# 第03讲 计算机视觉概述

信息学院（智能应用研究院）
欧 新 宇

# Welcome to Computer Vision

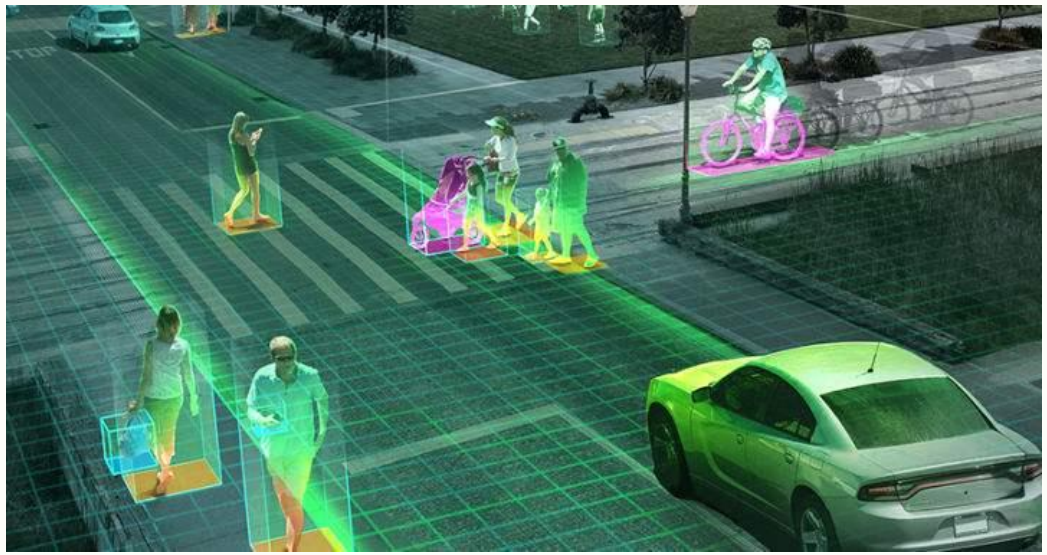
# 计算机视觉概述

# 计算机视觉概述

## What is Computer Vision?

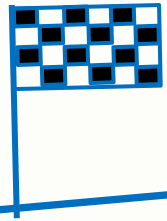
**计算机视觉(Computer Vision)**是一门研究如何使机器“**看**”的科学，即用摄影机和电脑代替人眼对目标进行**识别**、**跟踪**和**测量**等，并进一步做图形处理，使目标成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科，计算机视觉试图建立能够从图像或者多维数据中获取‘信息’的人工智能系统。这里所指的**信息**指Shannon定义的可以用来帮助做“决定”的信息。因为感知可以看作是从感官信号中提取信息，所以计算机视觉也可以看作是**研究如何使人工系统从图像或多维数据中“感知”的科学**。

有不少学科的研究目标与计算机视觉相近或与此有关。这些学科中包括图像处理、模式识别或图像识别、景物分析、图象理解等。具体到本课程，我们主要研究的是基于**深度学习**的图像分类、目标检测、图像分割、行人检测、图像搜索以及目标跟踪等。



# 计算机视觉概述

- 计算机视觉概述
- 计算机视觉简史
- 基于深度学习的视频内容理解
- 面向海量视频的视觉计算与识别



Part 01

# 计算机视觉概述

/ 计算机视觉在学科中的位置

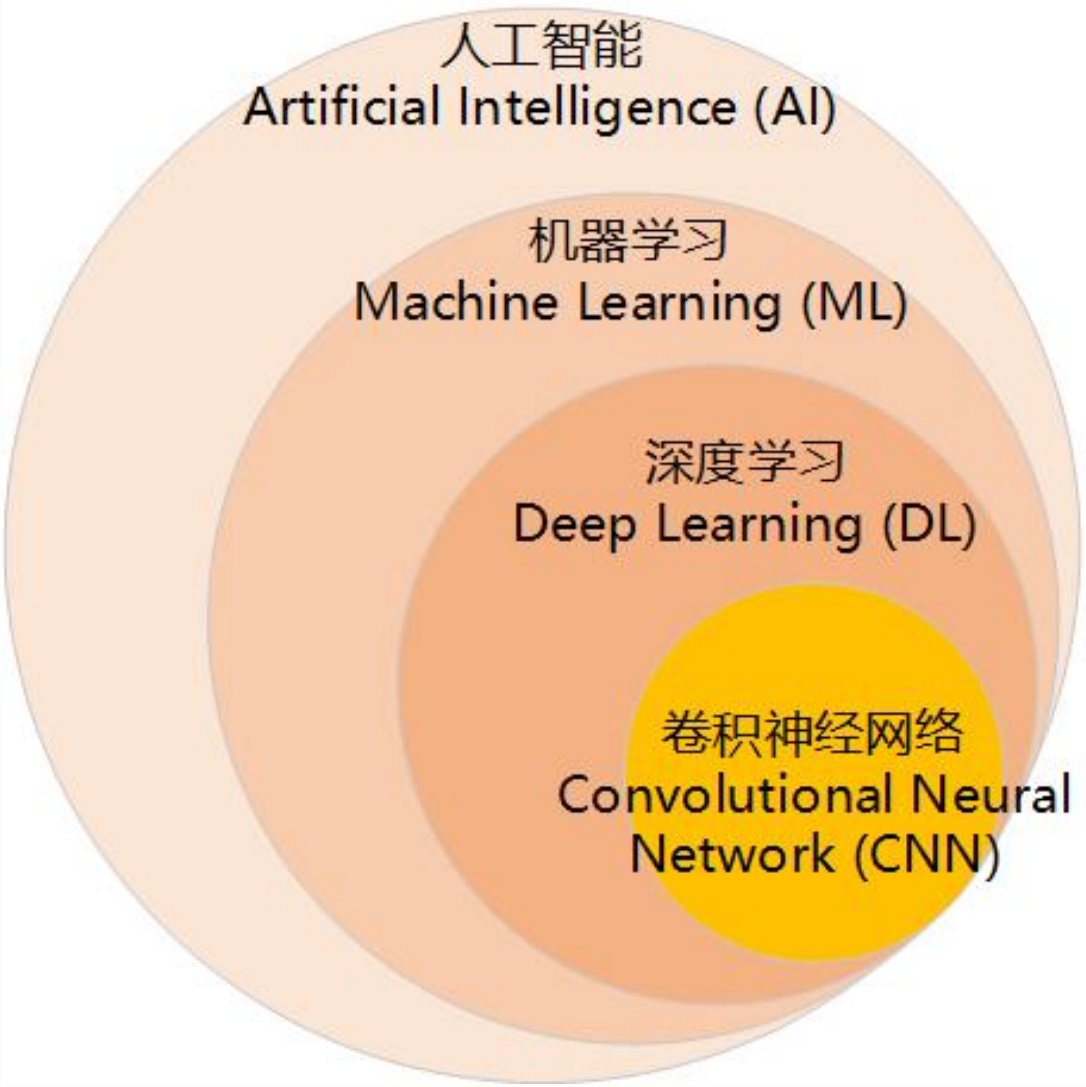
/ 视觉识别

/ 基于深度学习的视觉识别

/ 视觉领域的最高荣耀——图灵奖

# 1.1 计算机视觉在学科中的位置

# 1.1 计算机视觉在学科中的位置




**计算机视觉**
**Computer Vision (CV)**

- 目标检测 (Object Detection)
- 目标分类 (Object Classification)
- 场景理解 (Scene Understanding)
- 语义/实例分割 (Semantic/Instance Segmentation)
- 三维重建 (3D Reconstruction)
- 对象跟踪 (Object Tracking)
- 行人姿态估计 (Human Pose Estimation)
- 行为识别 (Activity Recognition)
- 视觉内容问答 (VQA)
- ...

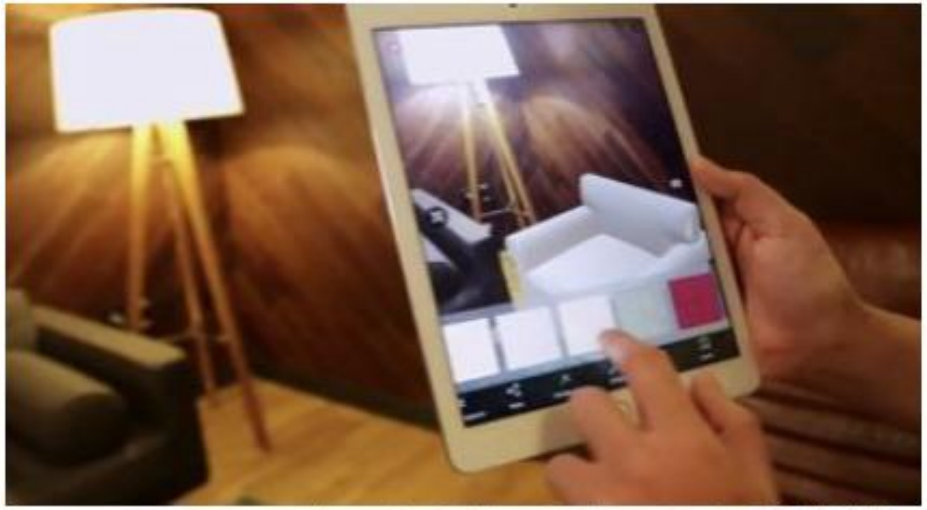
# 1.2 识别任务

## 视觉识别是计算机视觉的一个基本和普遍问题，源于认知科学







# 1.2 识别任务

## 视觉识别是计算机视觉的一个基本和普遍问题，源于认知科学

与**图像分类**(image classification)相关的视觉识别(visual recognition)问题有很多，如目标检测(object detection)、图像字幕(image captioning)、语义分割(semantic segmentation)、视觉问答(visual question answering)、视觉指令导航(visual instruction navigation)、场景图生成(scene graph generation)等。

