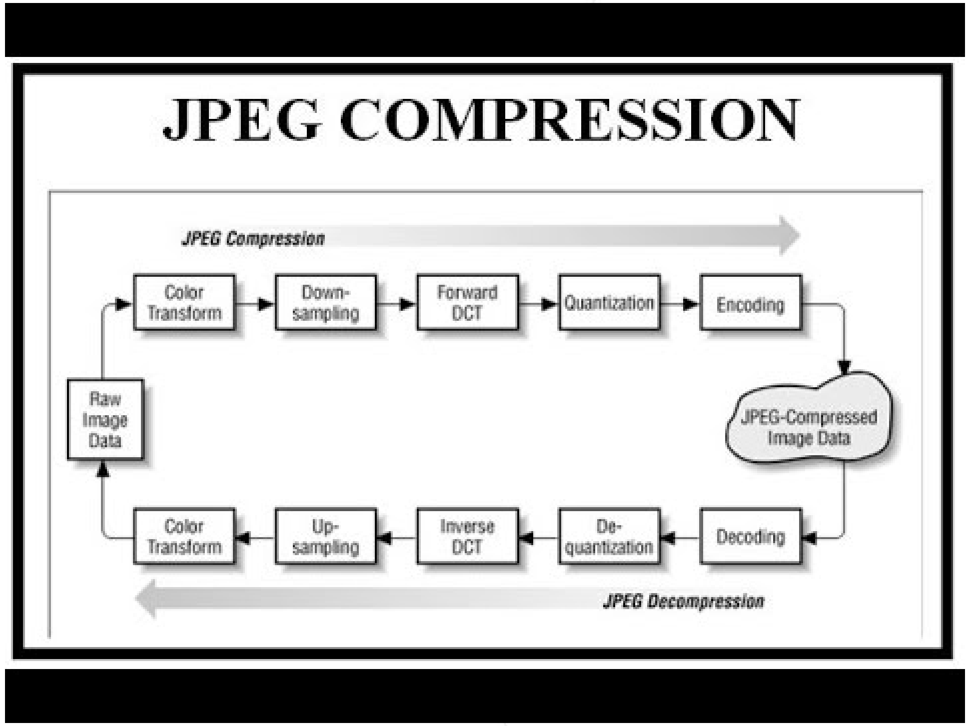
  - HSL



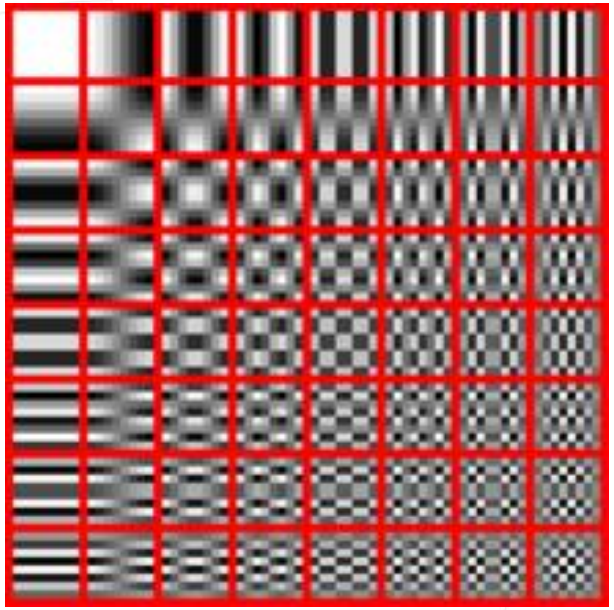
  - CMYK

Image & video Representation

- **Compression methods for image**
  - JPG, PNG, GIF, Webm

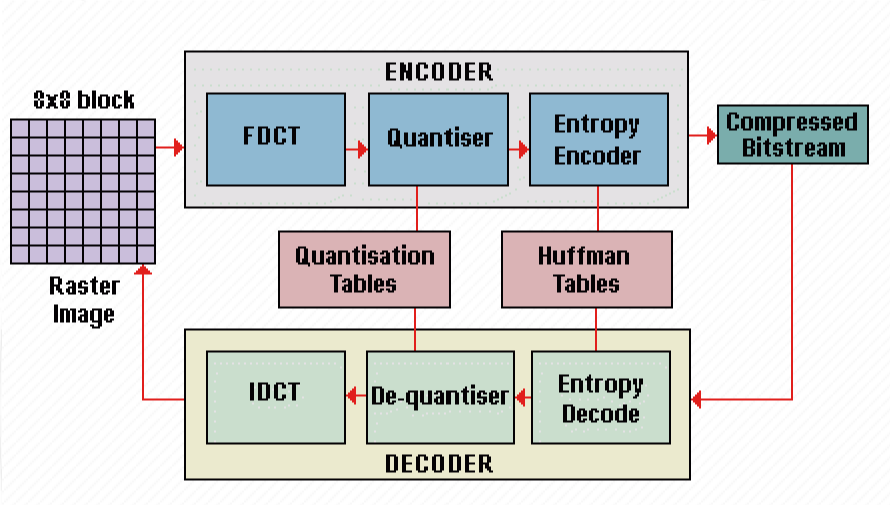


For JPG: discrete cosine transform



The DCT transforms an 8×8 block of input values to a linear combination of these 64 patterns. The patterns are referred to as the two-dimensional DCT basis functions, and the output values are referred to as transform coefficients.

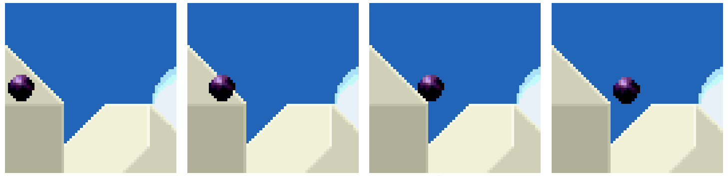
- **Compression method for video**
  - H.261, H.262, H.263, H.264, H.265, AV1, WMV



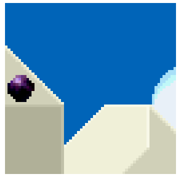
Example: Encoder decoder structure

Image & video Representation

- **Frame types of video**

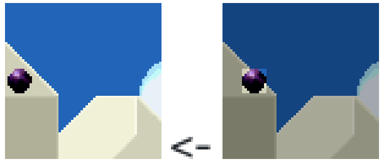
Input: 

- **I Frame (intra, keyframe)**



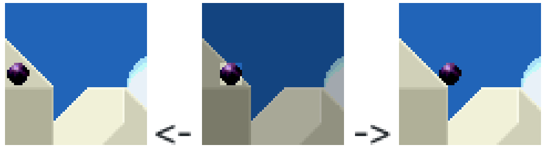
An I-frame (reference, keyframe, intra) is a self-contained frame. It doesn't rely on anything to be rendered, an I-frame looks similar to a static photo.

- **P Frame (predicted)**

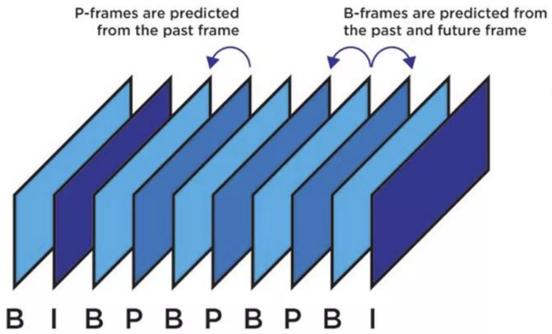


A P-frame takes advantage of the fact that almost always the current picture can be rendered using the previous frame.

- **B Frame (bi-predictive)**



B-frame refers the past and future frames to provide even a better compression



商汤 sensetime

清華大學 Tsinghua University

Outline

Part 1	Image and video representation
Part 2	Image processing
Part 3	Video processing

- **History of Digital Image Processing**

1960s: Improvements in computing technology and the onset of the **space race** led to a surge of work in digital image processing

- **1964**: Improve the quality of images of moon
- Such techniques were used in Apollo landings

Image Enhancement

A picture of the moon taken by the Ranger 7 probe minutes before landing

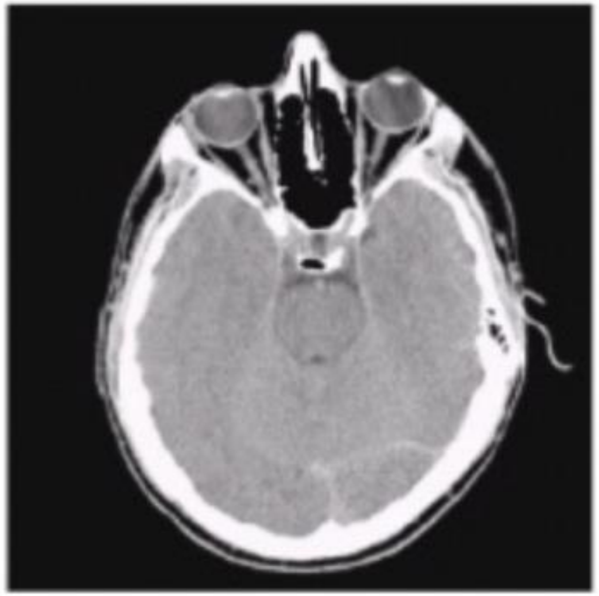
Image Processing

- **History of Digital Image Processing**

1970s: Digital Image processing begins to be used in medical applications

- **1979:** Sir Godfrey & Prof. Allan share the Nobel Prize in medicine for the tomography.