# 1.2 识别任务

## 视觉识别是计算机视觉的一个基本和普遍问题，源于认知科学

### 对象检测

汽车



### 行为识别

骑自行车




### 视觉关系检测

<男孩-拿着-锤子>



有什么 → 做什么 → 对象群体活动

# 1.3 基于深度学习的视觉识别

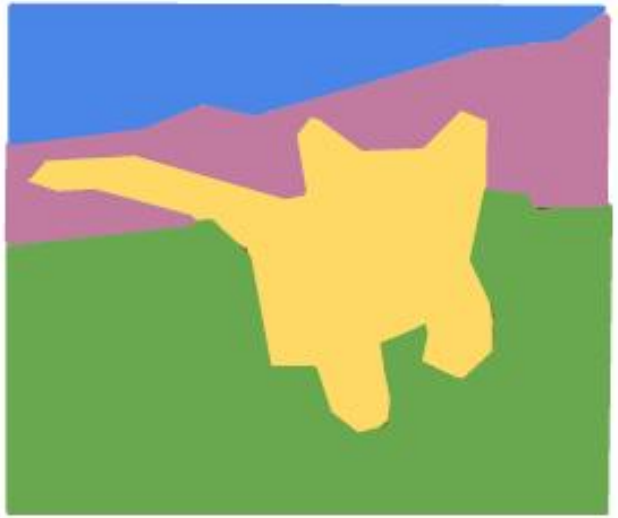
## 计算机视觉四大任务

**分　类**
**Classification**

CAT

无空间概念

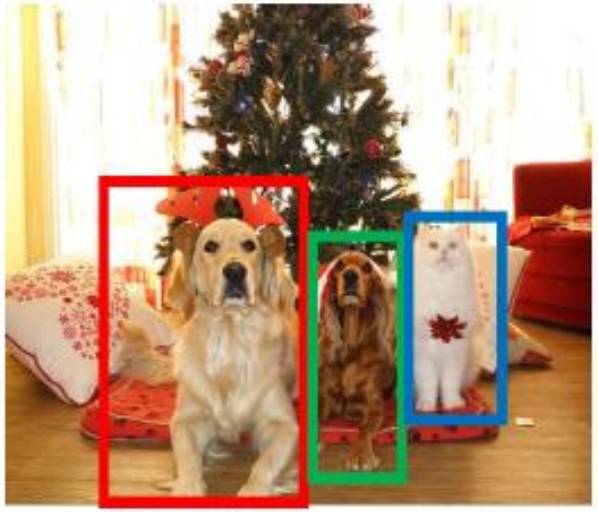
**语义分割**
**Semantic Segmentation**



GRASS, CAT, TREE, SKY

没有对象，只有像素

**目标检测**
**Object Detection**



DOG, DOG, CAT

**实例分割**
**Instance Segmentation**

DOG, DOG, CAT

多个目标对象

# 1.3 基于深度学习的视觉识别

## 一类神经网络已经成为视觉识别的重要工具

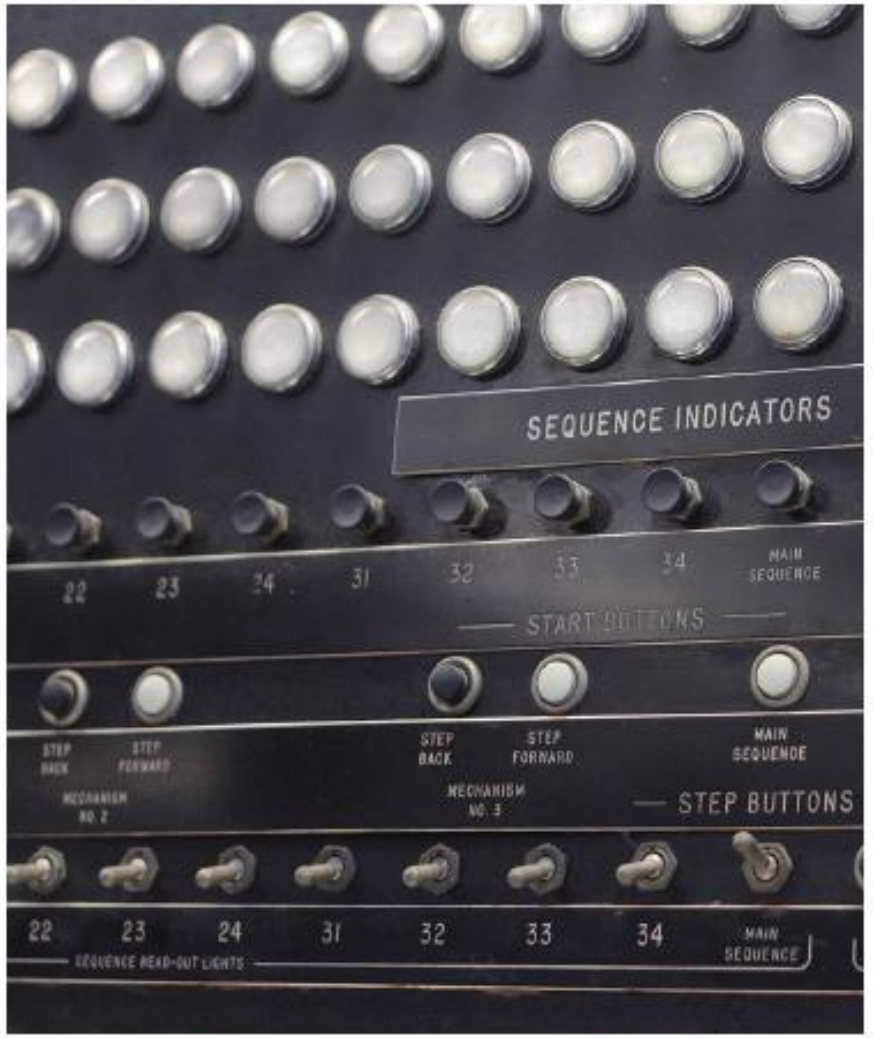
**核心思想可以追溯到几十年前！**

**Mark I感知机**是感知机算法的第一个实现。

这台机器与一台20×20像素的硫化镉相机相连，
可共同生成400像素的图像。

该系统可以用于识别信件中的英文字母。

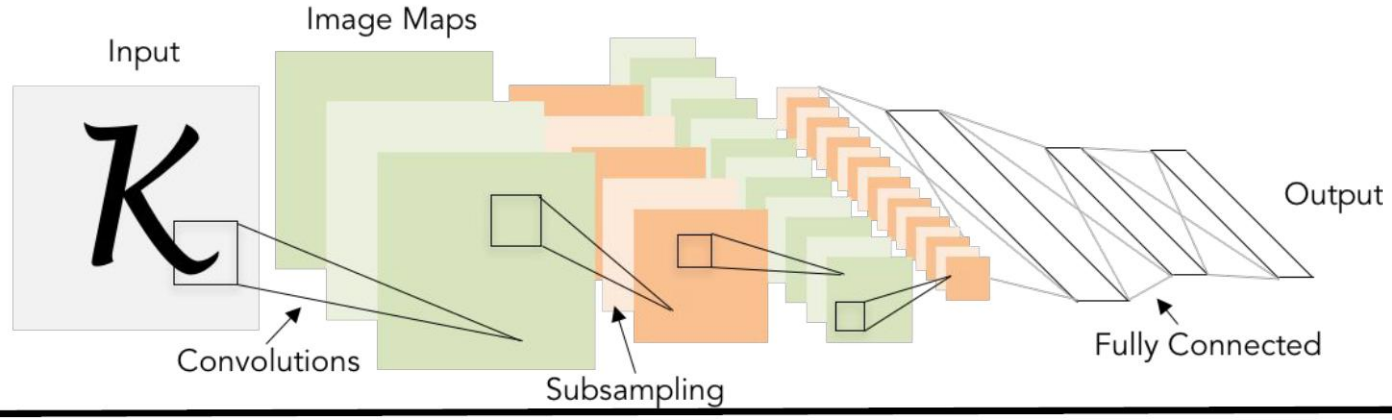
Frank Rosenblatt, ~1957: Perceptron



# 1.3 基于深度学习的视觉识别

## 一类神经网络已经成为视觉识别的重要工具

1998 LeCun et al.




晶体管数量


$10^6$

用于训练的像素数量

$10^7$ NIST

2012 Krizhevsky et al.

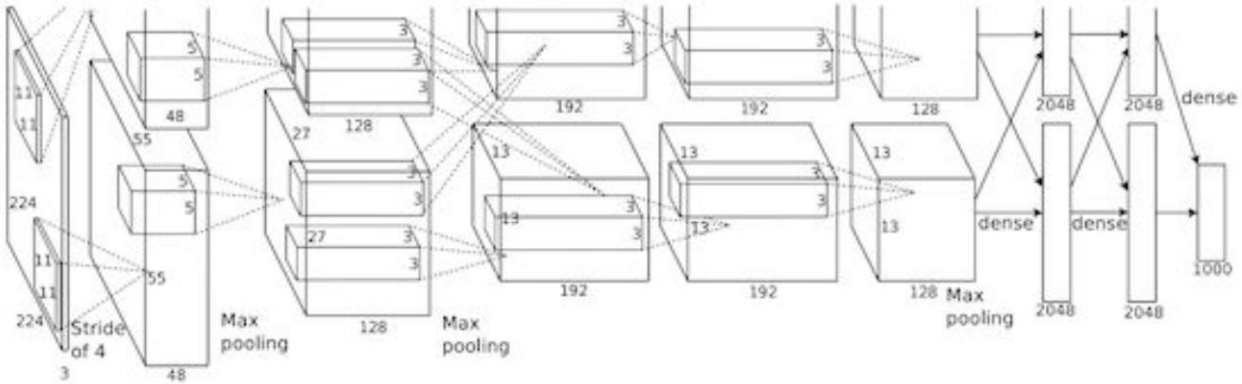



Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

晶体管数量

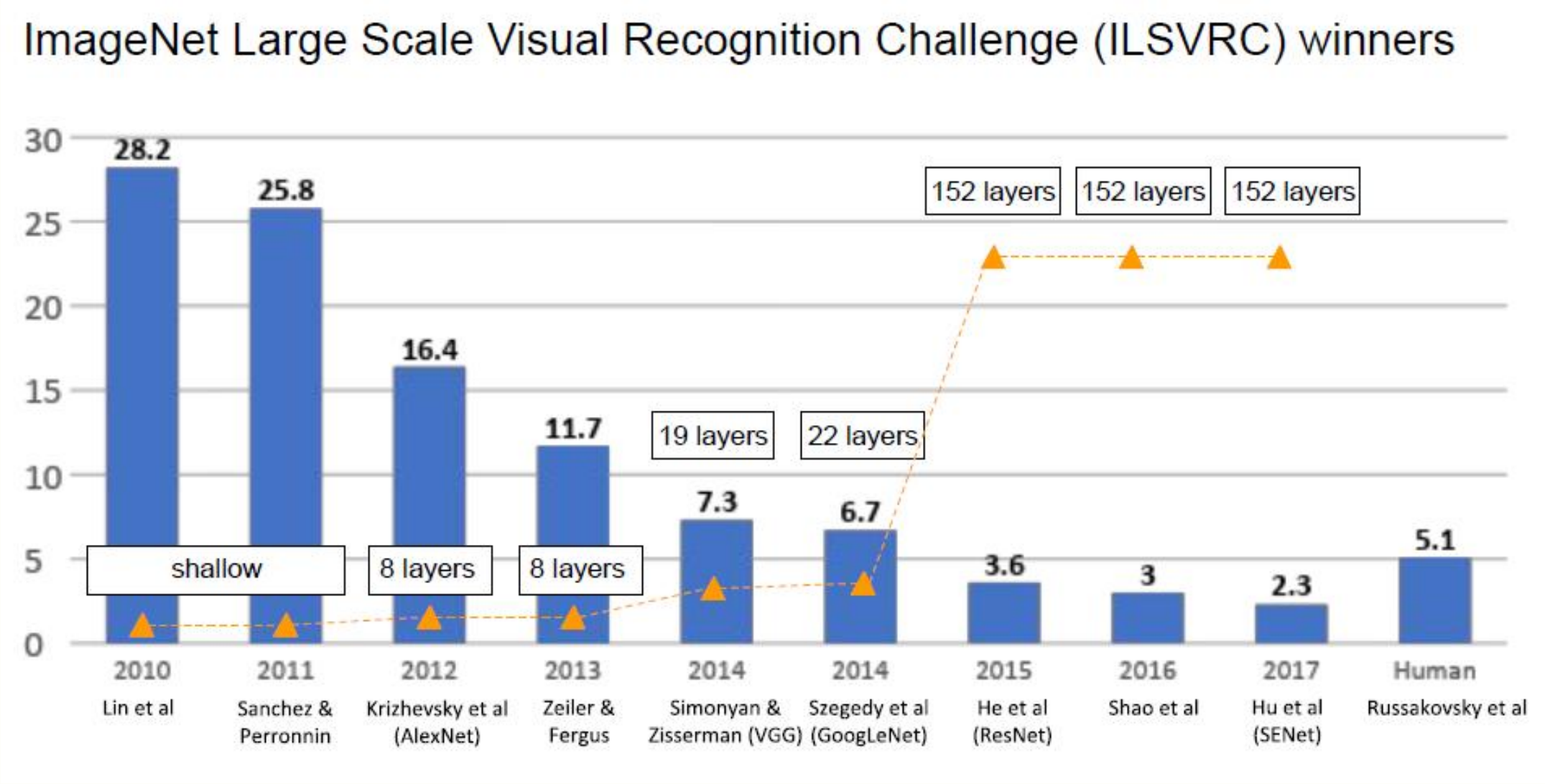
$10^9$

用于训练的像素数量

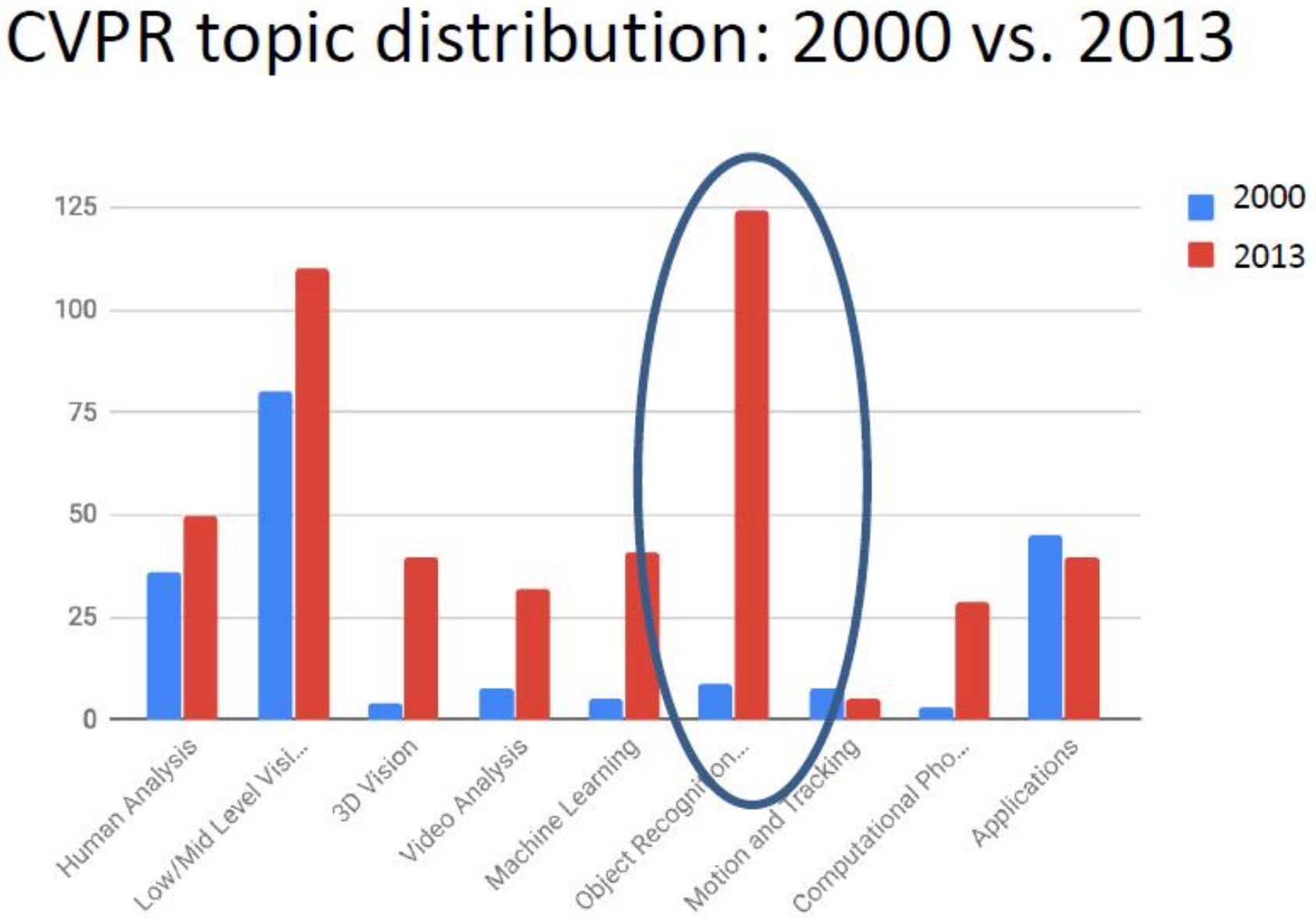
$10^{14}$ IMAGENET

# 1.3 基于深度学习的视觉识别

## CNN是具有许多“层”的分层计算系统，它受大脑的启发



# 1.3 基于深度学习的视觉识别



# 1.4 视觉领域的最高荣耀

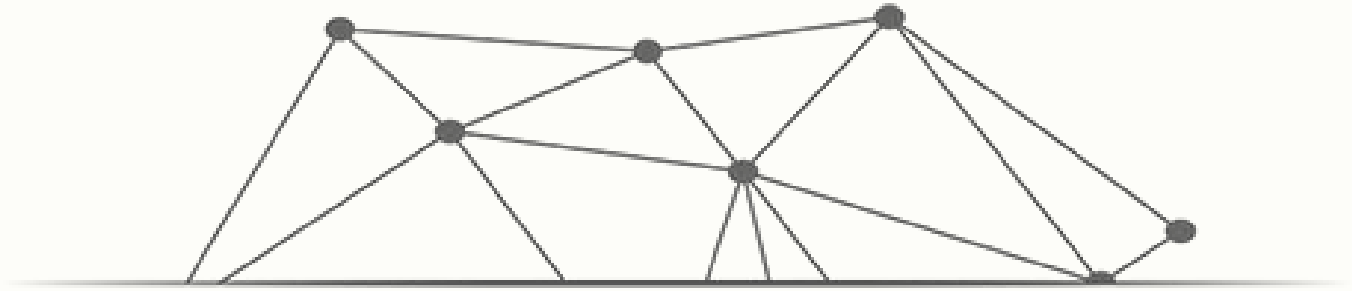
## 2018 Turing Award for deep learning

most prestigious technical award, is given for major contributions of lasting importance to computing.

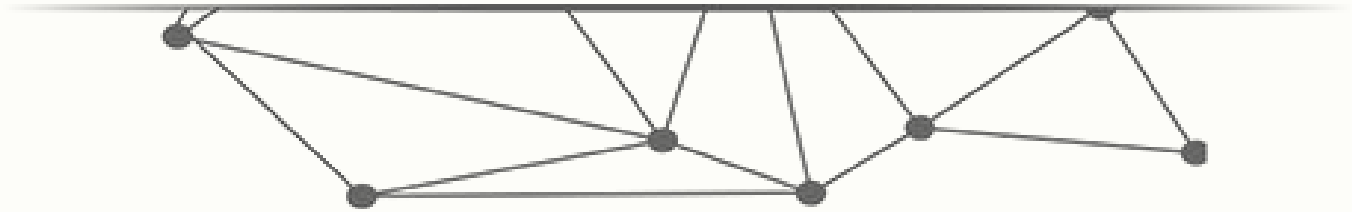

This image is CC0 public domain


This image is CC0 public domain


This image is CC0 public domain



# 课堂互动 13.1.1



Part
02

# 计算机视觉简史

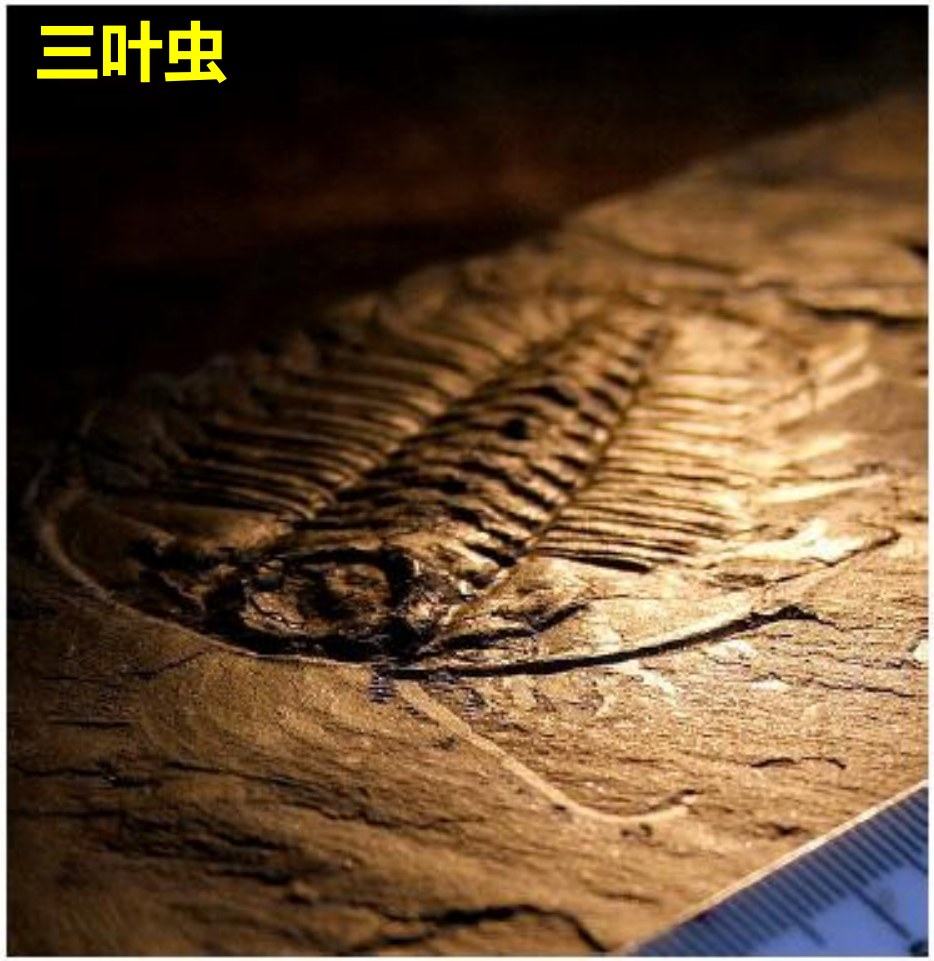
/ 视觉进化大爆炸

/ 计算机视觉的发展

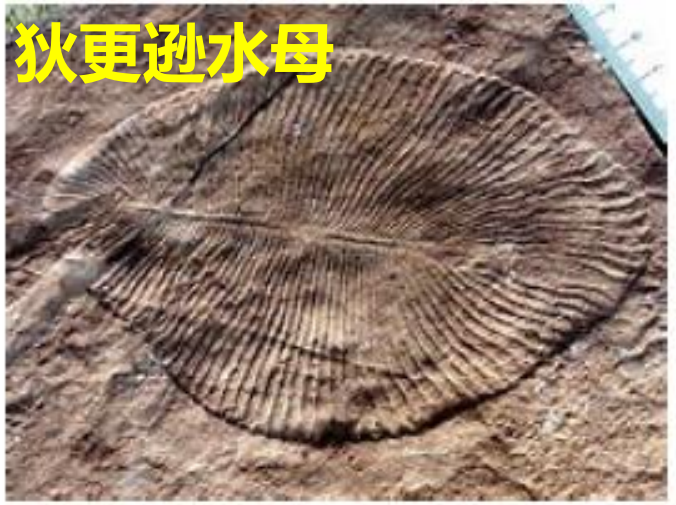
/ 人类在视觉方面的认识和努力

# 2.1 进化大爆炸

## 寒武纪大爆发, 530-540 million years, B.C.





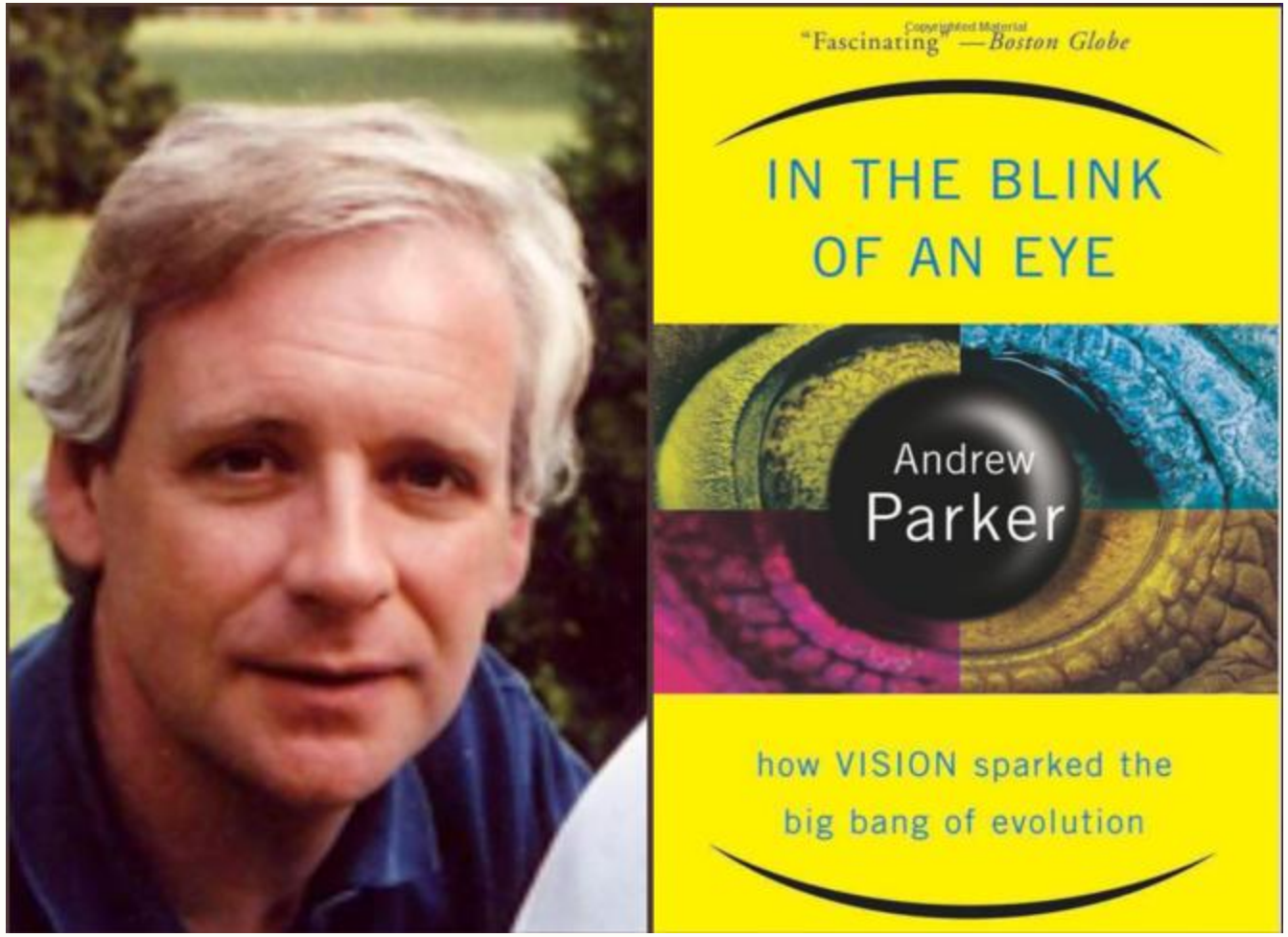



# 2.1 进化大爆炸

## 寒武纪大爆发, 530-540 million years, B.C.

“寒武纪大爆发是由视觉的突然进化引发的.“, 这引发了一场关于进化的军备竞赛，这个时期的动物要么进化，要么死亡。

——Andrew Parker, zoologist



# 2.1 进化大爆炸


结构复杂、高清且灵敏度高、多方向、
复杂颜色感知

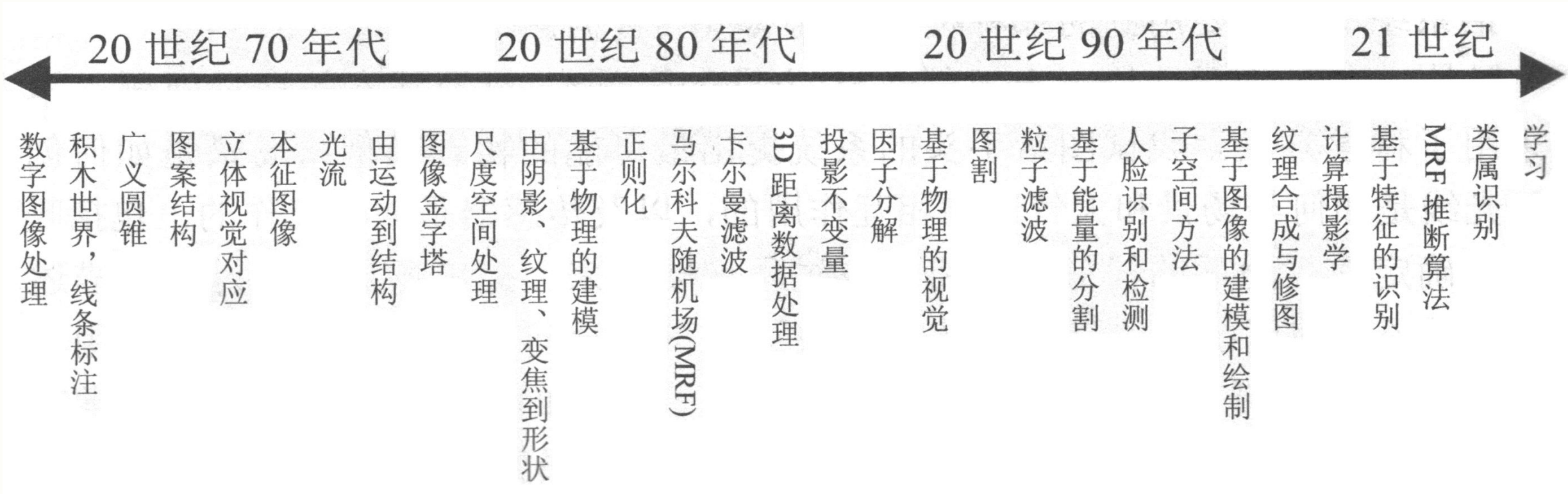

多方向复眼、高速感知、紫外感知、
偏振光导航


双眼独立运动、旋转角度大、紫外感
知、高清视觉系统


高分辨率、双眼立体视觉、高级色彩
感知、对环境高度适应性和调节能力、
与大脑连接具有强大处理能力

# 2.2 计算机视觉的发展

## 近50年来计算机视觉领域最活跃的主题

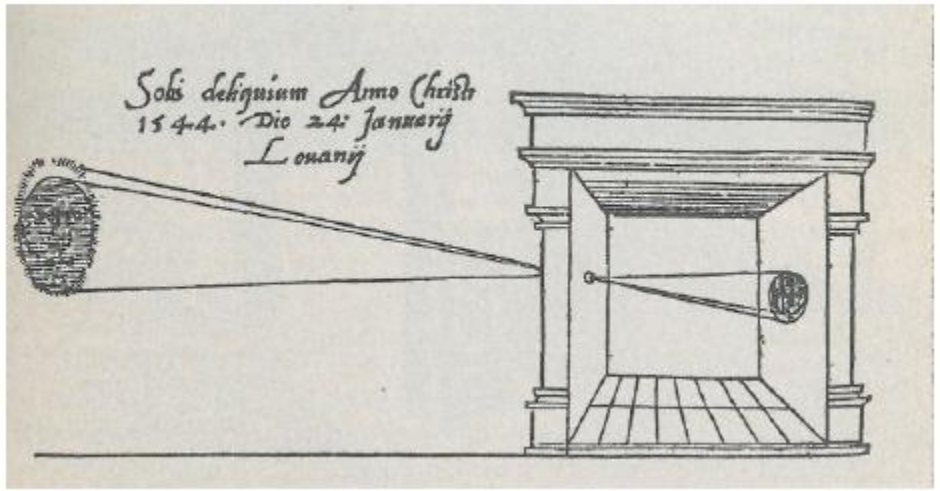


# 2.3 人类在视觉方面的认识和努力

## 照相机的发明

小孔成像

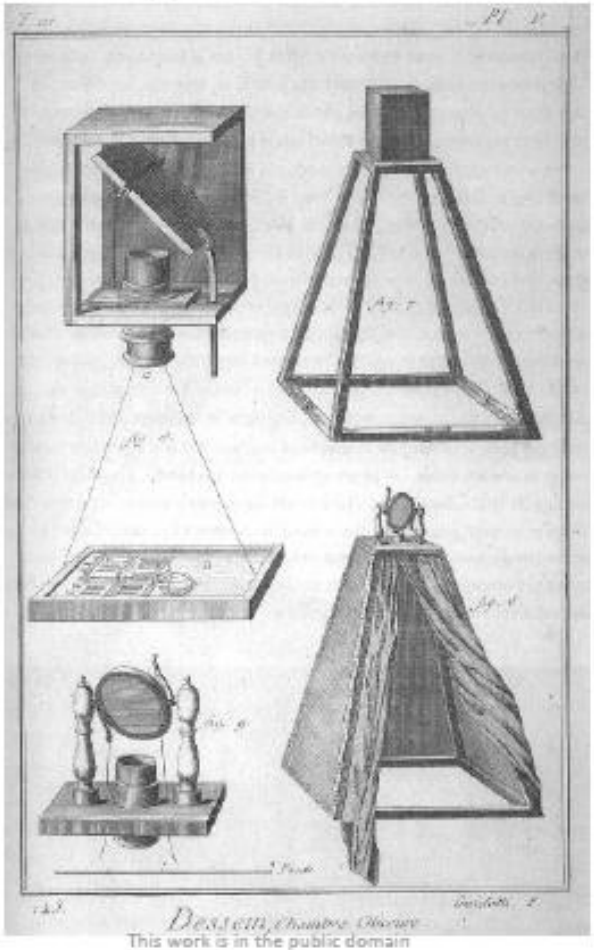
Gemma Frisius, 1545






暗箱摄影

Encyclopedia, 18th Century



# 2.3 人类在视觉方面的认识和努力

## Where did we come from?

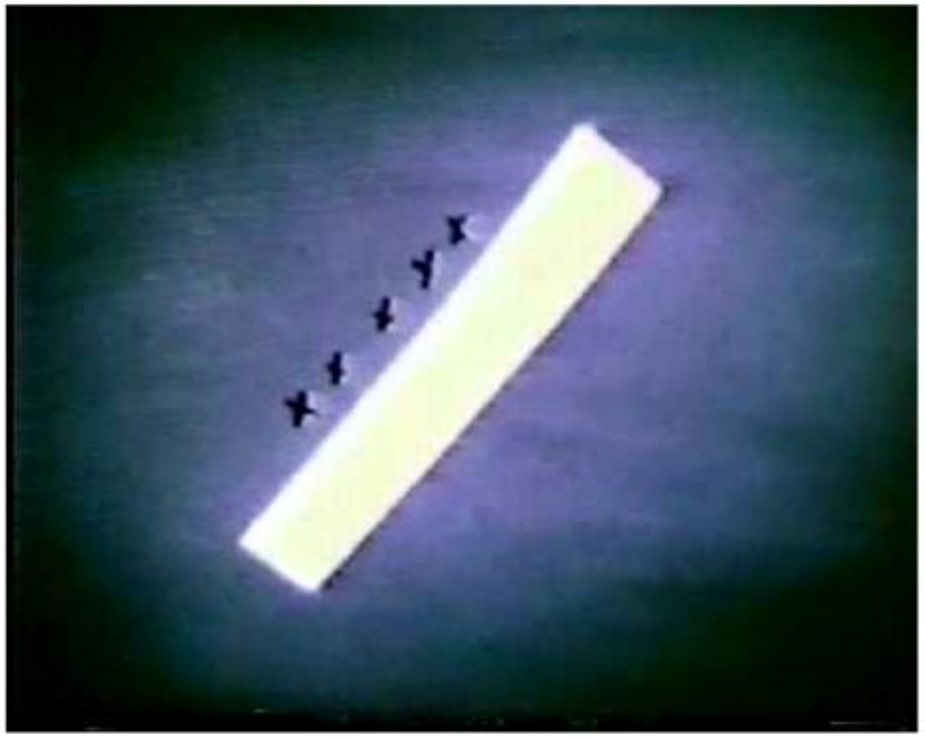
**The known story — Neuroscience inspired AI**

**已知的故事 — 受神经系统学启发的人工智能(AI)**

# 2.3 人类在视觉方面的认识和努力

## 移动的边缘对初级皮层的神经元具有刺激作用

Hubel and Wiesel, 1959



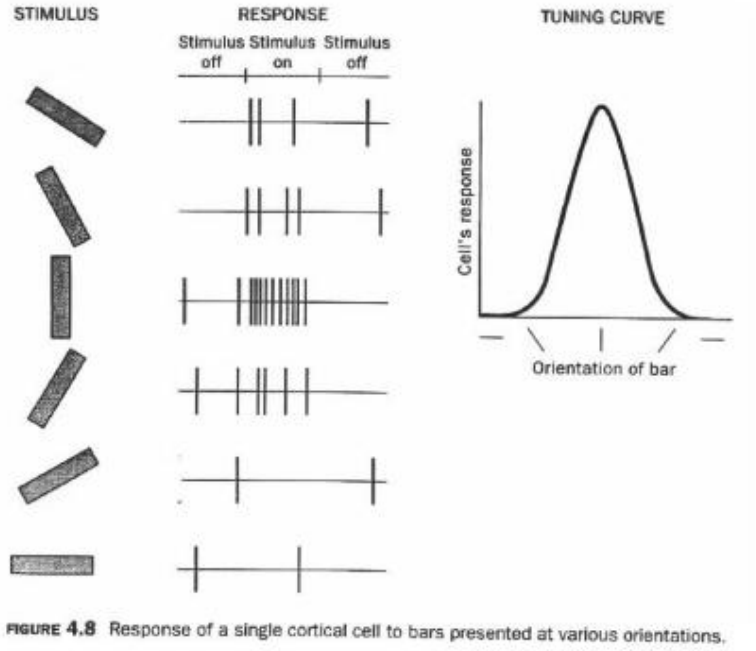
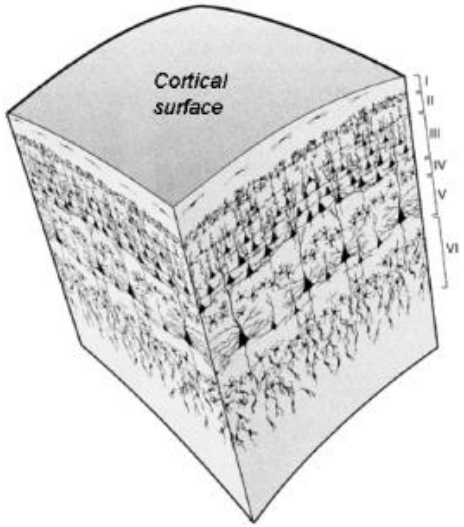



FIGURE 4.8 Response of a single cortical cell to bars presented at various orientations.