Image Restoration

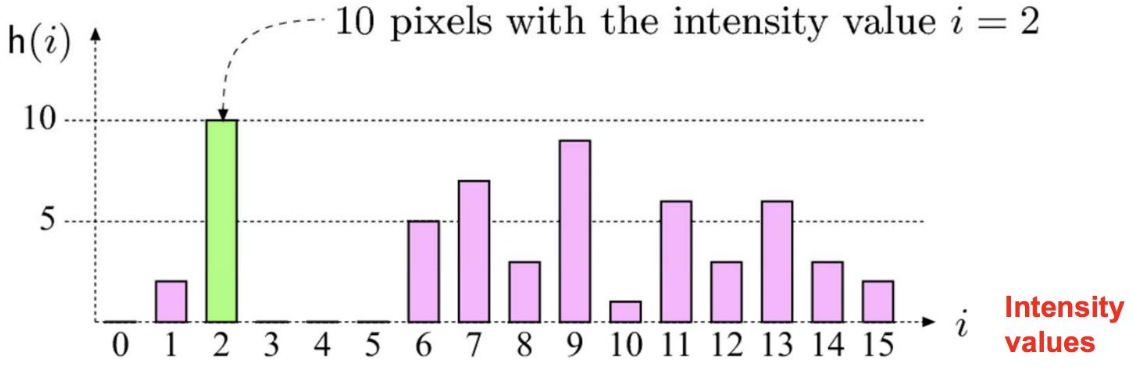


Typical head slice CAT image

- **Histograms**
  - Histograms plots how many times(frequency) each intensity value in image occurs
  - **Example:**
    - Image (left) has 256 distinct gray levels (8 bits)
    - Histogram (right) shows frequency (how many times) each gray level occurs

Image Processing

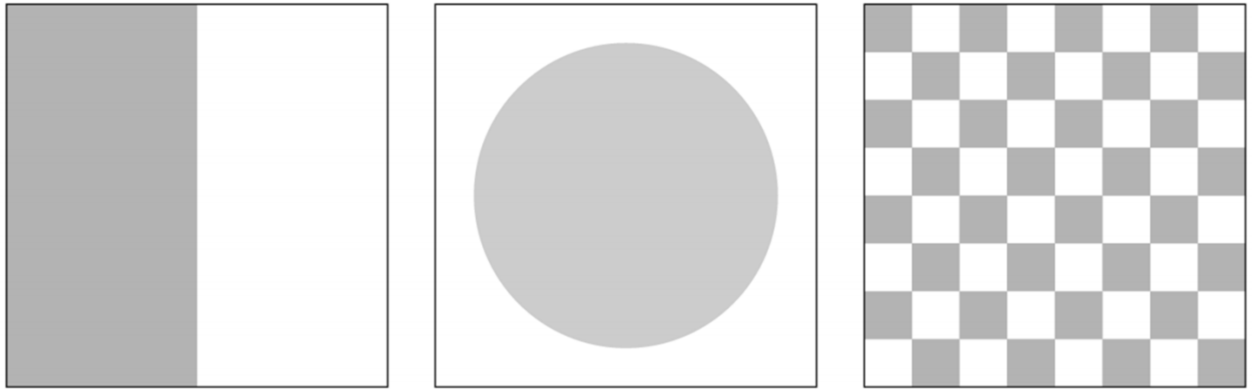
- **Histograms**



$h(i)$	0	2	10	0	0	0	5	7	3	9	1	6	3	6	3	2
i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

- Histograms: only statistical information
- No indication of location of pixels

- **Histograms**
  - Different images can have same histogram
  - 3 images below have same histogram



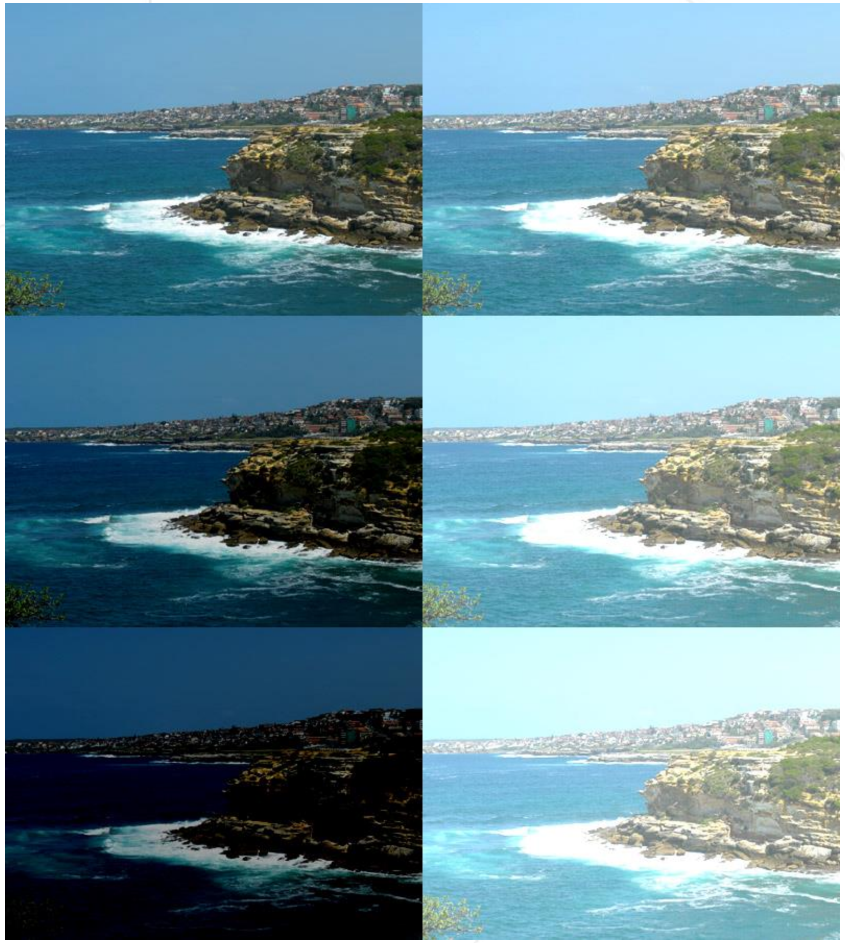
  - Half of pixels are gray, half are white
    - Same histogram = same statistics
    - Distribution of intensities could be different

- **Brightness**
  - Brightness of a grayscale image is the average intensity of all pixels in image

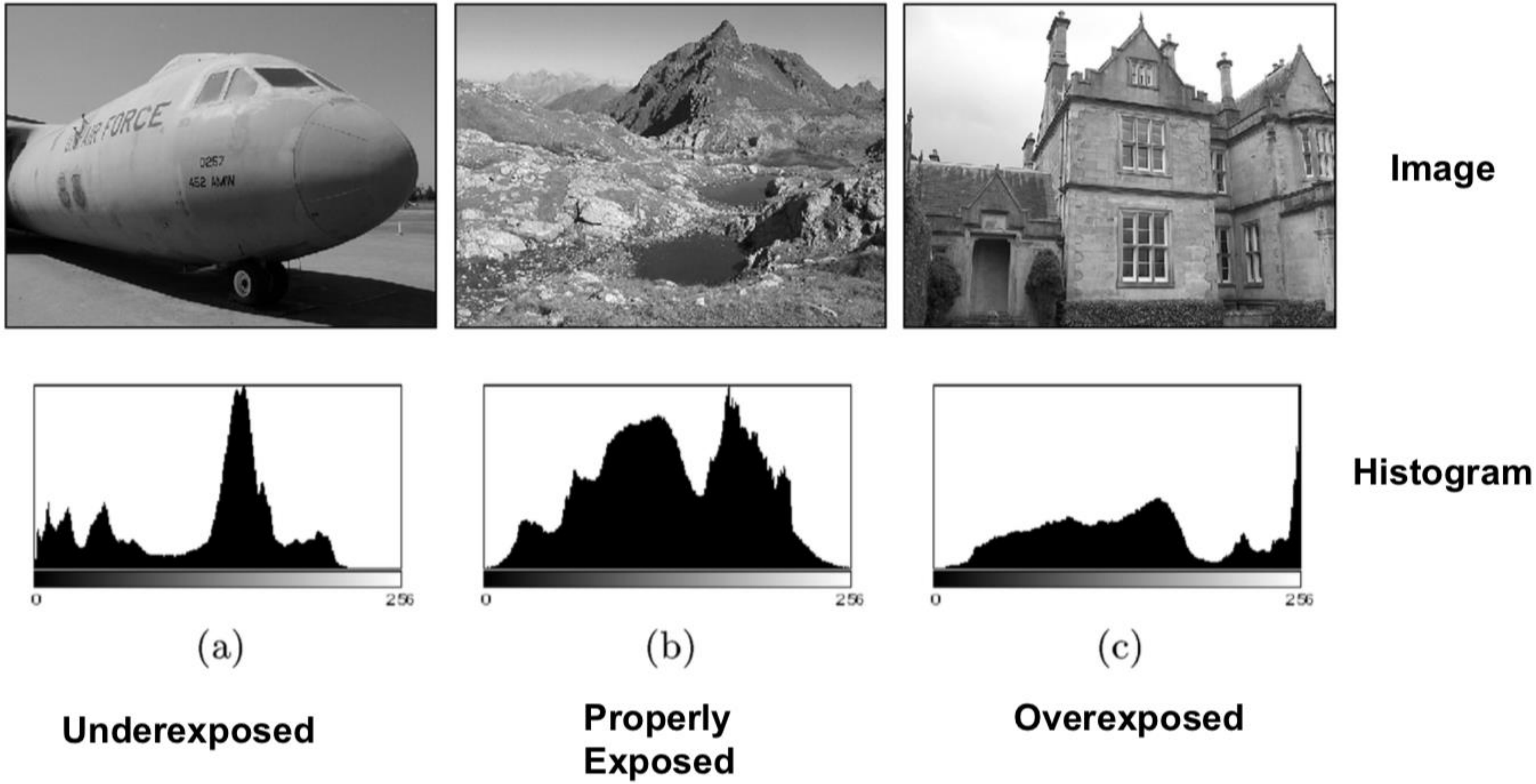
$$B(I) = \frac{1}{wh} \sum_{v=1}^{h} \sum_{u=1}^{w} I(u, v)$$

1. Sum up all pixel intensities

2. Divide by total number of pixels



- **Brightness and Histogram**



- **Image Contrast**
  - The contrast of a grayscale image indicates how easily objects in the image can be distinguished
    - **High contrast:** many distinct intensity values
    - **Low contrast:** image uses few intensity values
  - Many different equations for contrast exist

$$\text{Contrast} = \frac{\text{Change in Luminance}}{\text{Average Luminance}}$$