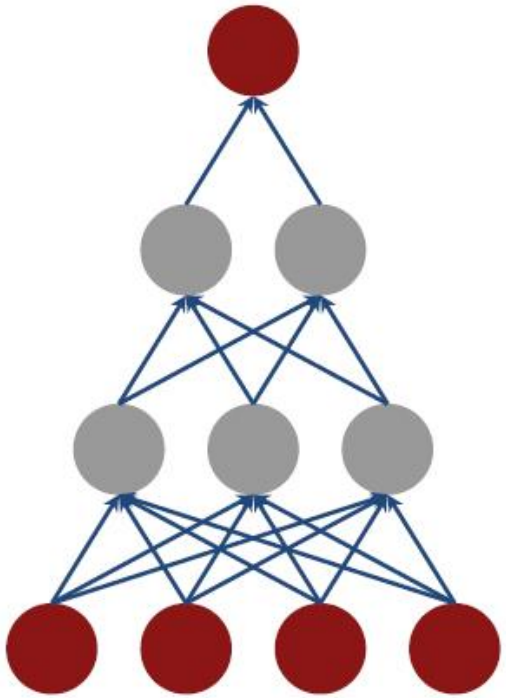
Hubel和Wiesel在1958年的猫视觉皮层实验中，首次观察到视觉初级皮层的神经元对移动的边缘刺激敏感，并定义了简单和复杂细胞，发现了视功能柱结构。此项工作为视觉神经研究奠定了重要的基础，两人在1981年共享了诺贝尔生理学或医学奖，以表彰他们在“**视觉系统信息加工**”的重要贡献。

# 2.3 人类在视觉方面的认识和努力

## 生物视觉神经的研究对神经网络的设计具有启发性




皮质柱
(生物学)



神经网络
(数字化)

高级模式
(语义)

低级模式
(像素)

# 2.3 人类在视觉方面的认识和努力

## 从感知机到BP神经网络

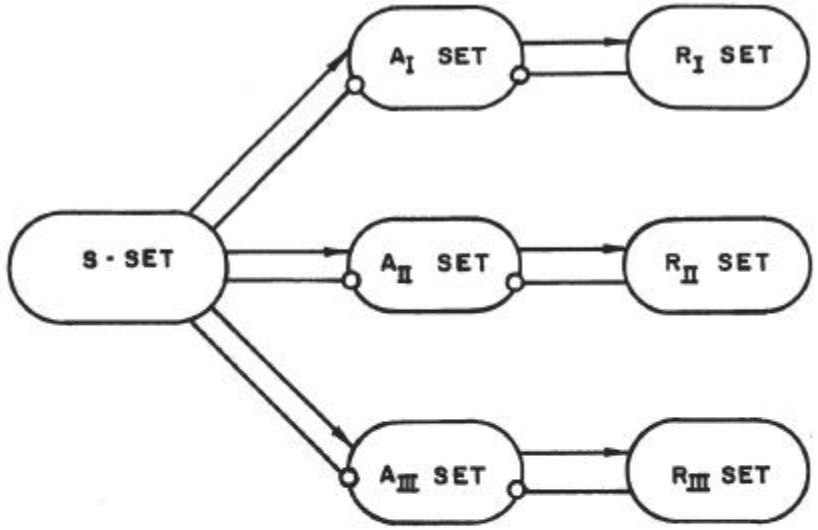



FIGURE 2

ORGANIZATION OF A PERCEPTRON WITH THREE INDEPENDENT OUTPUT-SETS

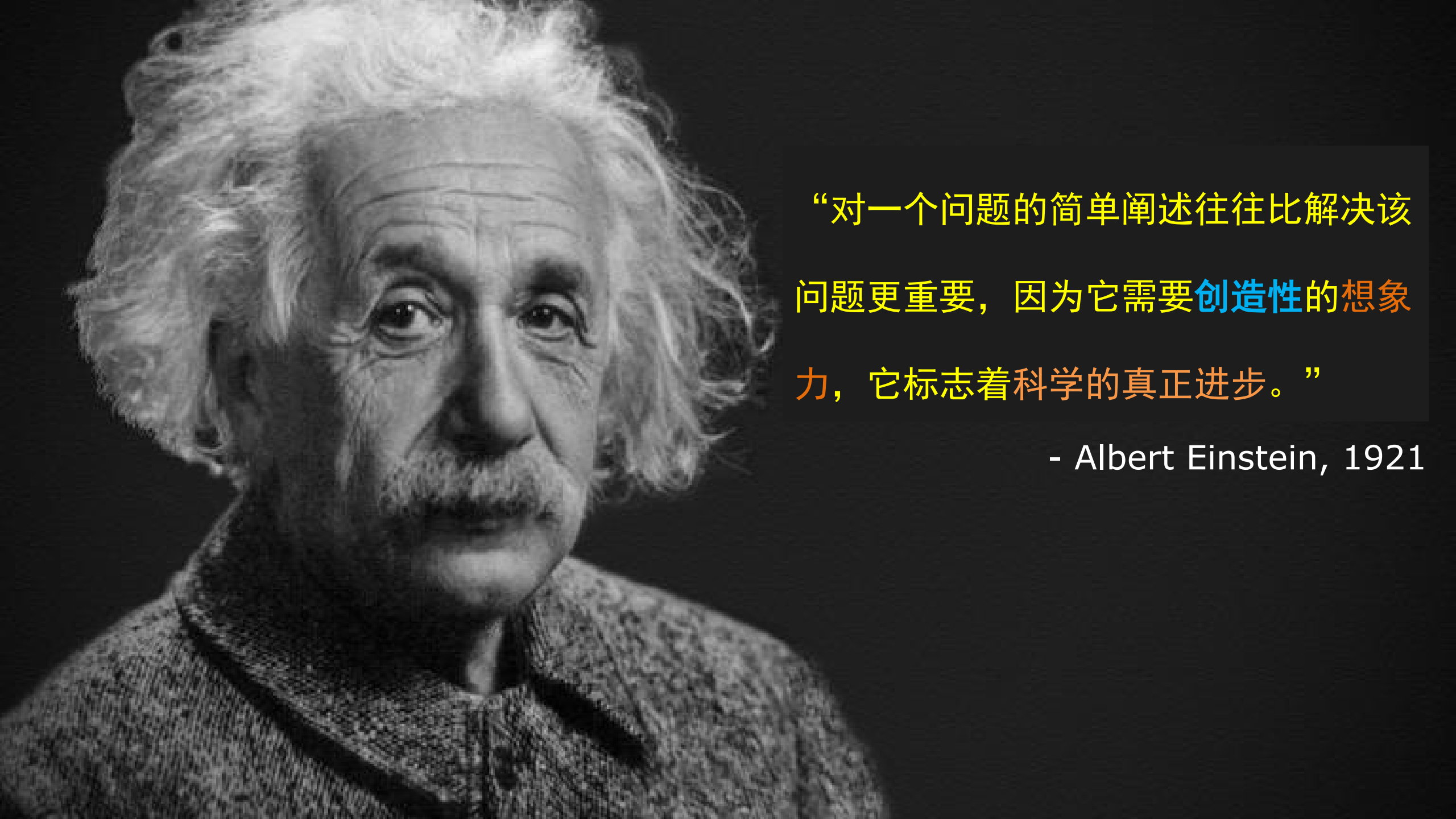
F. Rosenblatt, 1957

**Learning representations by back-propagating errors**

David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

Rumelhart, Hinton & Williams, 1986


“对一个问题的简单阐述往往比解决该问题更重要，因为它需要创造性的想象力，它标志着科学的真正进步。”
- Albert Einstein, 1921

# 2.3 人类在视觉方面的认识和努力

## Where did we come from?

**The not-so-known story – the search for computer vision's "North Star"**

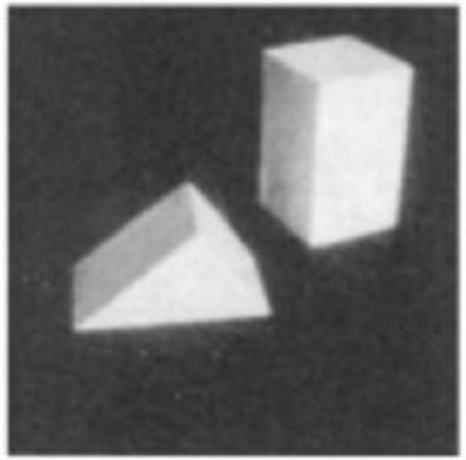
**不是那么广为人知的故事 — 寻找计算机视觉的“北极星”**

# 2.3 人类在视觉方面的认识和努力

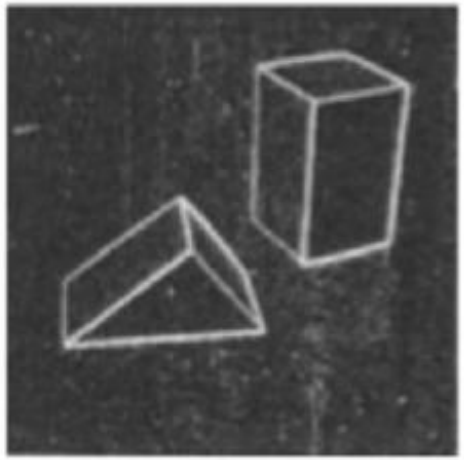
## 1960s: 对合成世界的理解

Larry Roerts
1963, 1st thesis of Computer Vision

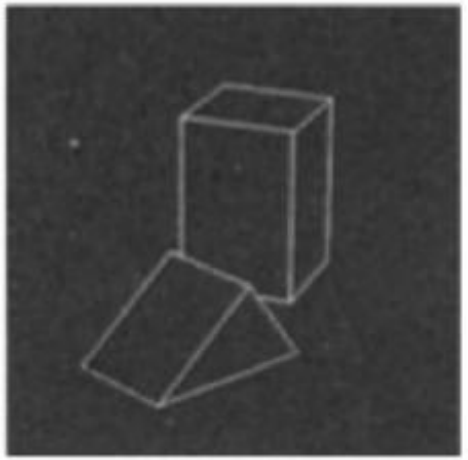
通过对输入图像的梯度因子建模构建新的3D模型



Input image



2x2 gradient operator



computed 3D model
rendered from new viewpoint

# 2.3 人类在视觉方面的认识和努力

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group　　July 7, 1966

Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

第一个计算机视觉项目于1966年在麻省理工大学成立。

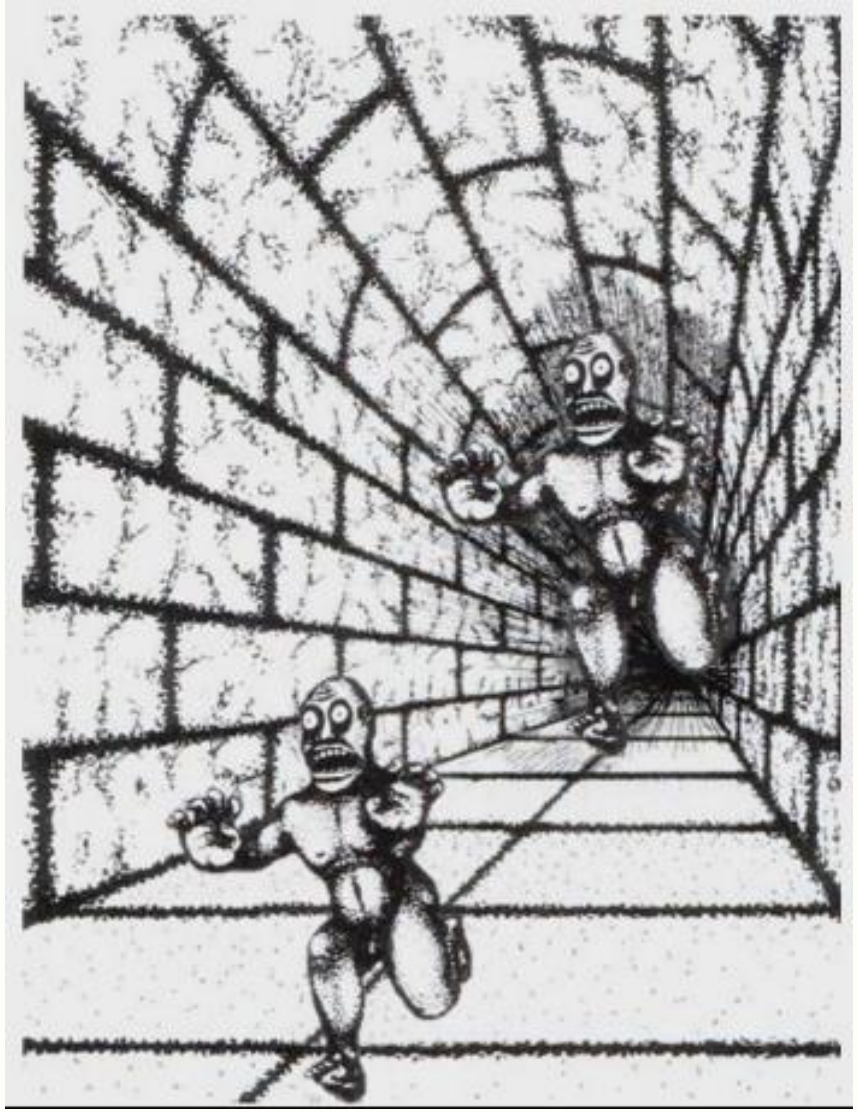

对于1966年的人类世界?

天方夜谭!

视觉获得的东西并不是那么容易去理解，甚至是一种假象。

# 2.3 人类在视觉方面的认识和努力

## 测量像素 vs. 场景理解

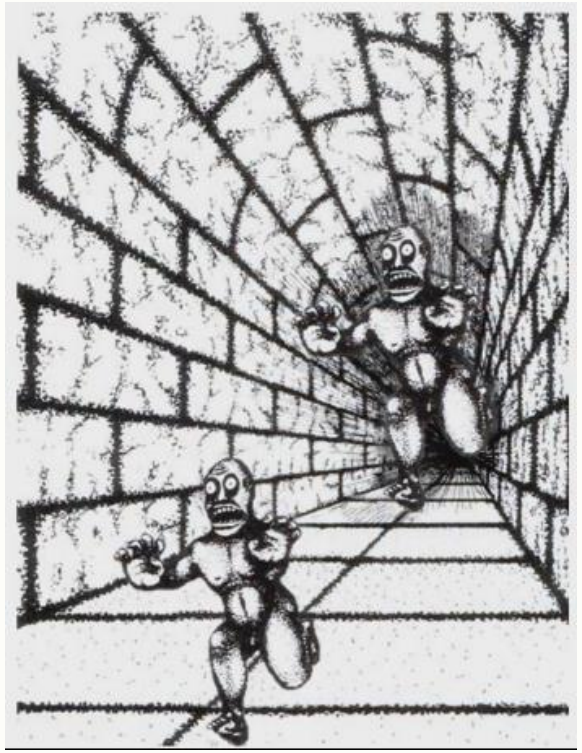


- **从这张图片中我们能看到什么？得到什么信息？**
  - ✓ 一个大怪物正在追逐一个小怪物
  - ✓ 小怪物在前面惊恐地逃跑
  - ✓ 大怪物在后面凶狠地追逐
  - ✓ 隧道非常的深邃，也许他们已经跑了很长的时间了
  - ✓ ……

- **真实情况是什么？**
  - ✓ 一张二维图片（图片本质）
  - ✓ 小怪物在图片的中下部分，大怪物在图片的中间靠右的部分（相对位置关系）
  - ✓ 两个怪物长相一模一样（外观特征）

# 2.3 人类在视觉方面的认识和努力

## 什么是计算机视觉？计算机视觉的根本任务是什么？



图片

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 49 | 159 | 185 | 235 | 58 | 105 | 154 | 222 | 221 |
| 143 | 36 | 165 | 112 | 44 | 236 | 224 | 185 | 217 |
| 32 | 247 | 196 | 234 | 39 | 123 | 87 | 221 | 236 |
| 62 | 64 | 210 | 213 | 239 | 245 | 156 | 54 | 133 |
| 76 | 23 | 188 | 243 | 27 | 72 | 20 | 21 | 48 |
| 221 | 179 | 102 | 241 | 235 | 203 | 141 | 56 | 236 |
| 38 | 218 | 217 | 57 | 213 | 164 | 157 | 31 | 109 |
| 17 | 193 | 171 | 217 | 106 | 75 | 92 | 97 | 112 |
| 90 | 196 | 198 | 104 | 131 | 215 | 9 | 125 | 70 |
| 104 | 31 | 162 | 4 | 229 | 201 | 103 | 137 | 48 |
| 221 | 226 | 207 | 154 | 14 | 131 | 148 | 237 | 64 |
| 168 | 64 | 36 | 49 | 69 | 49 | 182 | 220 | 48 |
| 254 | 53 | 59 | 57 | 123 | 137 | 16 | 211 | 220 |
| 53 | 171 | 249 | 155 | 124 | 147 | 210 | 210 | 230 |

真实数据

- **真实情况是什么？**
  - ✓ 一张二维图片（图片本质）
  - ✓ 小怪物在图片的中下部分，大怪物在图片的中间靠右的部分（相对位置关系）
  - ✓ 两个怪物长相一模一样（外观特征）
- **从这张图片中我们能看到什么？得到什么信息？**
  - ✓ 一个大怪物正在追逐一个小怪物
  - ✓ 小怪物在前面惊恐地逃跑
  - ✓ 大怪物在后面凶狠地追逐
  - ✓ 隧道非常的深邃，也许他们已经跑了很长的时间了
  - ✓ ……

从表面的语义到知识

**跨越“语义鸿沟”，建立像素到语义的映射，实现知识的获取**

# 2.3 人类在视觉方面的认识和努力

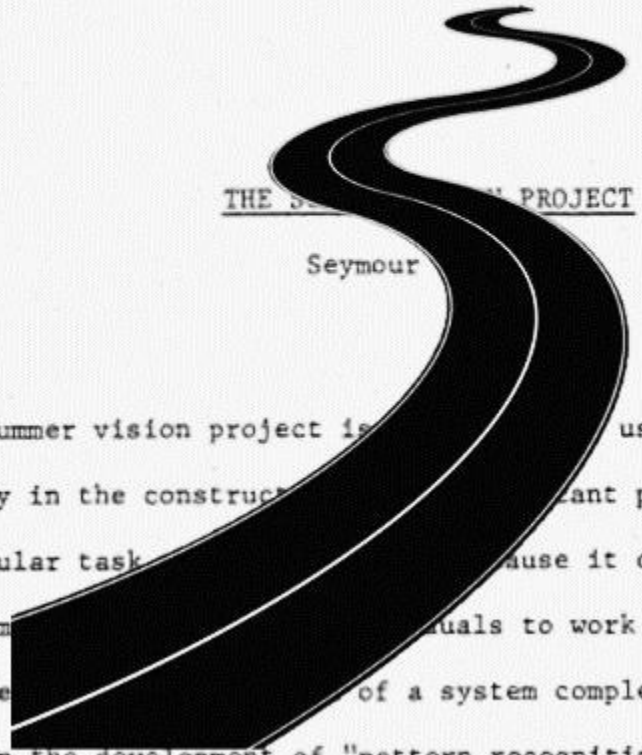
## 计算机视觉要像人的视觉一样，要走的路还很远

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966

Vision Memo. No. 100.