- **Contrast and Histogram**

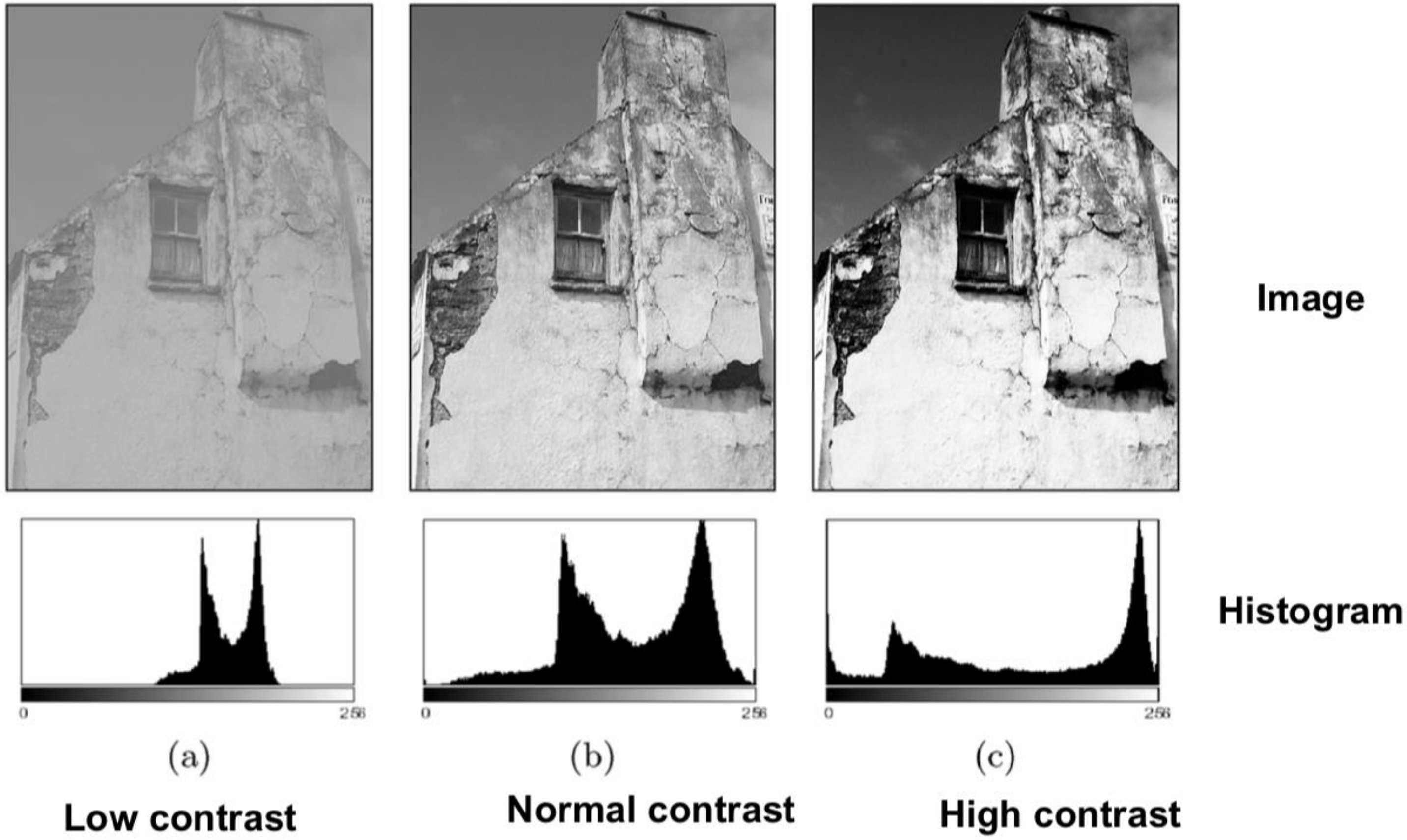
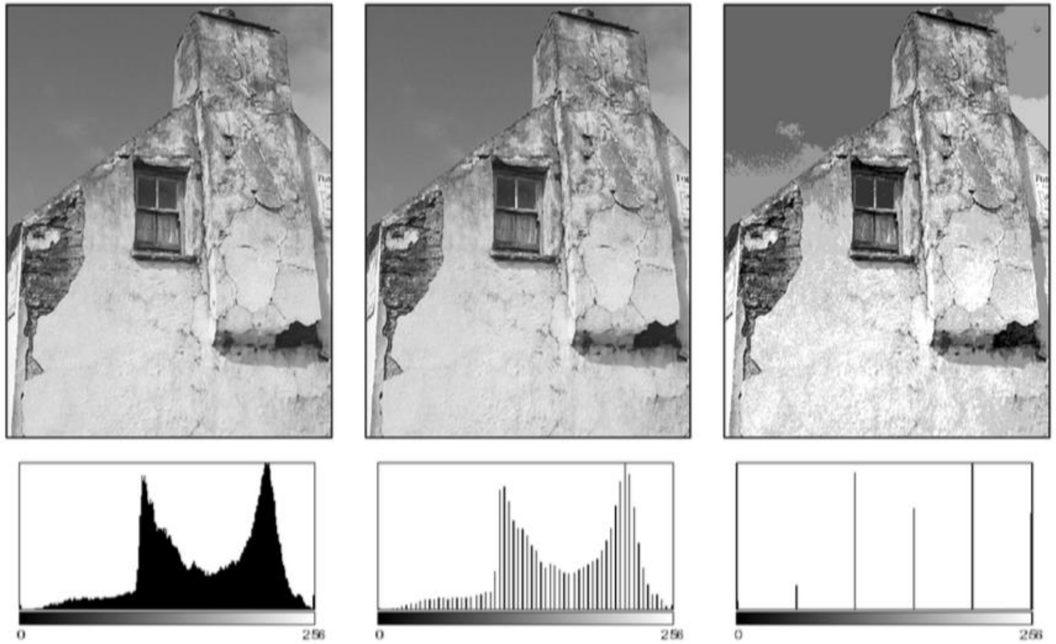


Image Processing

- ## Dynamic Range and Histogram

- **Dynamic Range:** Number of distinct pixels in image



(a) High Dynamic Range

(b) Low Dynamic Range (64 intensities)

(c) Extremely low Dynamic Range (6 intensity values)

Image Processing

- **Image Enhancement - intensity transformation**

  - **Image negatives**
  - Transform function T : g(x, y) = L - f(x, y),

    where L is the max intensity.

```python
import cv2
import numpy as np
# Load the image
img = cv2.imread('D:/downloads/forest.jpg')
# Check the datatype of the image
print(img.dtype)
# Subtract the img from max value(calculated from dtype)
img_neg = 255 - img
# Show the image
cv2.imshow('negative',img_neg)
cv2.waitKey(0)
```

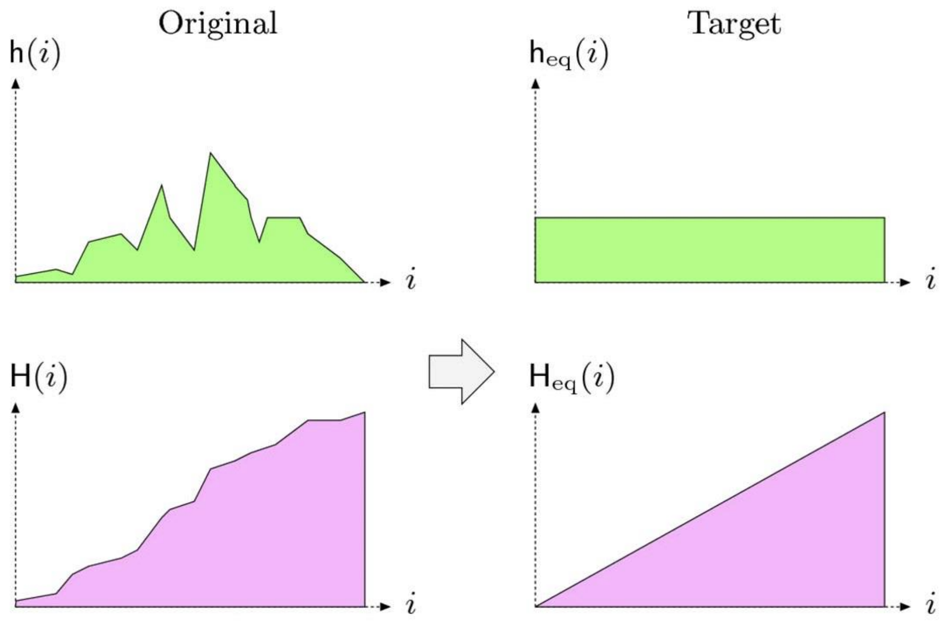
Original

Negative

- **Image Enhancement - Histogram equalization**
  - Apply a point operation that changes histogram of modified image into **uniform distribution**

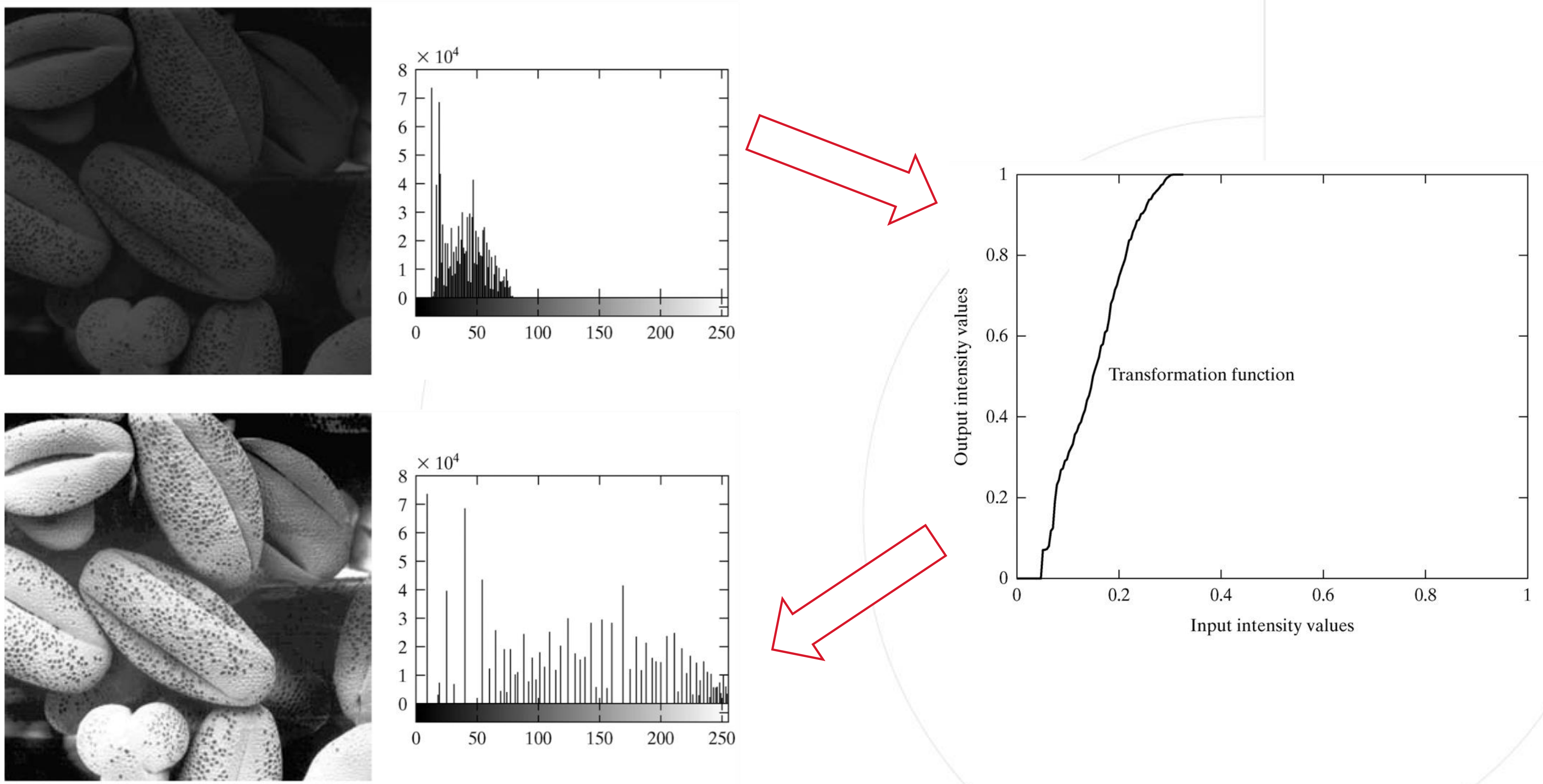
Histogram

Cumulative Histogram




```
img = cv2.imread('test.jpg',0)
equ = cv2.equalizeHist(img)
res = np.hstack((img,equ)) #stacking images side-by-side
cv2.imwrite('output.png',res)
```

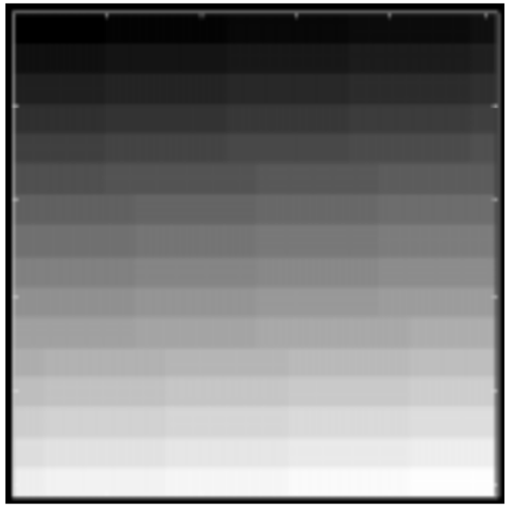
- **Image Enhancement - Histogram equalization**



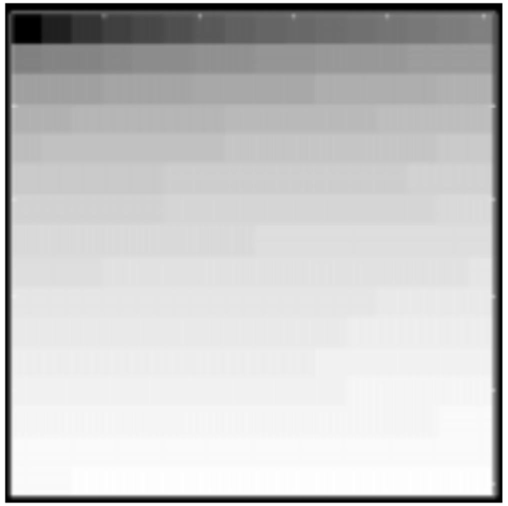
- **Image Enhancement - Compression of dynamic range**

  s = c log(1+|r|)

  - where c is a scaling constant, and the logarithm function performs the desired compression.



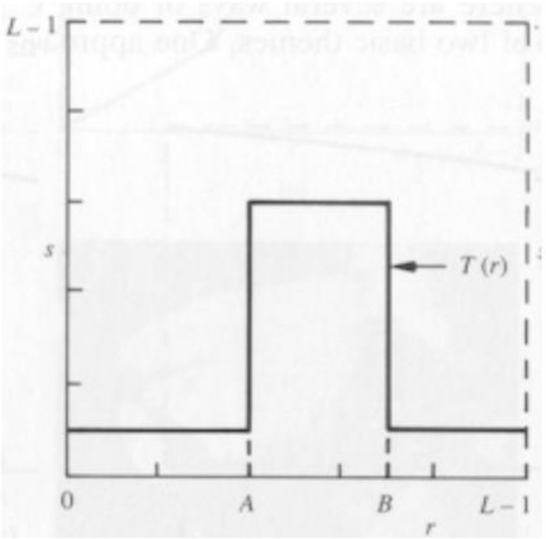
Original

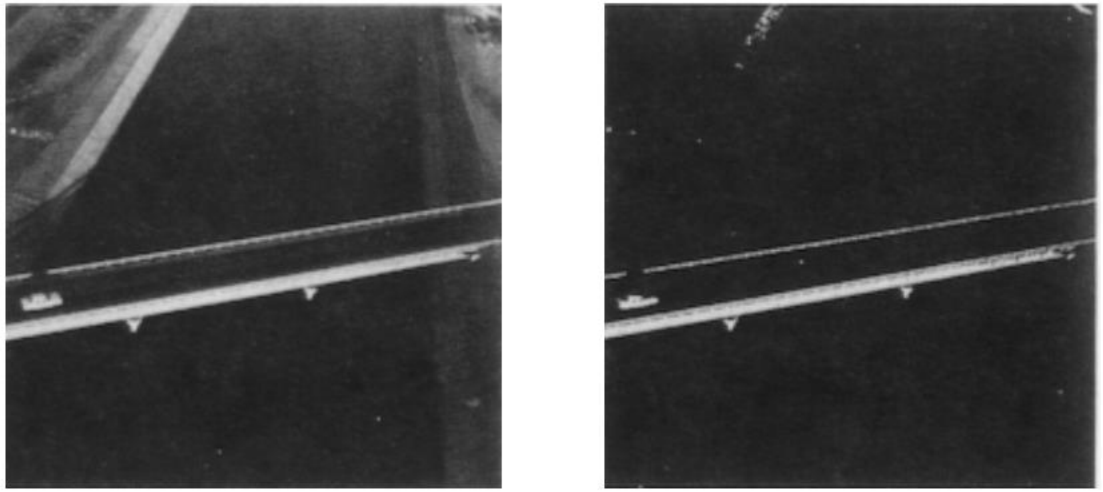


Processed output

Image Enhancement - Gray-level slicing

- A function that highlights a range [A,B] of transformation intensities while diminishing all others to a constant.





(b) (c)

Fig 1. (a) Transfer function, (b) Original image, (c) Processing output.

- **Image Enhancement - Spatial Filtering**

1. Low pass filtering

$$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

2. Median filtering

replacing each point with the median of neighboring points.

3. Sharpening Filter

$$\frac{1}{9}\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Original with (a) spike noise (b) white noise

Median filtering output

Low-pass filtering output

- **Image Enhancement in the frequency domain**

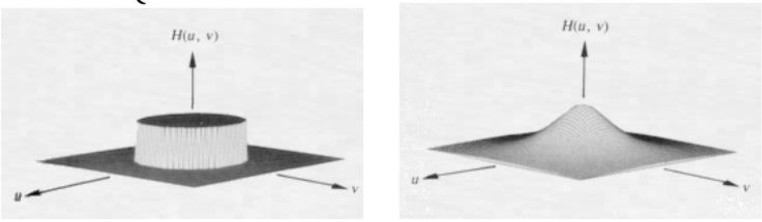
Spatial domain: $g(x,y)=f(x,y)*h(x,y)$

$$\Updownarrow$$

Frequency domain: $G(w_1,w_2)=F(w_1,w_2)H(w_1,w_2)$

- Lowpass filtering

$$H(u,v)=\begin{cases}1 & \text{if } D(u,v)\le D_o\\ 0 & \text{else}\end{cases}$$

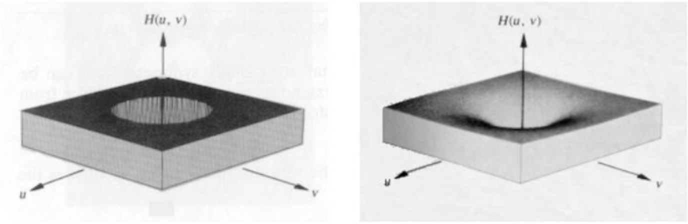


(a) (b)

Fig 6. (a) Ideal LPF; (b) Butterworth LPF.

- Highpass filtering

$$H(u,v)=\begin{cases}0 & \text{if } D(u,v)\le D_o\\ 1 & \text{else}\end{cases}$$



(a) (b)

Fig 7. (a) Ideal HPF; (b) Butterworth HPF.