THE S PROJECT

Seymour

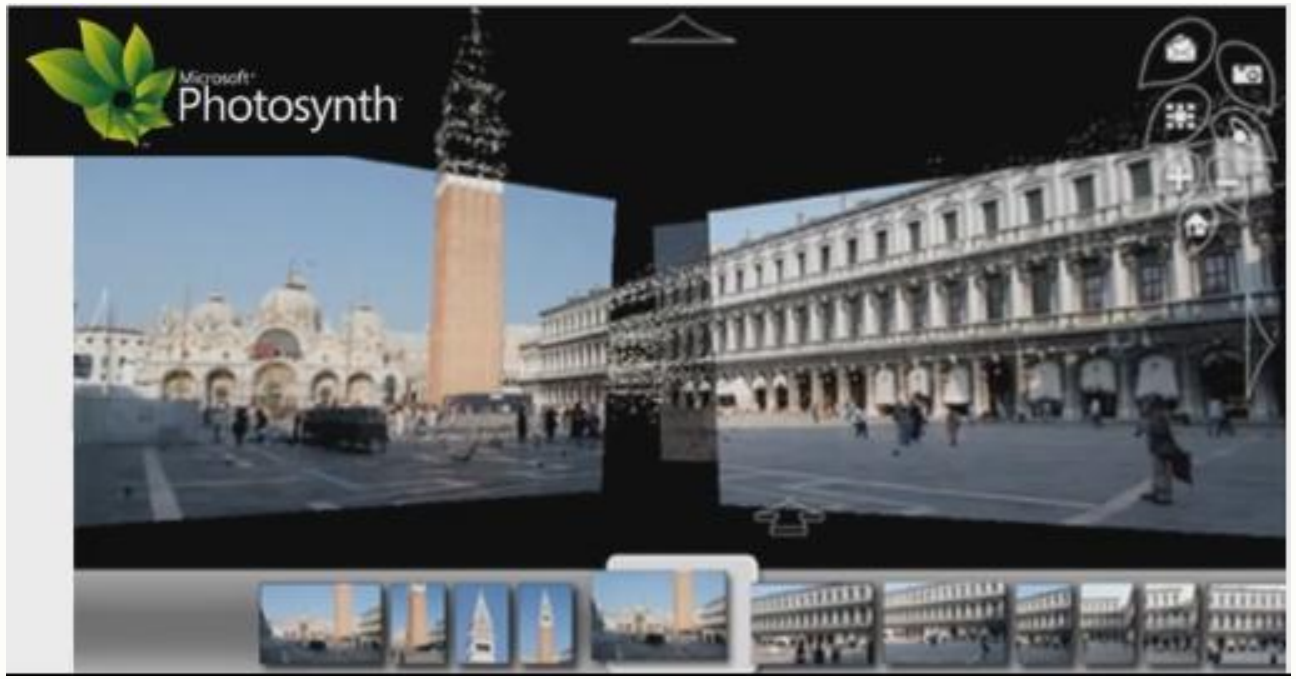
The summer vision project is use our summer workers effectively in the construct ant part of a visual system. The particular task use it can be segmented into sub-problem uals to work independently and yet participate of a system complex enough to be a real landmark in the development of "pattern recognition".

# 2.3 人类在视觉方面的认识和努力

## 即使那些今天看起来很牛的应用依然不完善


实景导航

图像拼接



# 2.3 人类在视觉方面的认识和努力

## 即使那些今天看起来很牛的应用依然不完善


体感游戏

# 2.3 人类在视觉方面的认识和努力

## 一个简单的搜索应用也可能会出现巨大的偏差

# 2.3 人类在视觉方面的认识和努力

原因在哪里？

袋熊

**各种各样的形态**

# 2.3 人类在视觉方面的认识和努力

**解决对象识别的三个基础元素：**

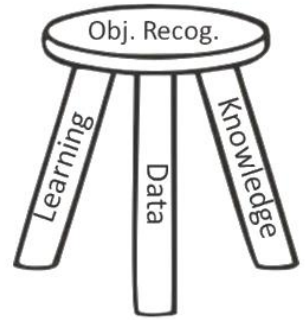
**数据、学习和知识。**

# 2.3 人类在视觉方面的认识和努力

## 对象识别的三个重要元素

# 2.3 人类在视觉方面的认识和努力

## 起源阶段：计算机和互联网出现之前

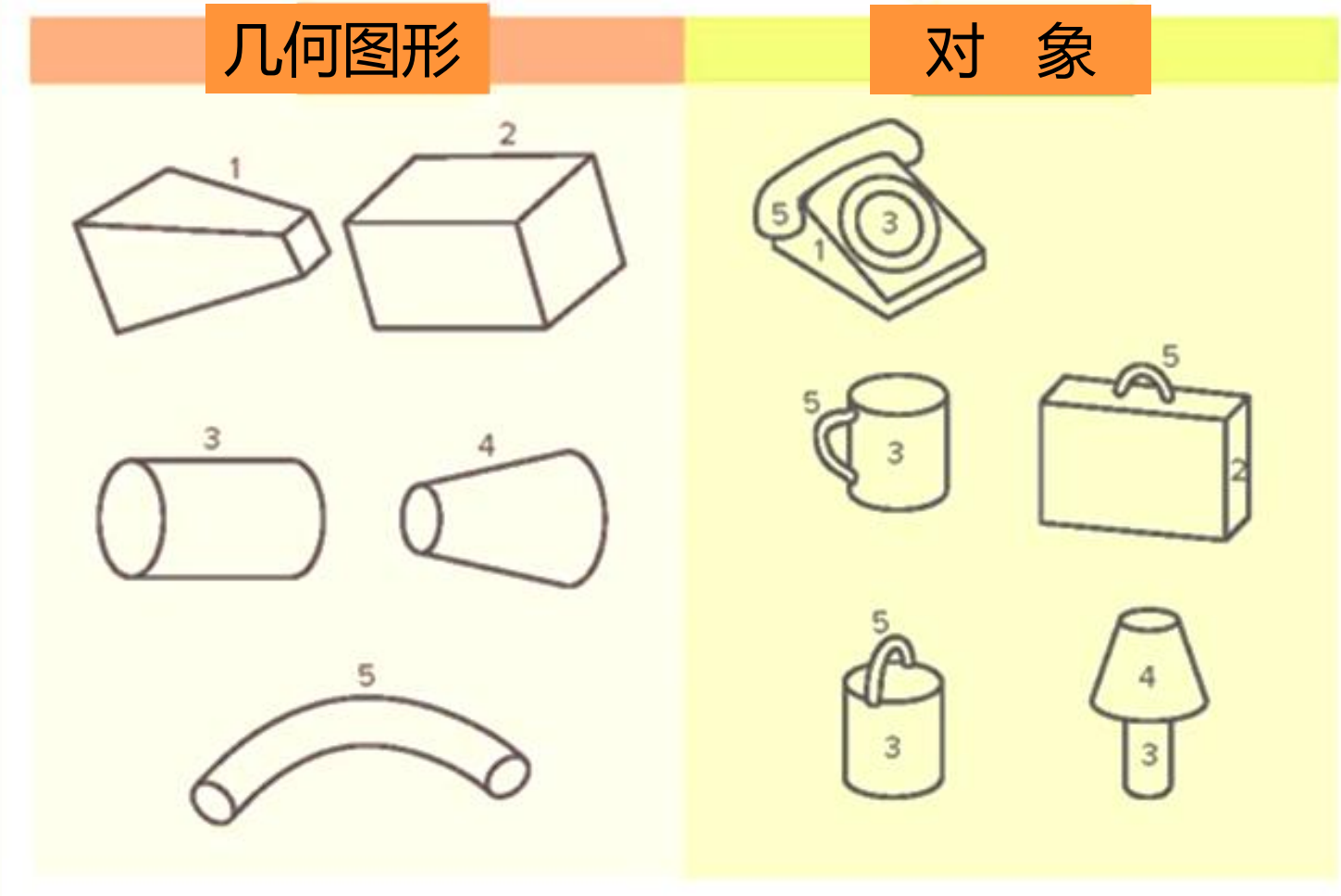




- ✓ 没有数据
- ✓ 拥有少量的统计学习的知识
- ✓ “机器学习Machine Learning”几乎没有
- ✓ 互联网也没有出现
- ✓ 在你们所有同学出生之前

# 2.3 人类在视觉方面的认识和努力

## 起源阶段：从几何图像到对象

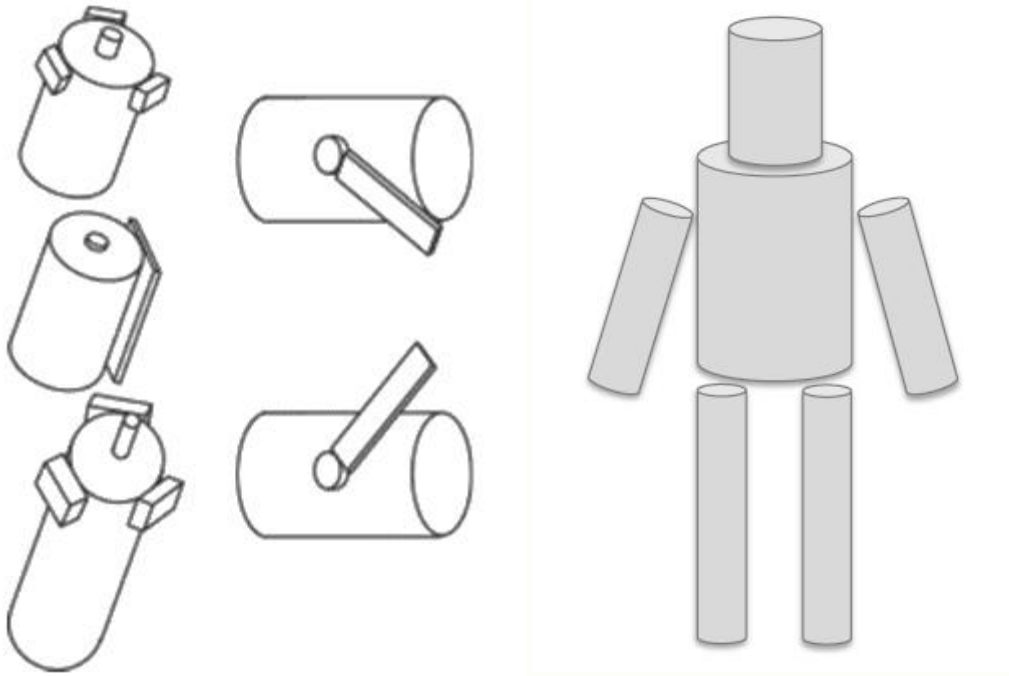



Biederman, 1970s-80s

# 2.3 人类在视觉方面的认识和努力

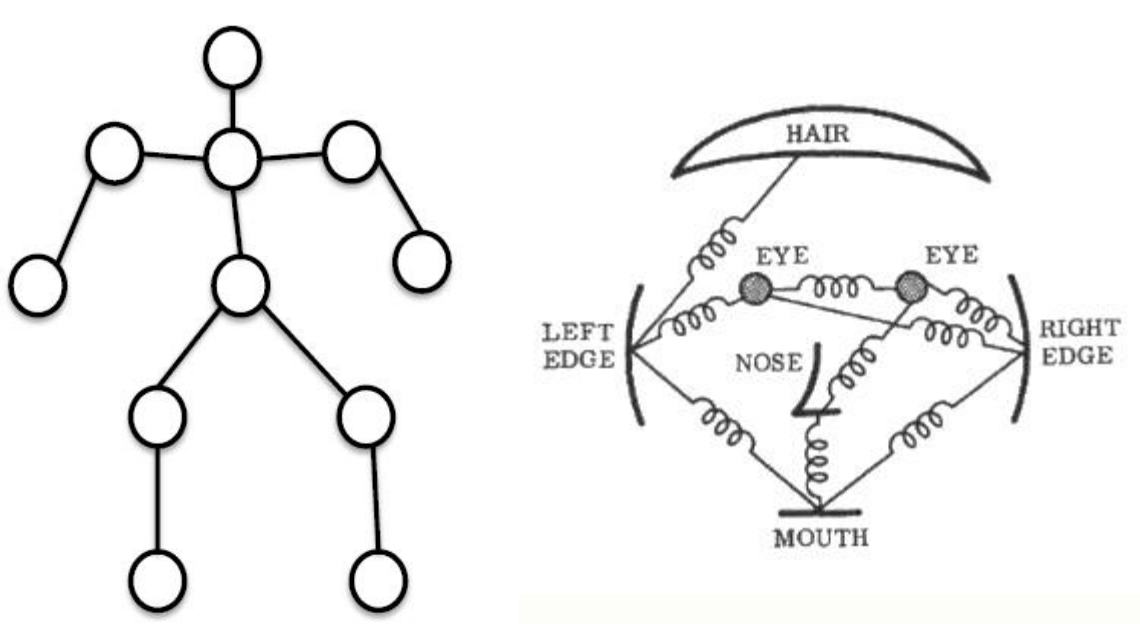
## 起源阶段：基于基本结构组合

Brooks & Binford, 1979



广义圆柱

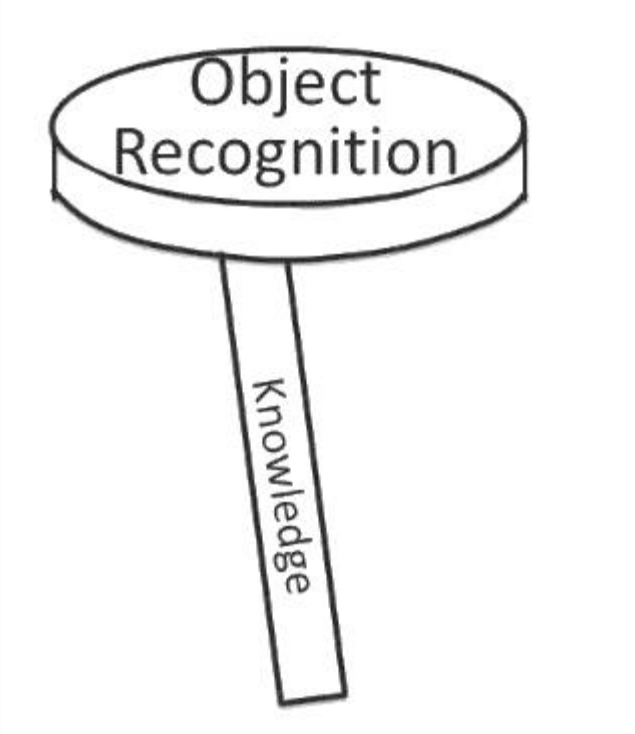
Brooks & Binford, 1979




图结构模型

# 2.3 人类在视觉方面的认识和努力

## 起源阶段：对象识别的黎明（1970s-1980s)



# 2.3 人类在视觉方面的认识和努力

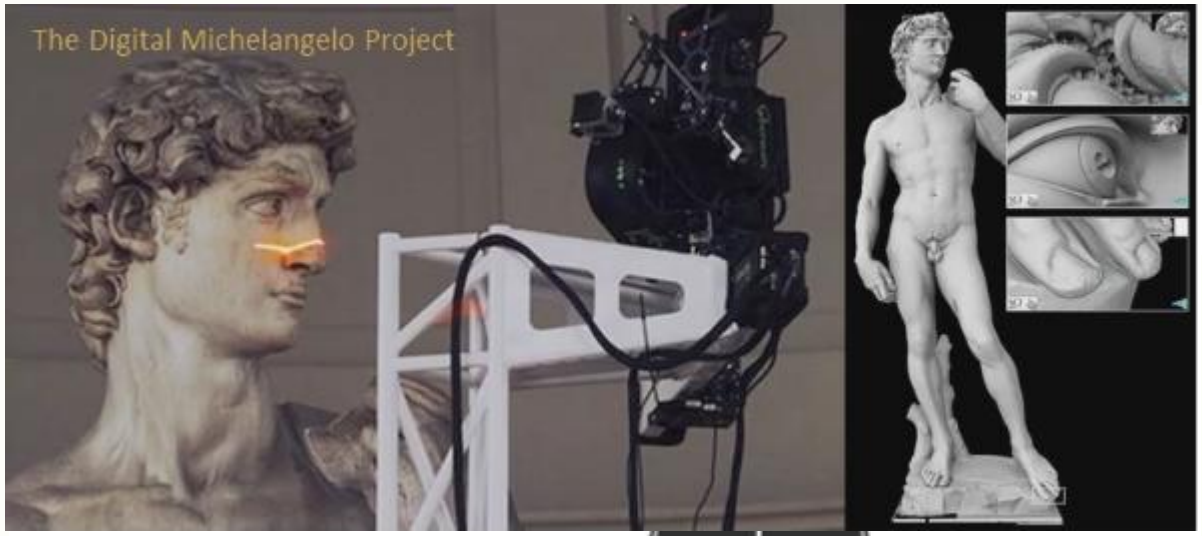
## 起源阶段：对象识别的黎明（1970s-1980s)


虽然这个时代的计算机视觉是残缺的，但请记住这些先先驱所作出的努力。

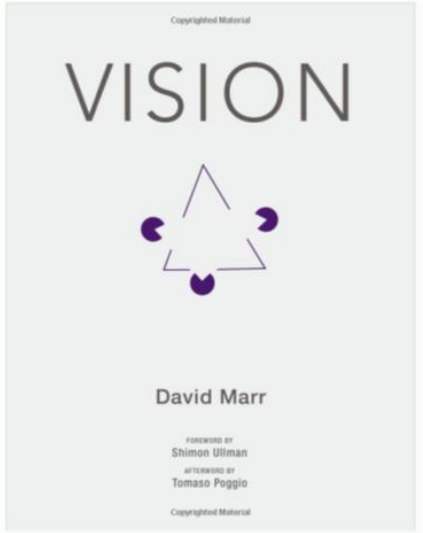
APRANet 分组网络　视觉表象　计算机视觉 机器人学家　模式识别

# 2.3 人类在视觉方面的认识和努力

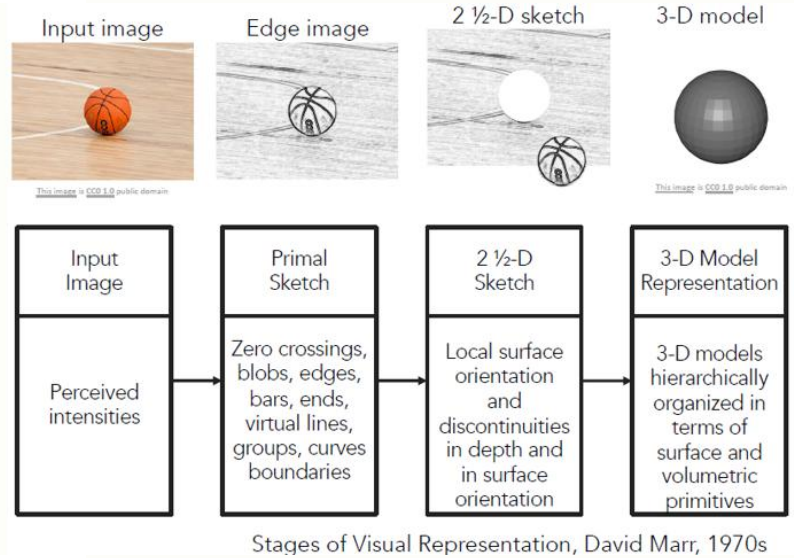
## 起源之后的20年，是三维重建的时代




数字米开朗基罗工程



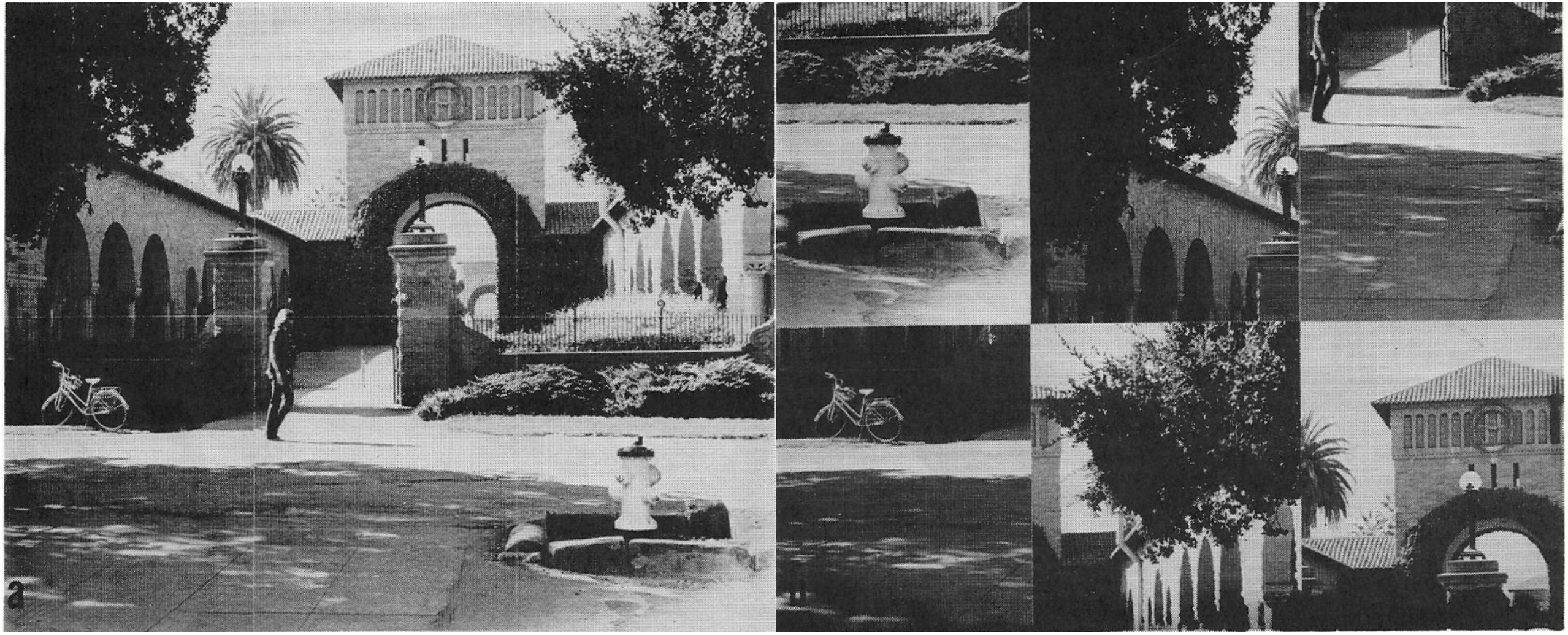

第一本关于计算机视觉的专著




Stages of Visual Representation, David Marr, 1970s

# 2.3 人类在视觉方面的认识和努力

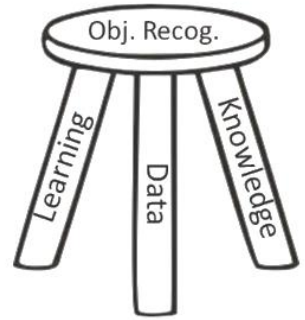
## 感知真实世界的场景



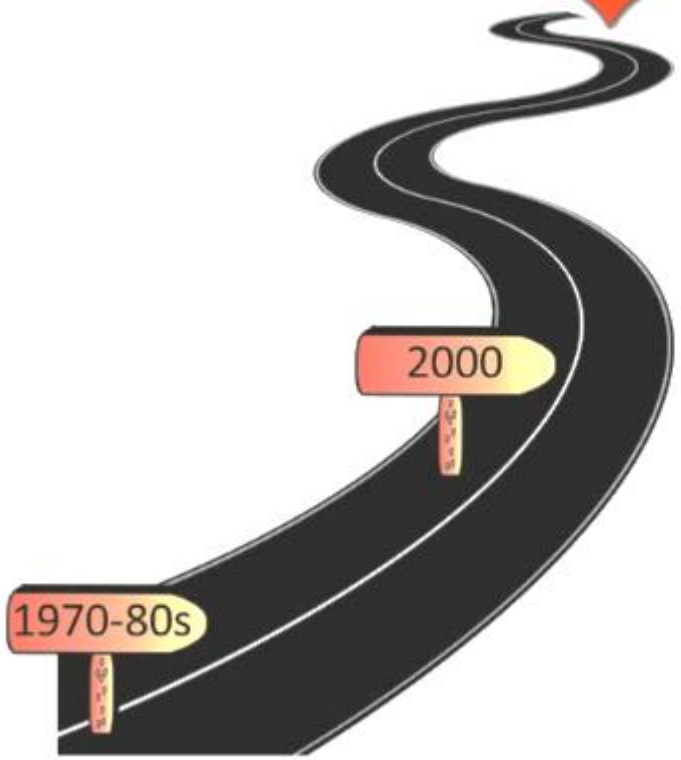

Irving Biederman, *Science*, 1972

# 2.3 人类在视觉方面的认识和努力

## 发展阶段：机器学习时代的到来




**计算机视觉 Computer Vision** ❤ **机器学习 Machine Learning**




***机器学习得到了前所未有的发展***

- ✓ **支持向量机(SVM):** Vapnik et al. 1995
- ✓ **自适应增强(AdaBoost):** Freund & Schapire 1995
- ✓ **图模型(Graphical models):** Pearl 1988, Bishop 1995
  - 马尔可夫随机场(MRF), 条件随机场(CRF), 蒙特卡罗(MCMC), Gibbs采样, 变分(Variational), 非参贝叶斯(Non-parametric Bayes)
- ✓ **神经网络(Neural network)：** 许多研究者 1950s
- ✓ ……