- **Image Detection**

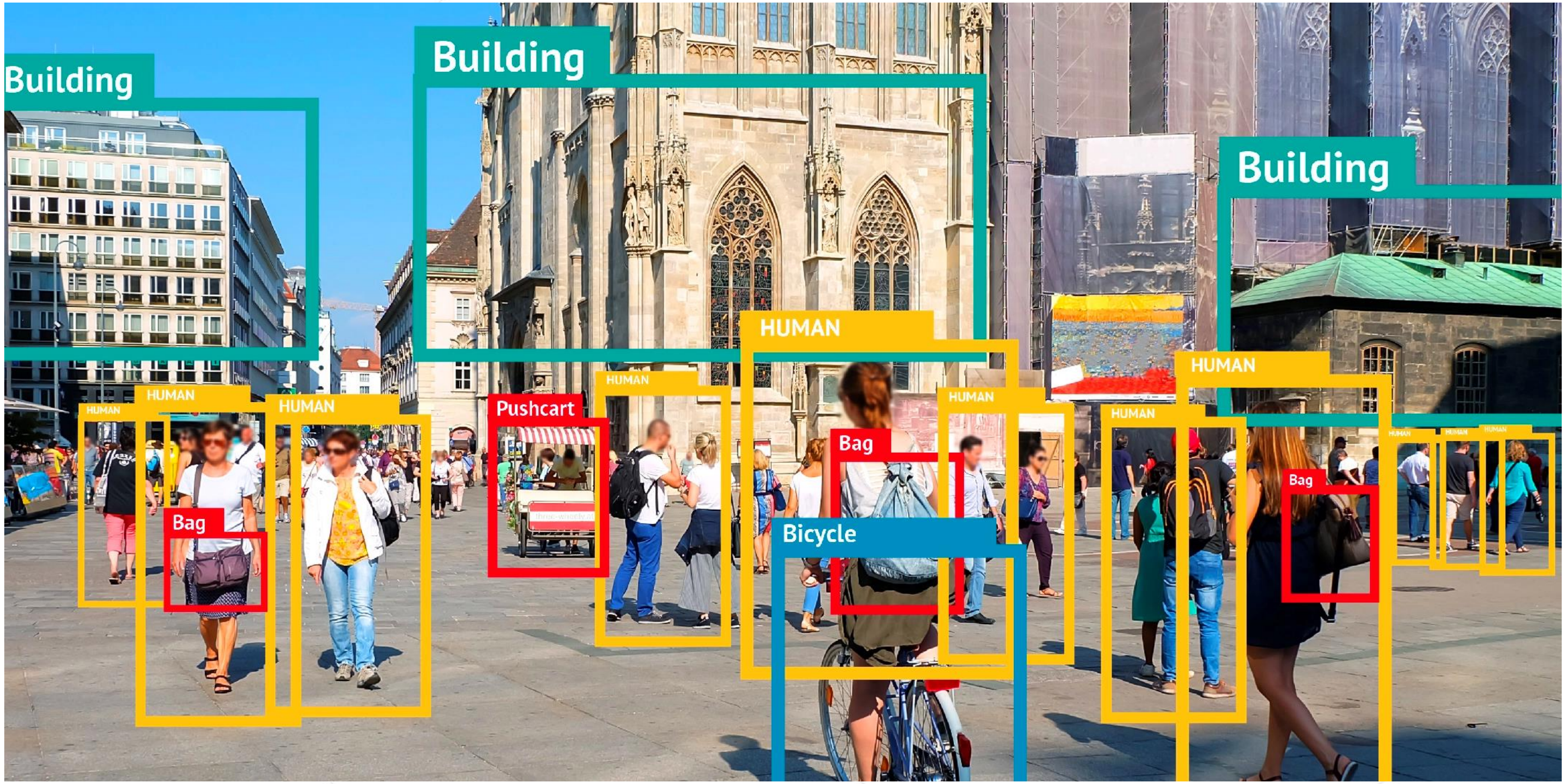
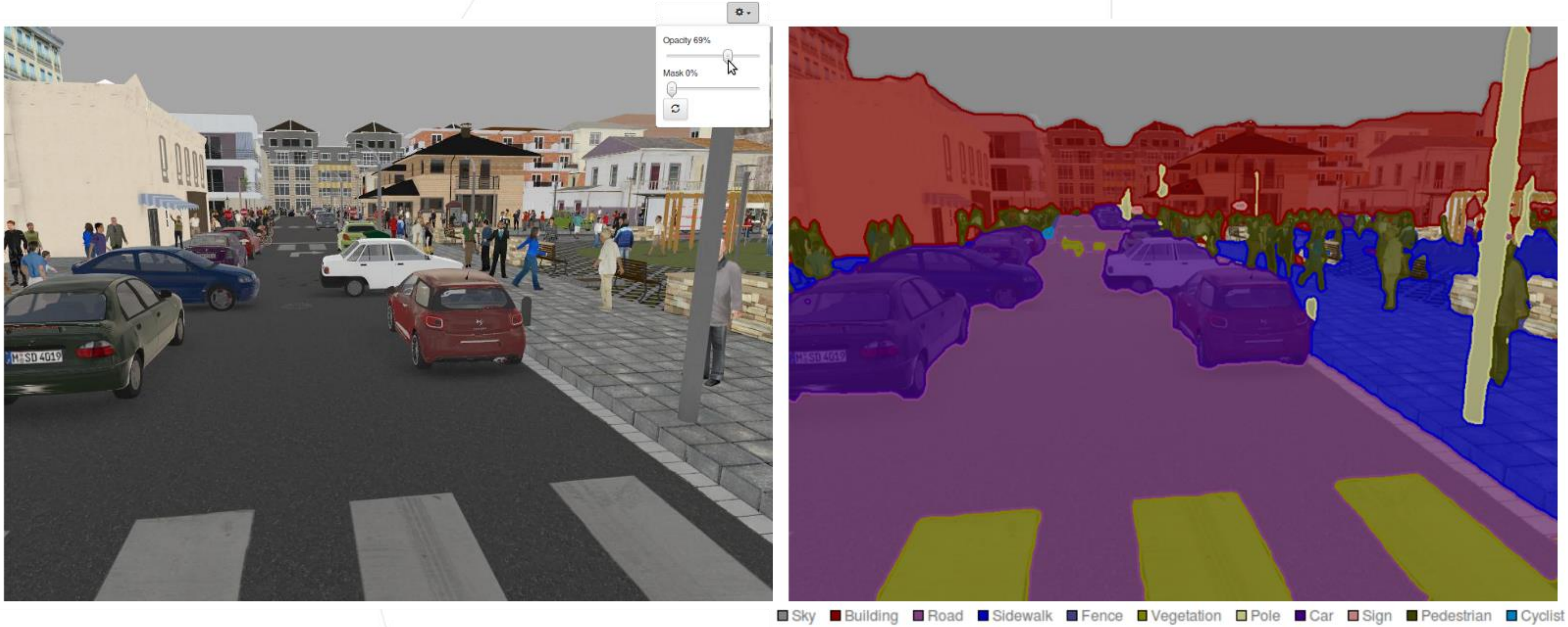


Image Processing

- **Image Segmentation**



Outline

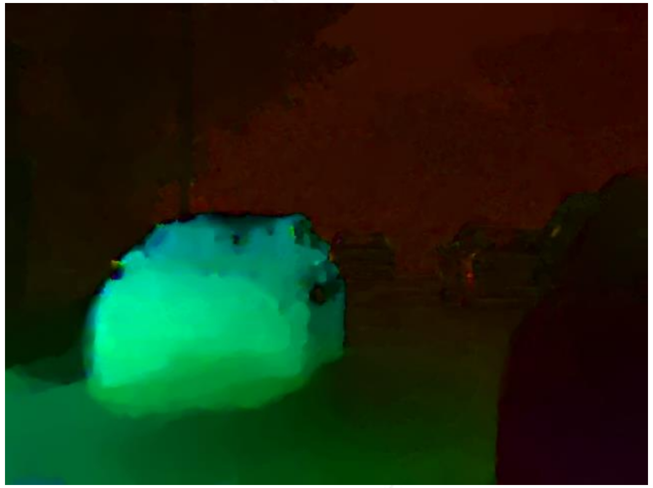
Part 1	Image and video representation
Part 2	Image processing
Part 3	Video processing

- **Optical Flow**

Optical flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene.

T = t

T = t + 1



Optical flow

- **Two types of Optical Flow**

Sparse

Dense

Gif by: https://nanonets.com/blog/optical-flow/

- **Optical Flow demo**

Gif by: https://gfycat.com/fr/wetcreepygecko

Tsinghua University sensetime

- **Optical Flow Estimation**

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t)$$

- Assuming the movement is small

$$I(x+\Delta x, y+\Delta y, t+\Delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \text{higher-order terms}$$

- By truncating the higher order terms, a linearization, it follows that

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \qquad \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0$$

- Thus

$$I_x V_x + I_y V_y = -I_t$$

- This is an equation in two unknowns and cannot be solved as such. This is known as the aperture problem of the optical flow algorithms
- To find the optical flow another set of equations is needed, given by some additional constraint. All optical flow methods introduce additional conditions for estimating

- **Lucas–Kanade method (Sparse, Local)**
  - It assumes that the flow is essentially constant in a local neighborhood of the pixel under consideration, and solves the basic optical flow equations for all the pixels in that neighborhood, by the least squares criterion

$$\begin{aligned} I_x(q_1)V_x + I_y(q_1)V_y &= -I_t(q_1) \\ I_x(q_2)V_x + I_y(q_2)V_y &= -I_t(q_2) \\ &\vdots \\ I_x(q_n)V_x + I_y(q_n)V_y &= -I_t(q_n) \end{aligned}$$

$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \vdots \\ I_x(q_n) & I_y(q_n) \end{bmatrix} \quad v = \begin{bmatrix} V_x \\ V_y \end{bmatrix} \quad b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(q_n) \end{bmatrix}$$

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \sum_i I_x(q_i)^2 & \sum_i I_x(q_i)I_y(q_i) \\ \sum_i I_y(q_i)I_x(q_i) & \sum_i I_y(q_i)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i I_x(q_i)I_t(q_i) \\ -\sum_i I_y(q_i)I_t(q_i) \end{bmatrix}$$

  - Since it is a purely local method, it cannot provide flow information in the interior of uniform regions of the image.

- **Horn–Schunck method (Dense, Global)**

The Horn-Schunck algorithm assumes smoothness in the flow over the whole image. Thus, it tries to minimize distortions in flow and prefers solutions which show more smoothness.

Let the image be p = (x,y) and the underlying flow field be w(p) = (u(p),v(p), 1), where u(p) and v(p) are the horizontal and vertical components of the flow field, respectively.

$$\mathrm{E(u,v)} = \int |\mathrm{I}_2(\mathbf{p}+\mathbf{w}) - \mathrm{I}_1(\mathbf{p})|^2 + \lambda(|\nabla u|^2 + |\nabla v|^2)d\boldsymbol{p}$$

To solve Eq. (1), we use an iterative flow framework. It assumes that an estimate of the flow field is w, and one needs to estimate the best increment dw(dw=(du,dv)), to update w. The objective function in Eq. (1) is then changed to

$$\mathrm{E(du,dv)} = \int |\mathrm{I}_2(\mathbf{p}+\mathbf{w}+\mathbf{dw}) - \mathrm{I}_1(\mathbf{p})|^2 + \lambda(|\nabla(u+du)|^2 + |\nabla(v+dv)|^2)d\boldsymbol{p}$$

The main idea to solve the above equation is to find dU,dV so that the gradient

$$\left[\frac{\partial E}{\partial dU}; \frac{\partial E}{\partial dV}\right] = 0$$

- **Horn–Schunck method**

We can derive

$$\frac{\partial E}{\partial dV} = 2((\boldsymbol{I}_y^2 + \lambda \boldsymbol{L})dV + \boldsymbol{I}_x\boldsymbol{I}_y dU + \boldsymbol{I}_y I_z + \lambda \boldsymbol{L}V)$$

where L is a Laplacian filter defined as

$$I_z(\boldsymbol{p}) = I_2(\boldsymbol{p}+\boldsymbol{w}) - I_1(\boldsymbol{p})$$

$$\boldsymbol{L} = \boldsymbol{D}_x^T\boldsymbol{D}_x + \boldsymbol{D}_y^T\boldsymbol{D}_y \qquad I_x(\boldsymbol{p}) = \frac{\partial}{\partial x} I_2(\boldsymbol{p}+\boldsymbol{w})$$

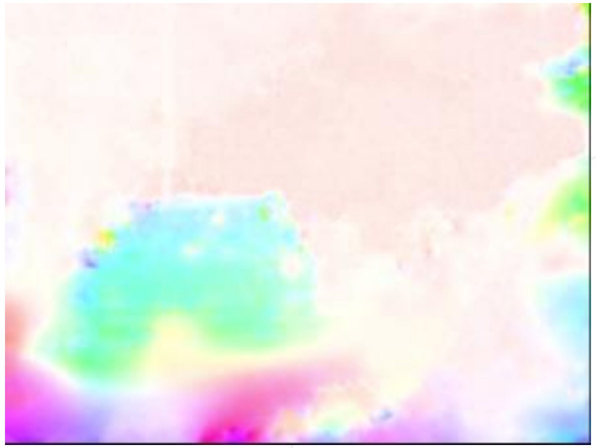
$$I_y(\boldsymbol{p}) = \frac{\partial}{\partial y} I_2(\boldsymbol{p}+\boldsymbol{w})$$

The term of dU in gradient is derived similarly. Therefore, solving the gradient equation can be performed in the following linear system

$$\begin{bmatrix} \boldsymbol{I}_x^2 + \lambda \boldsymbol{L} & \boldsymbol{I}_x\boldsymbol{I}_y \\ \boldsymbol{I}_x\boldsymbol{I}_y & \boldsymbol{I}_y^2 + \lambda \boldsymbol{L} \end{bmatrix} \begin{bmatrix} dU \\ dV \end{bmatrix} = -\begin{bmatrix} \boldsymbol{I}_x I_z + \lambda \boldsymbol{L}U \\ \boldsymbol{I}_y I_z + \lambda \boldsymbol{L}V \end{bmatrix}$$

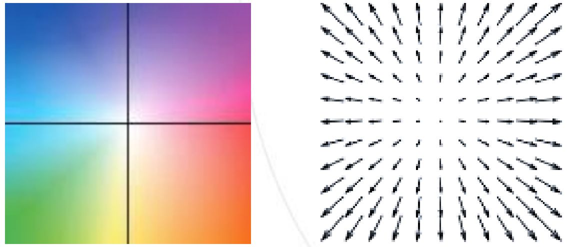
- **Horn–Schunck method**

Input two frames



Dense optical flow

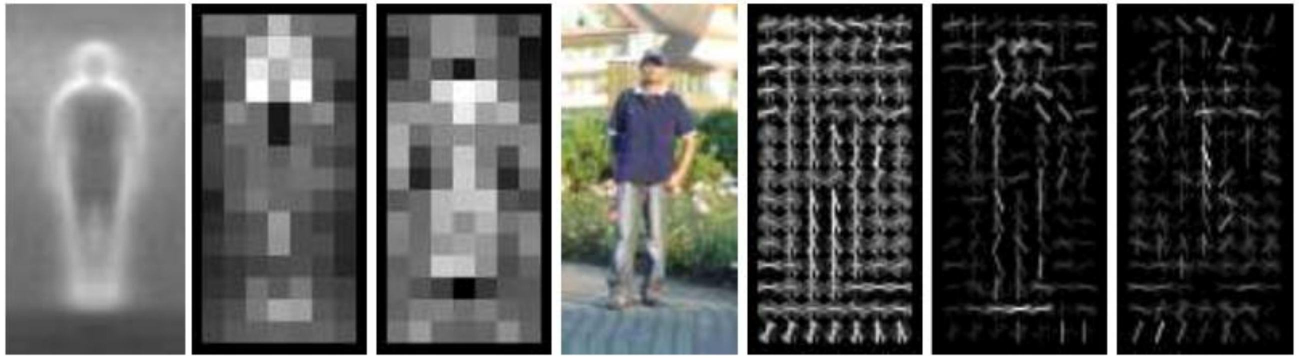
Wrapped frame



Flow Visualization

- **Video Descriptors**

HOG: Histogram of oriented spatial grad



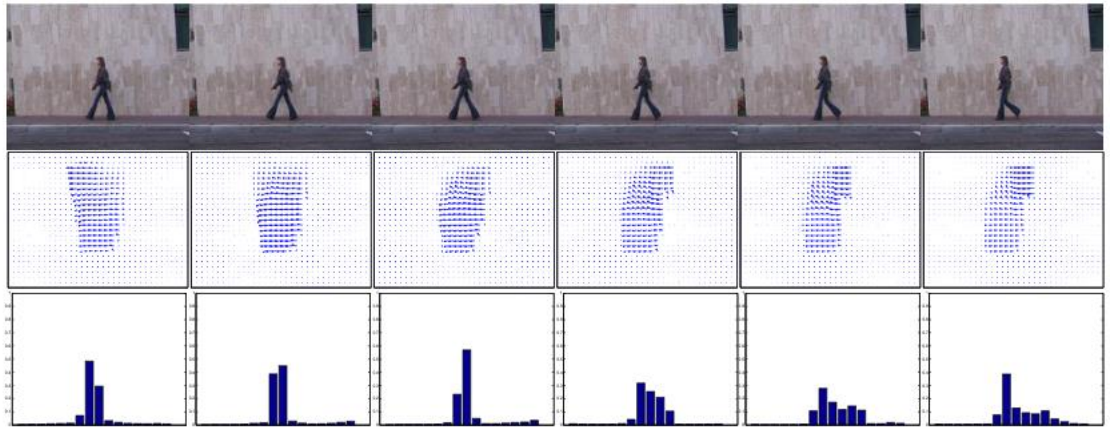
(a) (b) (c) (d) (e) (f) (g)

(a) The average gradient image over the training examples.
(b) Each ‘pixel’ shows the maximum positive SVM weight in the block centred on the pixel.
(c) Likewise for the negative SVM weights.
(d) A test image.
(e) It’s computed R-HOG descriptor.
(f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights.

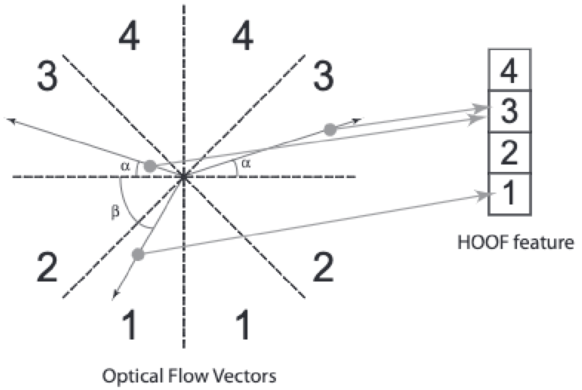
Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision & Pattern Recognition (CVPR ’05), Jun 2005, San Diego, United States. pp.886–893, 10.1109/CVPR.2005.177. inria-00548512

- **Video Descriptors**

HOF: Histogram of oriented optical flow



Optical flows and HOF feature trajectories



Histogram formation with four bins, B= 4

Chaudhry R, Ravichandran A, Hager G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 1932-1939.

Video Processing

- ## Video Descriptors

MBH: Motion Boundary Histograms

(a) (b) (c) (d) (e) (f) (g) (h)

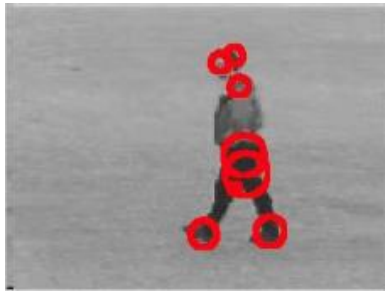
Illustration of the MBH descriptor.
(a,b) Reference images at time t and $t + 1$.
(c,d) Computed optical flow, and flow magnitude showing motion boundaries. (e,f) Gradient magnitude of flow field $\mathcal{I}^x, \mathcal{I}^y$ for image pair (a,b). (g,h) Average MBH descriptor over all training images for flow field $\mathcal{I}^x, \mathcal{I}^y$.

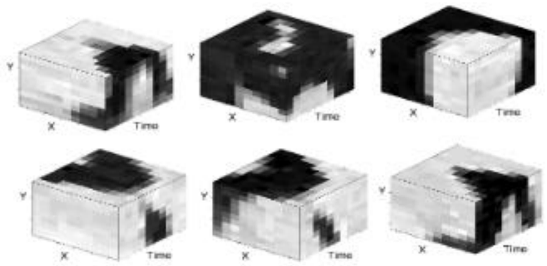
Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2006: 428-441.