# 2.3 人类在视觉方面的认识和努力

## 机器学习时代的到来：边缘，分割和感知

D. Lowe. *IJCV*, 1992

# 2.3 人类在视觉方面的认识和努力

## Normalized Cut (Shi & Malik, 1997)

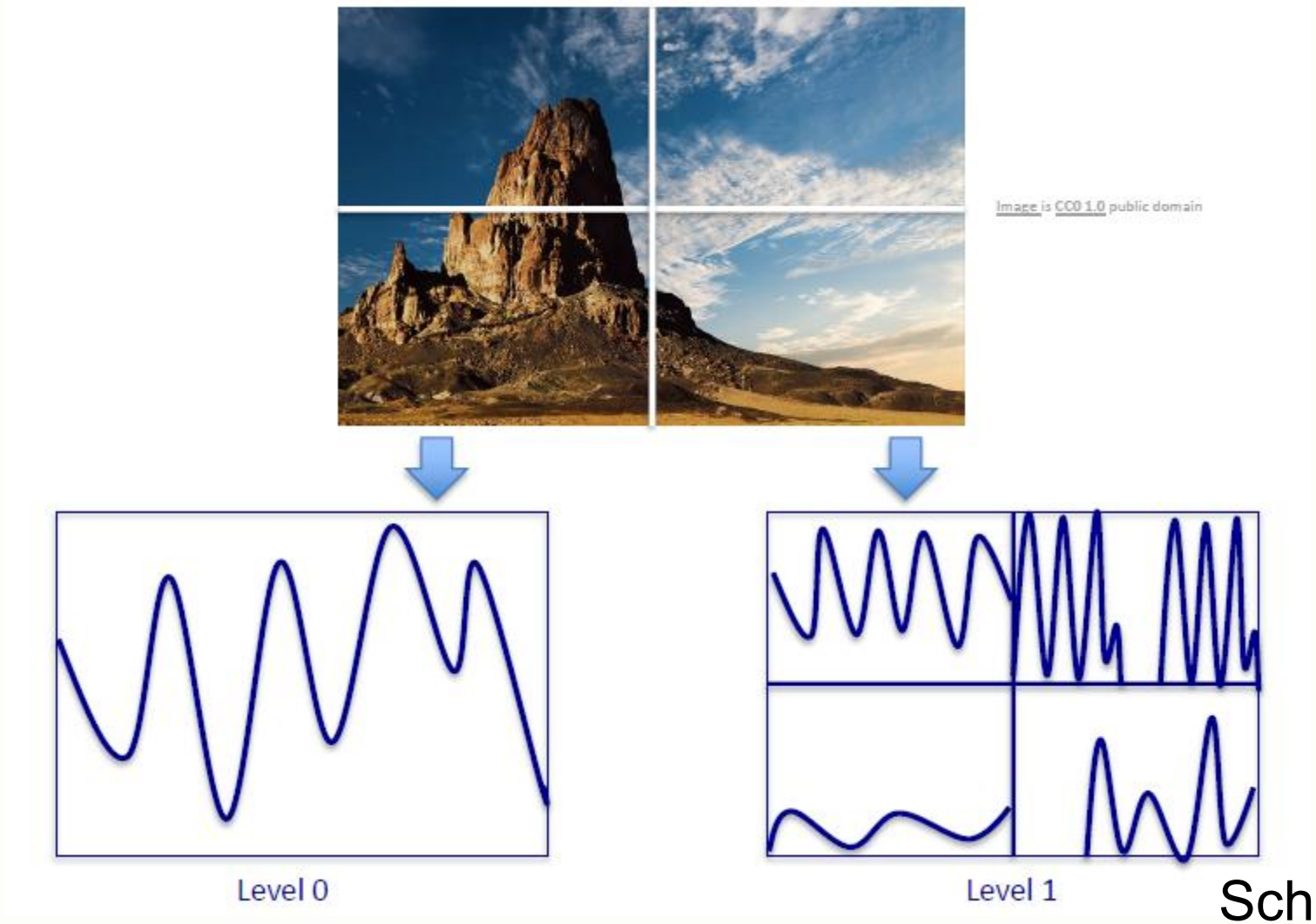
# 2.3 人类在视觉方面的认识和努力

## 单目标识别 —— 尺度不变特征变换(SIFT)


D. Lowe. *ICCV*, 1999

# 2.3 人类在视觉方面的认识和努力

## 空间金字塔匹配

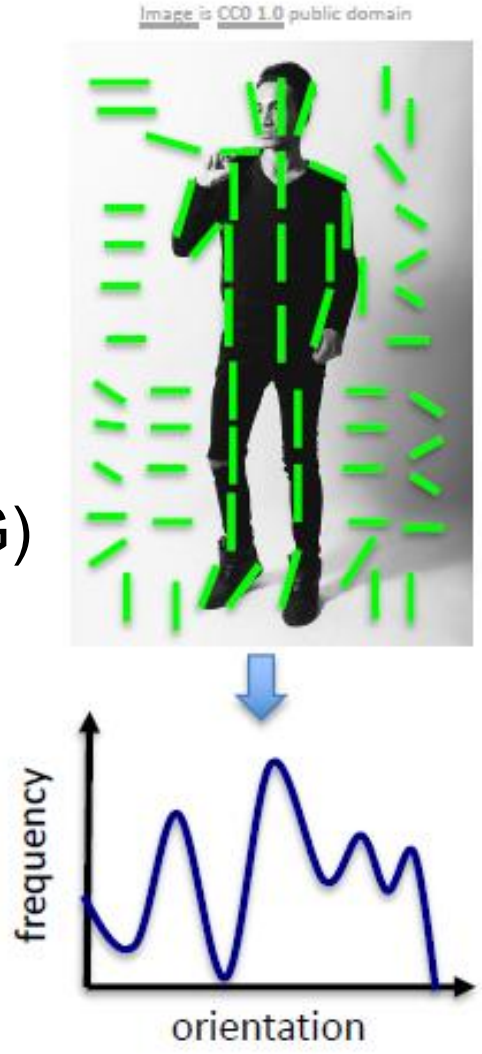


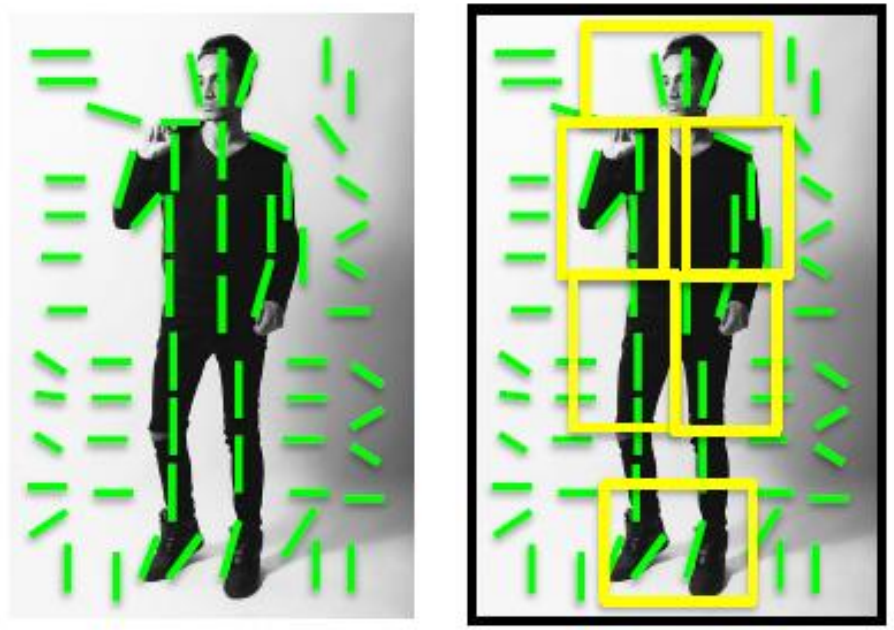

Schmid & Ponce, 2006

# 2.3 人类在视觉方面的认识和努力

## 基于梯度直方图的HoG和DPM

Histogram of Gradients (HoG)
Dalal & Triggs, 2005



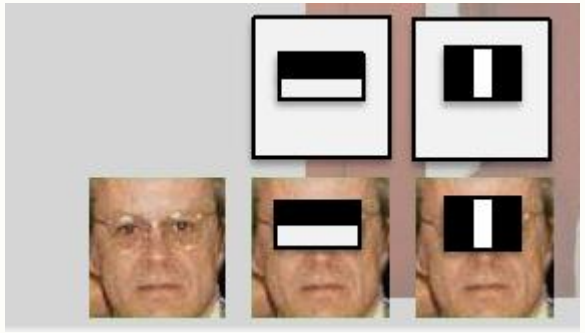



Deformable Part Model
Felzenswalb, McAllester, Ramanan, 2009

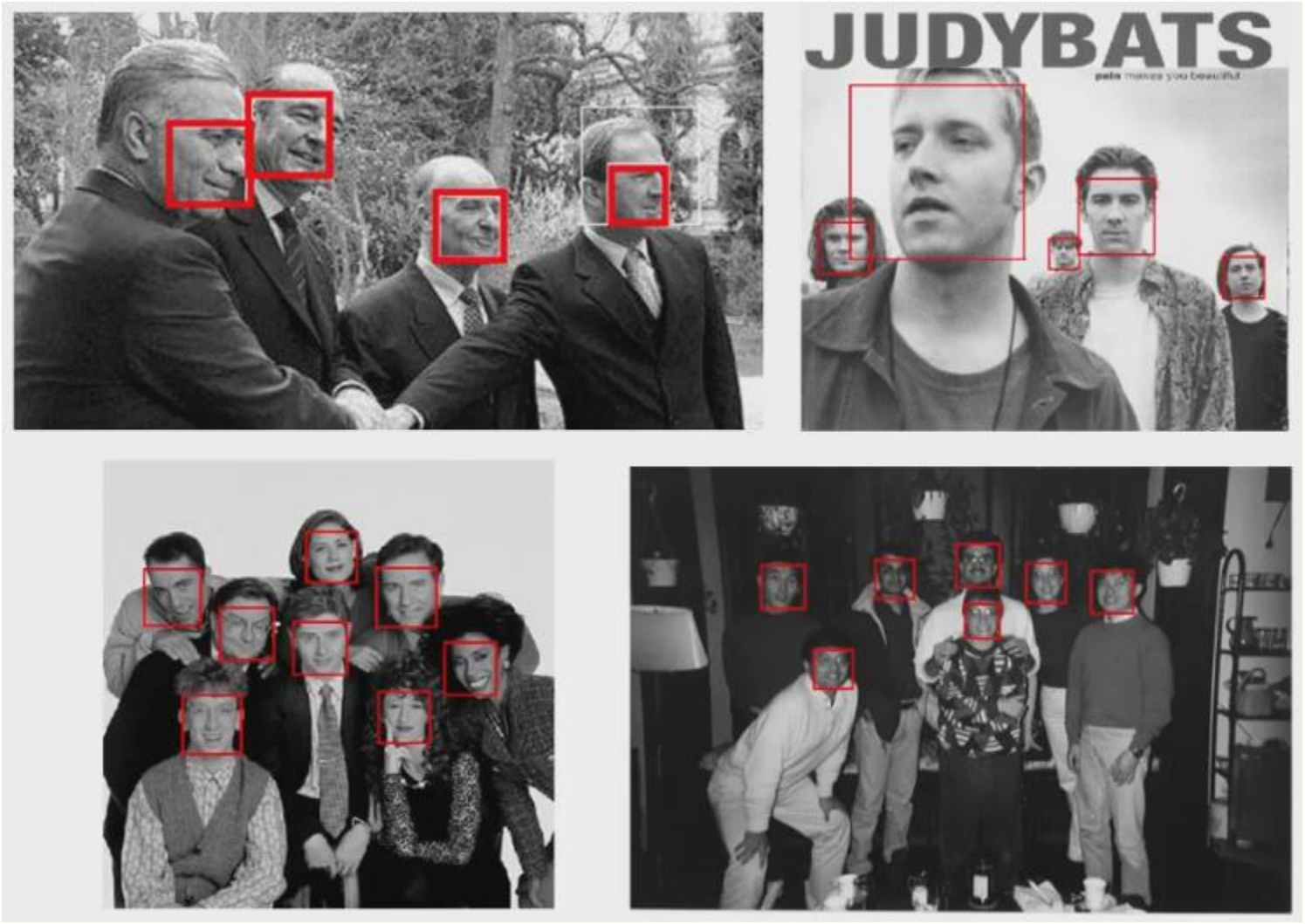
# 2.3 人类在视觉方面的认识和努力

## 机器学习时代的到来：实时人脸检测

基于人脸检测技术，FujiFilm于2006年推出了第一台人脸识别数码相机。

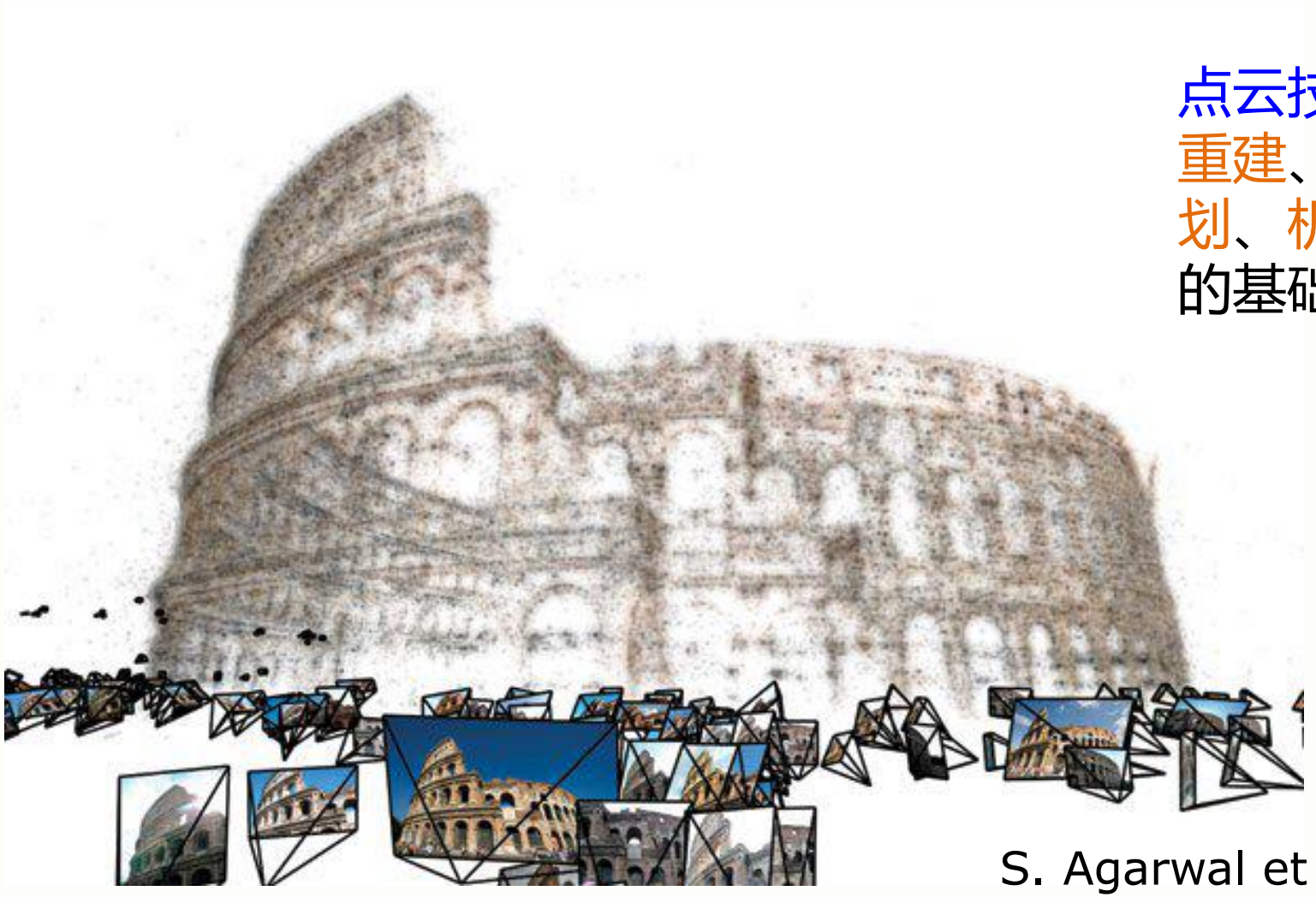


Viola & Jones, 2001



# 2.3 人类在视觉方面的认识和努力

## 三维重建（3D Reconstruction）



点云技术是当今三维重建、无人机路线规划、机器人路线规划的基础。

S. Agarwal et al. ICCV, 2009

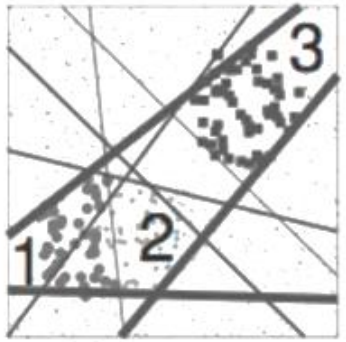
# 2.3 人类在视觉方面的认识和努力

## 多种经典机器学习模型不断涌现（2000-2010）

**图模型&星座模型**

Felzenszwalb et al. 2000
Fergus et al. 2003
Fei-Fei et al. 2003

**提升(Boosting)**

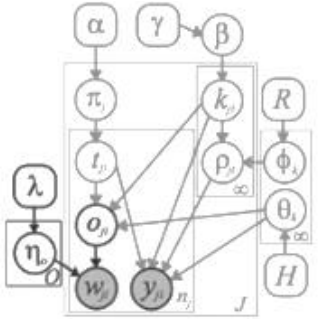



Viola & Jones 2001
Torralba et al. 2004

**词袋(Bag of Words)**

Leung et al. 1999; Sivic et al. 2003;
Grauman et al. 2005; Lazebnik et al. 2006;
Fei-Fei et al. 2005

**非参贝叶斯**

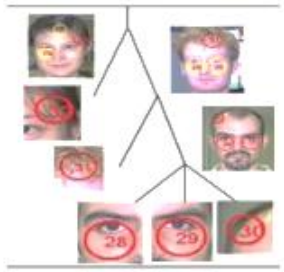


Sudderth et al. 2005
Li et al. 2007

**条件随机场**

Kumar et al. 2003
Gould et al. 2009

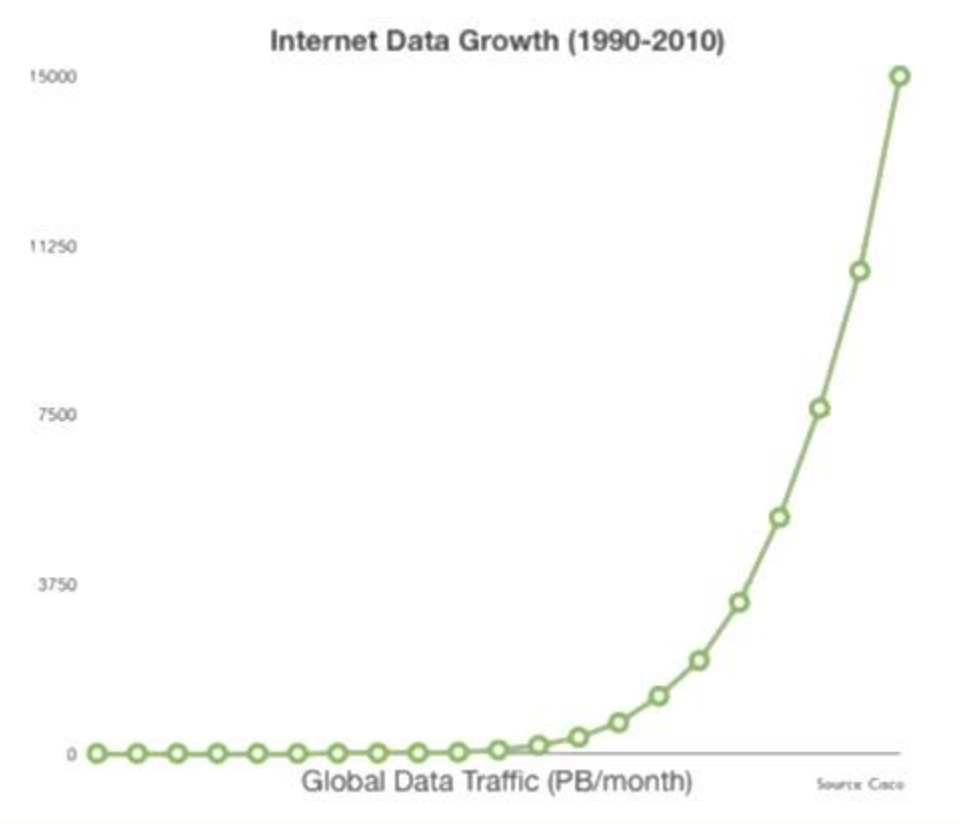
**与或图模型**



Chen et al. 2006
Zhu et al. 2007

# 2.3 人类在视觉方面的认识和努力

## 互联网数据暴涨（1990-2010）

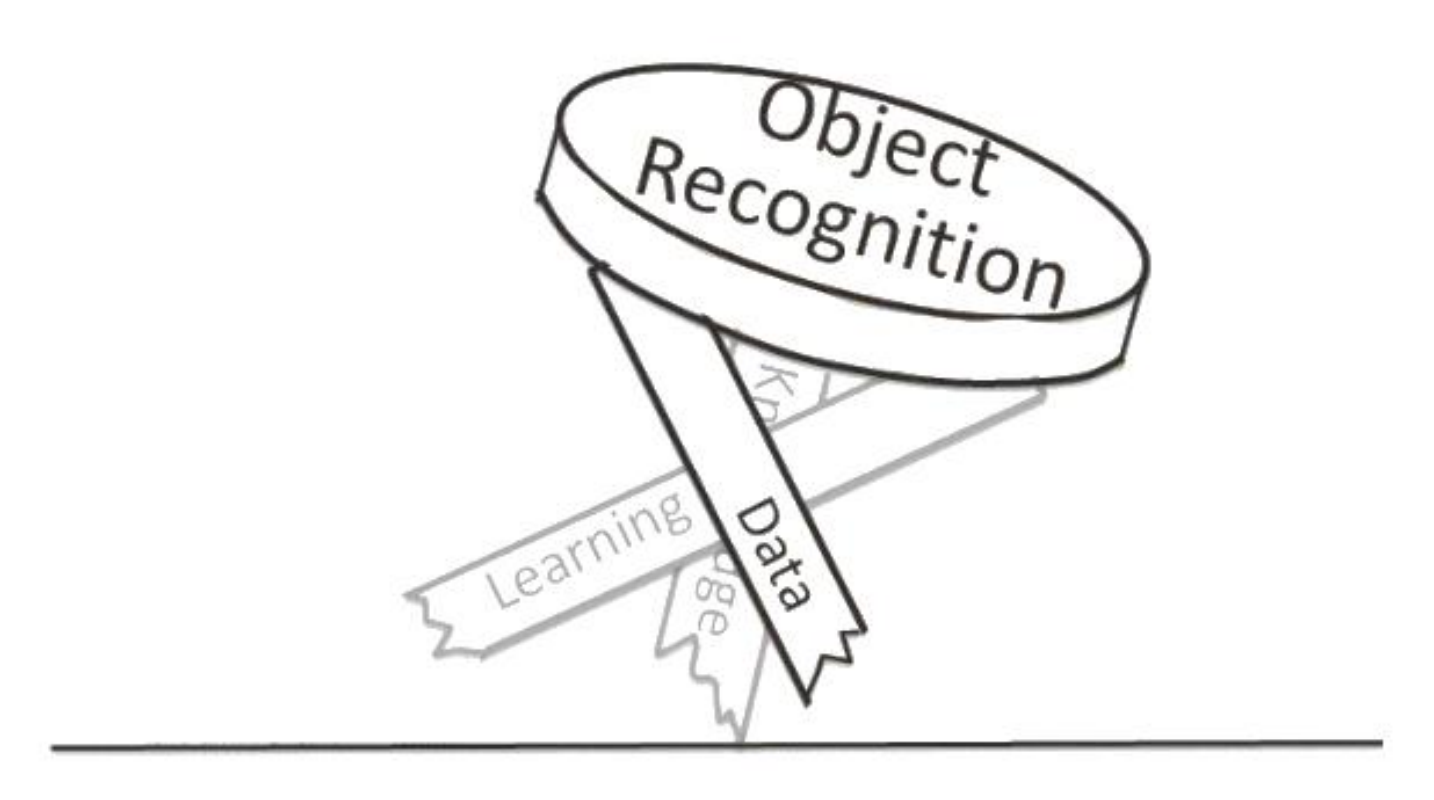



86%的数据为多媒体数据


**面向像素的计算机视觉识别成为互联网时代的重要任务**

# 2.3 人类在视觉方面的认识和努力

## 数据的暴涨并不意味着数据的可用




**数据需要能支持学习，能产生知识**

# 2.3 人类在视觉方面的认识和努力

## 计算机视觉/人工智能 的“圣杯” —— 标注数据


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

# 2.3 人类在视觉方面的认识和努力

## George A. Miller

Psychology, Cognitive Science
Princeton University

WORDNET

G. A. Miller, *Communications of the ACM*, 1995

# 2.3 人类在视觉方面的认识和努力

IMAGENET

22,000 categories ⋮ 15,000,000 images

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li & L. Fei-Fei. *CVPR*, 2009.