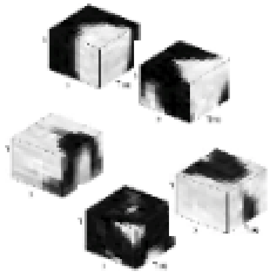
Video Processing

- **Traditional Action classification**
  - Bag of space-time features + SVM

Collection of space-time patches

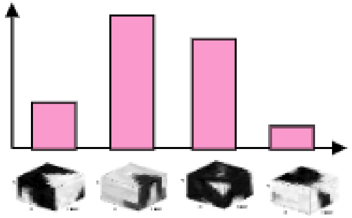






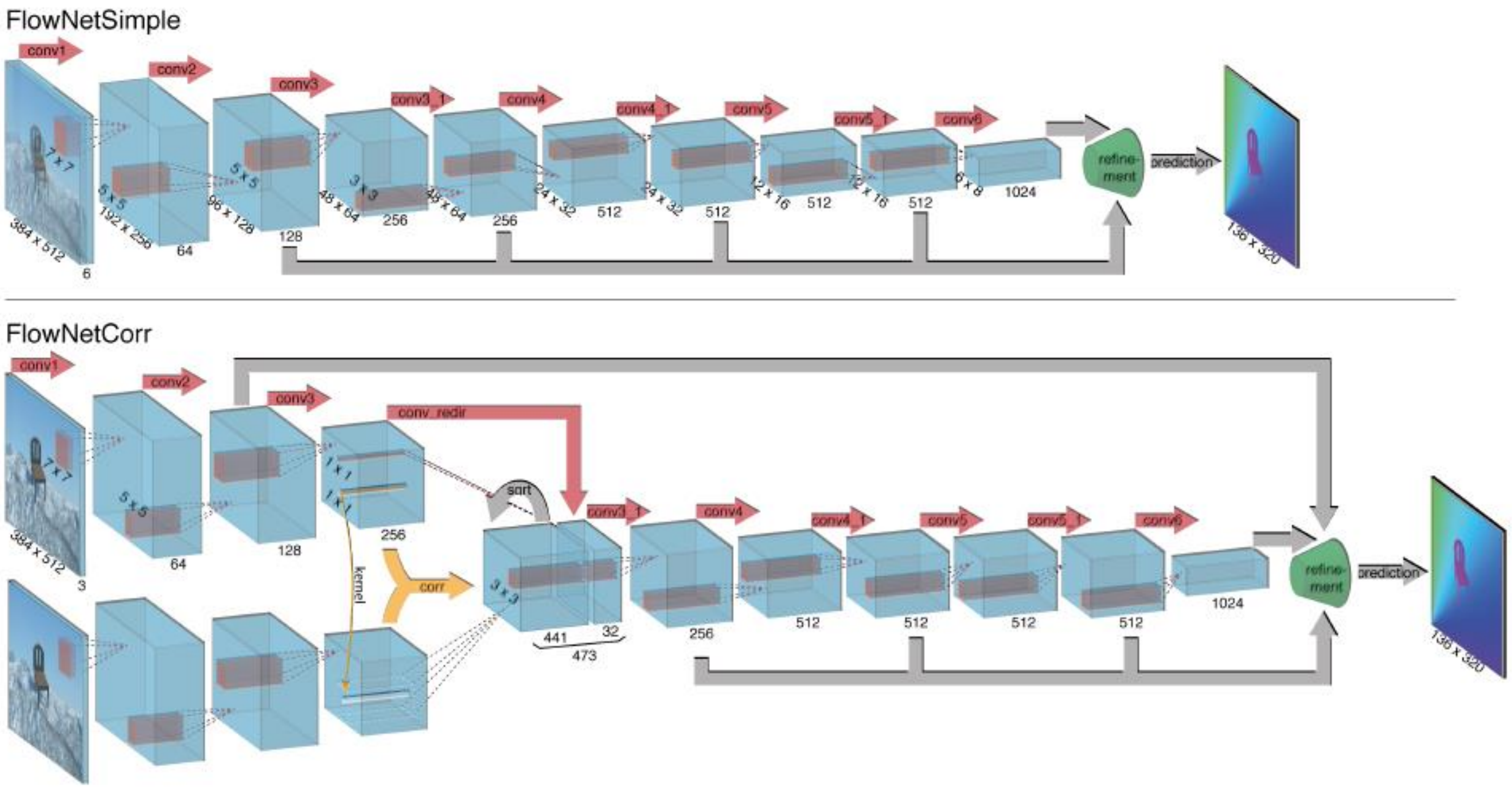
HOG & HOF patch& Others descriptors

Histogram of visual words



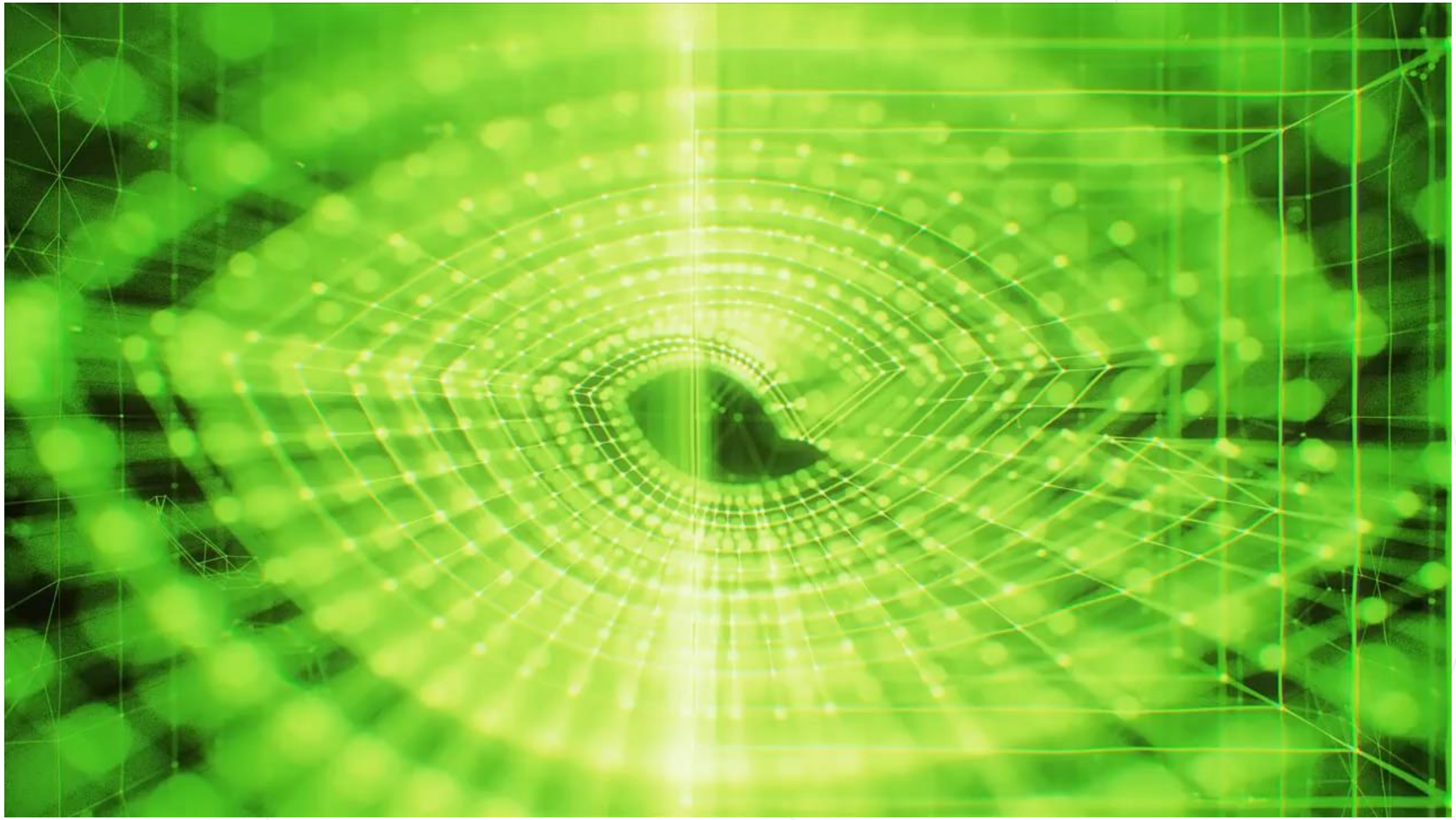
SVM Classifier

- **Optical flow estimation via deep networks**



Fischer et al.: FlowNet: Learning Optical Flow with Convolutional Networks, ICCV 2015.

- **Video Interpolation**



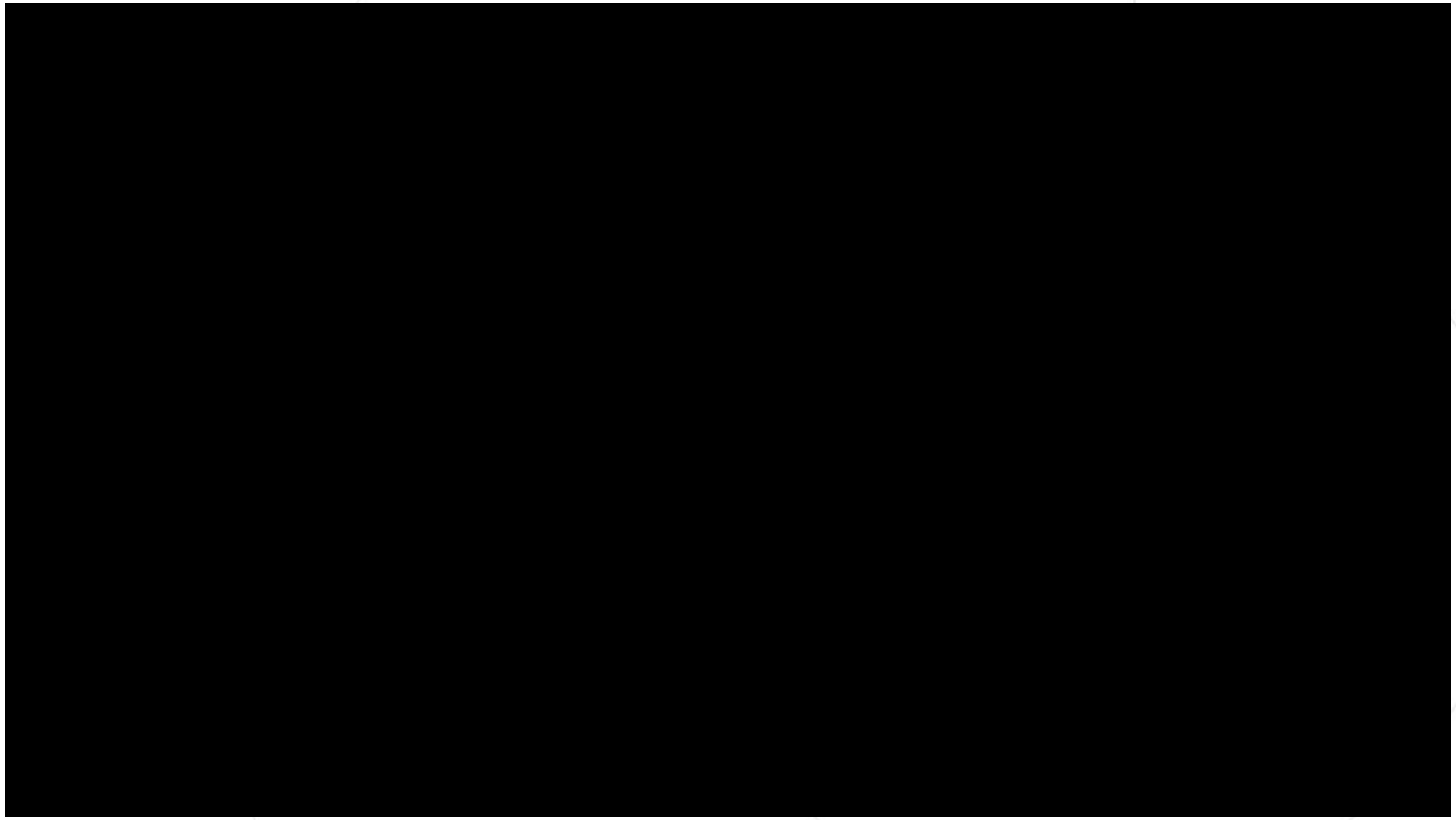
http://jianghz.me/projects/superslomo/

- **Video Stabilization**

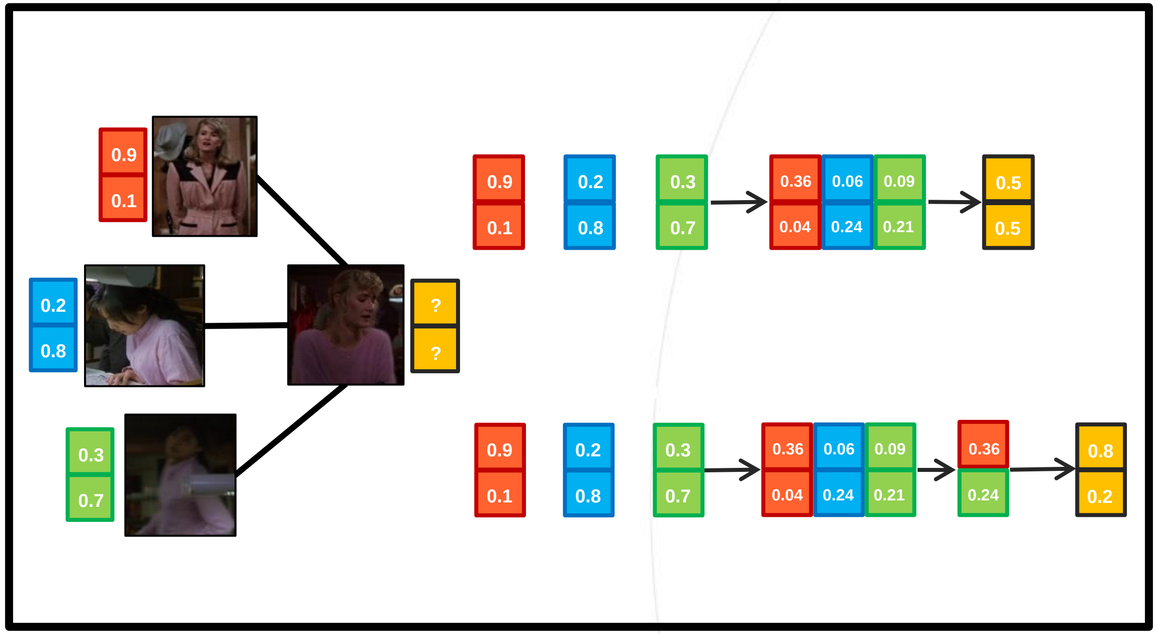
- **Video Denosing**

Video Processing

- **Video Super-Resolution**



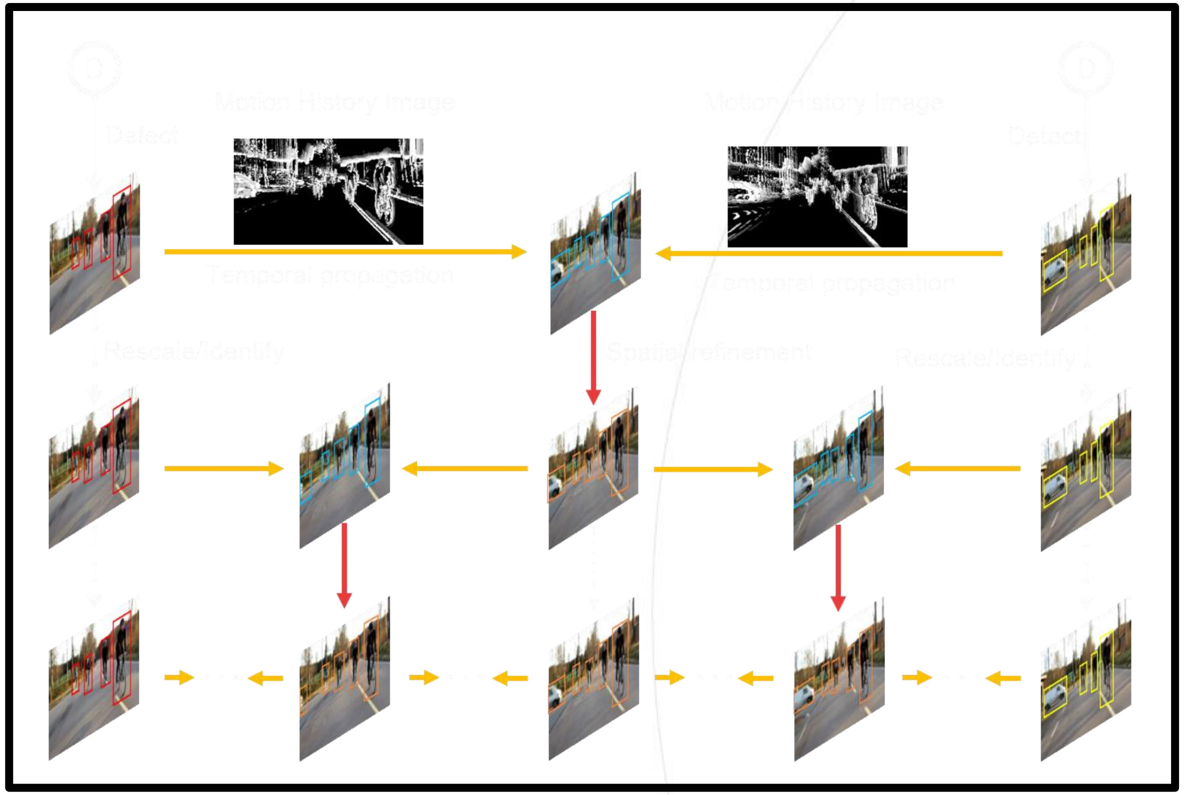
- **Video Understanding - Human**

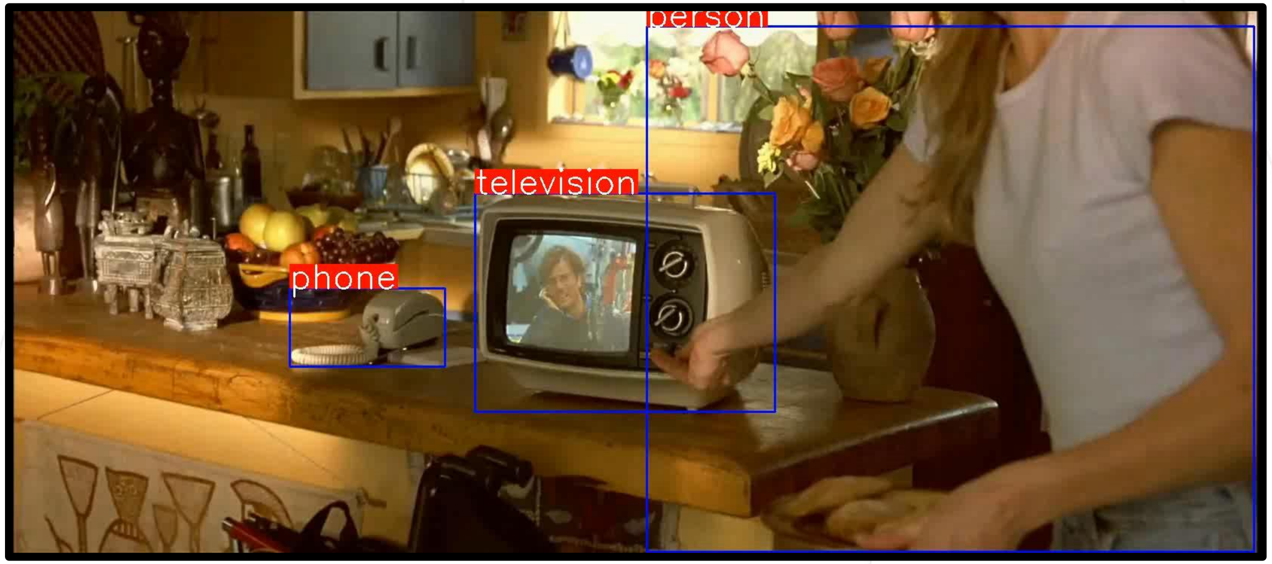


Q. Huang, W. Liu, D. Lin. Person Search in Videos with One Portrait Through Visual and Temporal Links. ECCV 2018

Video Processing

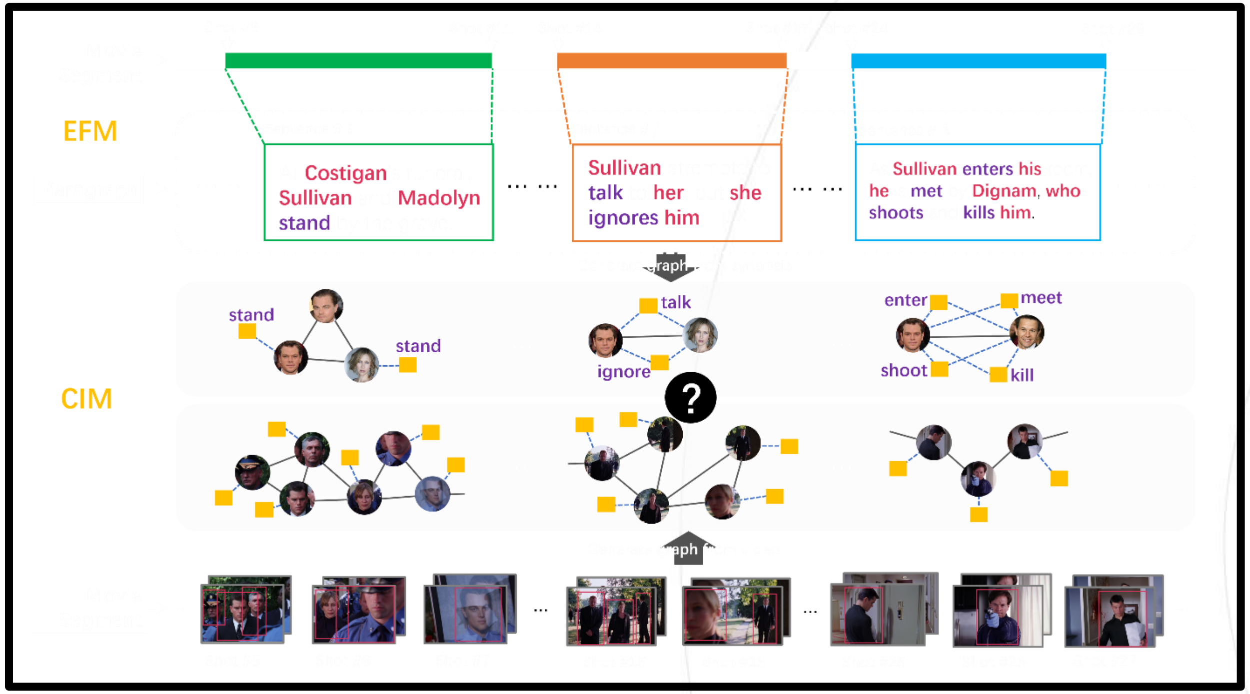
- ## Video Understanding - Object





K. Chen et al. Optimizing Video Object Detection via a Scale-Time Lattice. CVPR 2018.

- **Video Understanding - Context**



Y. Xiong, Q. Huang, L. Guo, H. Zhou, B. Zhou, D. Lin. A Graph-based Framework to Bridge Movies and Synopses. ICCV 2019.