# 2.3 人类在视觉方面的认识和努力

IMAGENET Large Scale Visual Recognition Challenge

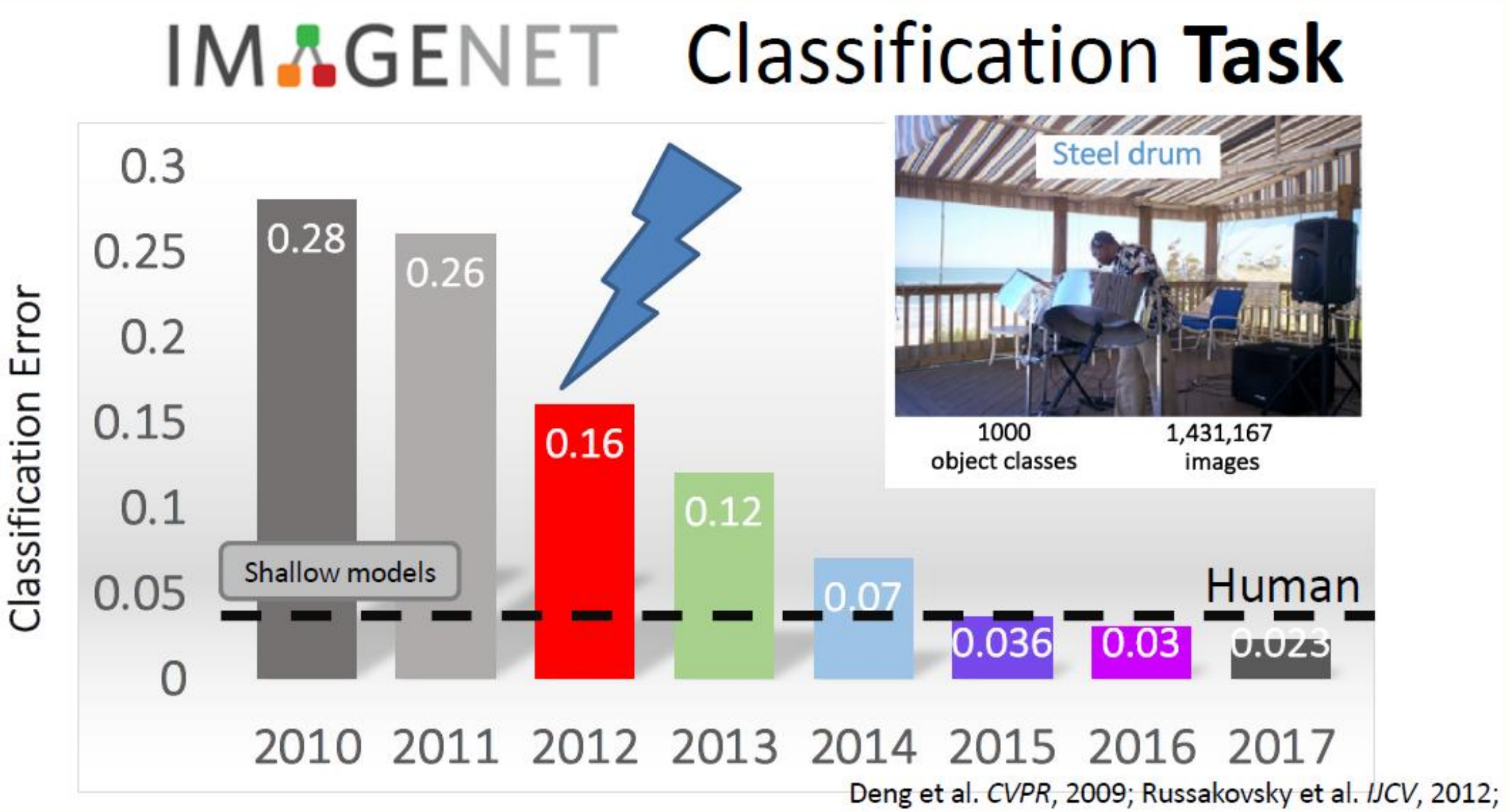
The Image Classification Challenge:
1,000 object classes
1,431,167 images

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

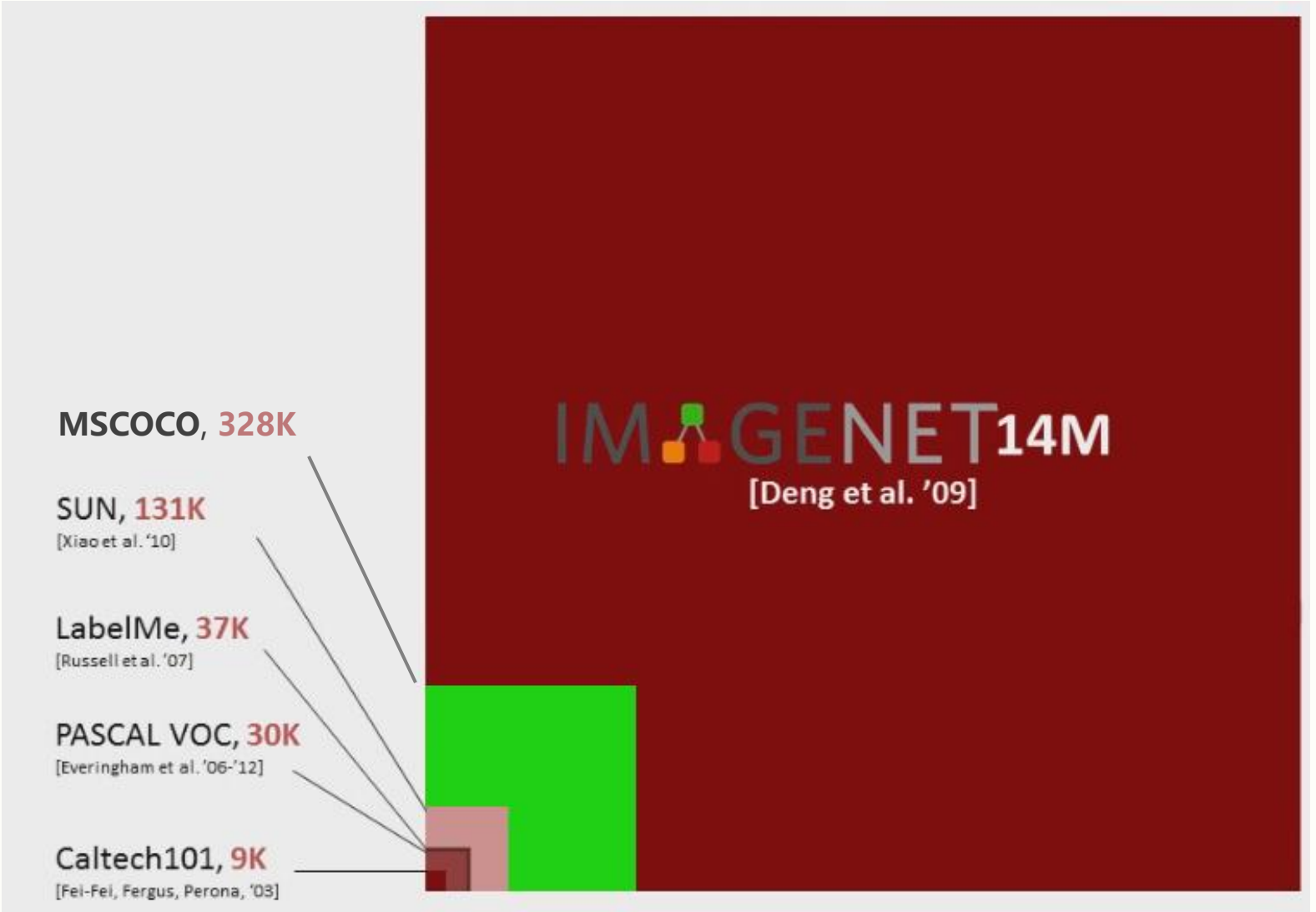
Russakovsky et al. IJCV 2015

# 2.3 人类在视觉方面的认识和努力



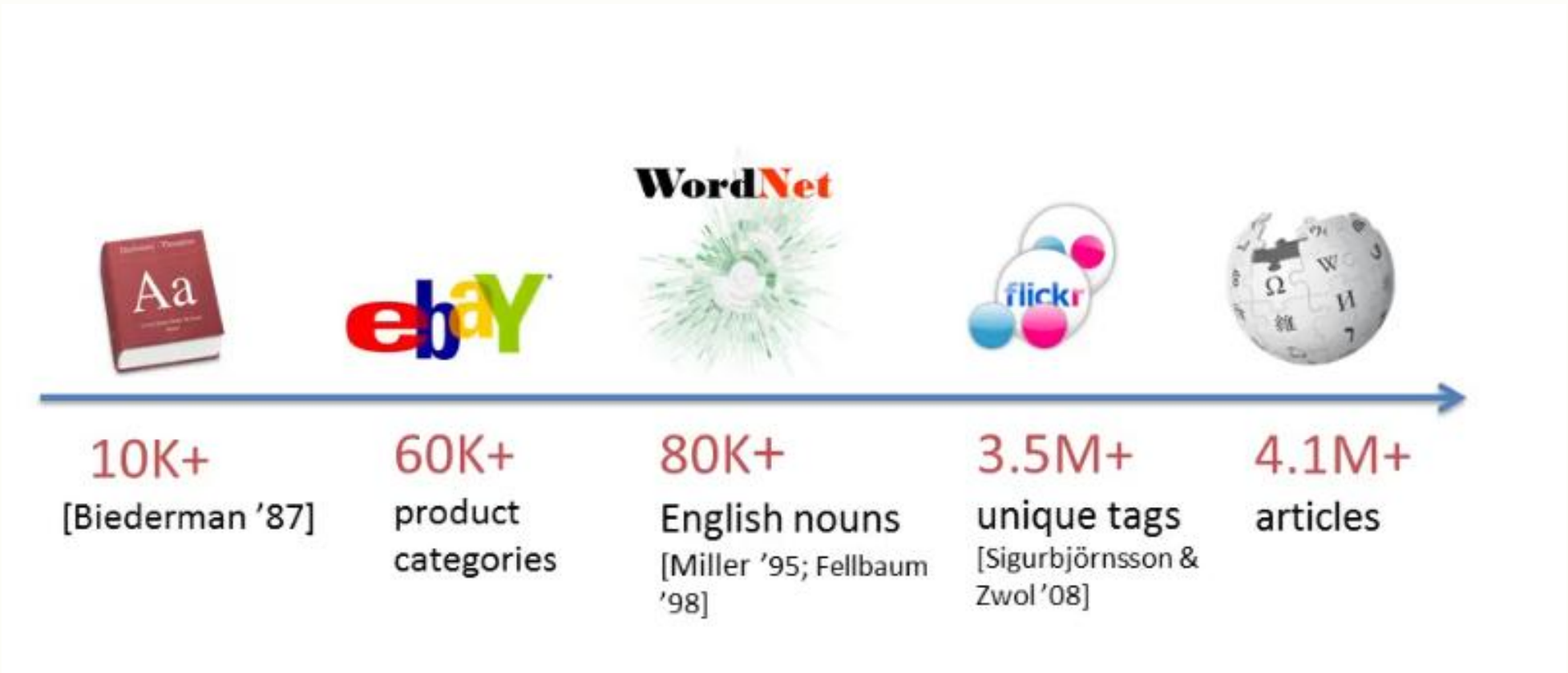
# 2.3 人类在视觉方面的认识和努力

## 二十世纪的前20年：常见标注图像数据集的样本数量对比



# 2.3 人类在视觉方面的认识和努力

## 其他常见的数据集



# 2.3 人类在视觉方面的认识和努力

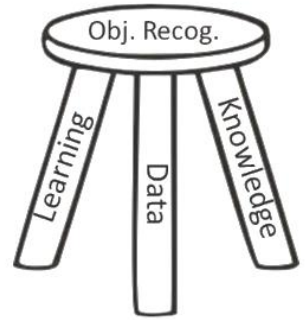
**大数据之后，我们需要什么？**

**更有用的知识**

**更好的学习算法、更强大的运算设备GPU**

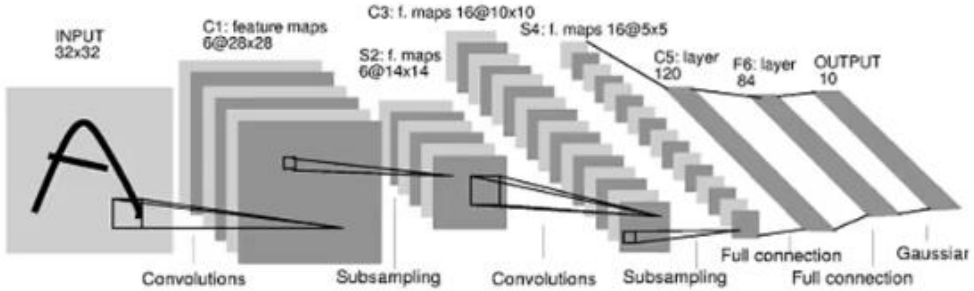
# 2.3 人类在视觉方面的认识和努力

## 近代的飞跃：深度学习时代的到来




1998

LeCun et al.




# of transistors

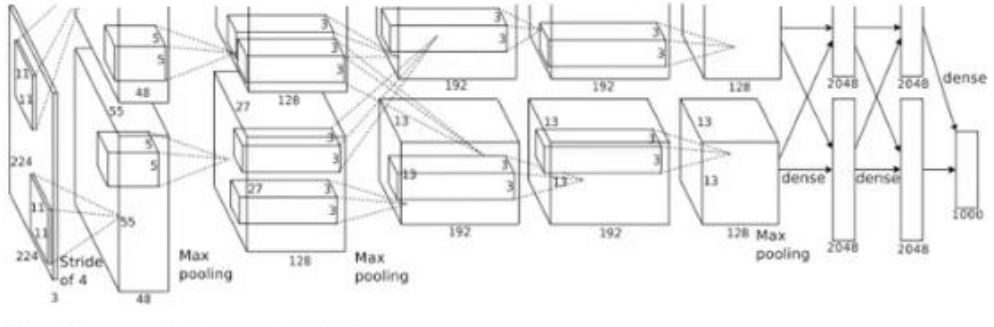

$10^6$

# of pixels used in training

$10^7$ NIST

2012

Krizhevsky et al.




深度
卷积神经网络

GPU

# of transistors GPUs

$10^9$

# of pixels used in training

$10^{14}$ IMAGENET

大数据

# 2.3 人类在视觉方面的认识和努力

## IMAGENET Large Scale Visual Recognition Challenge

### Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

Pooling, SPM

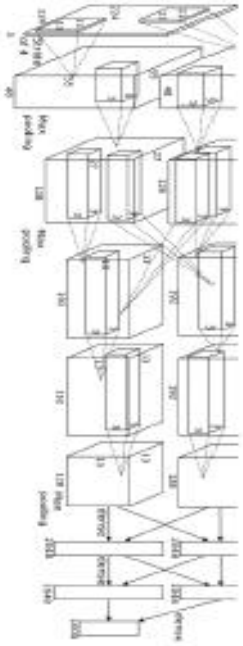
Linear SVM

[Lin CVPR 2011]

Lion image by Swissfrog is licensed under CC BY 3.0

### Year 2012

SuperVision

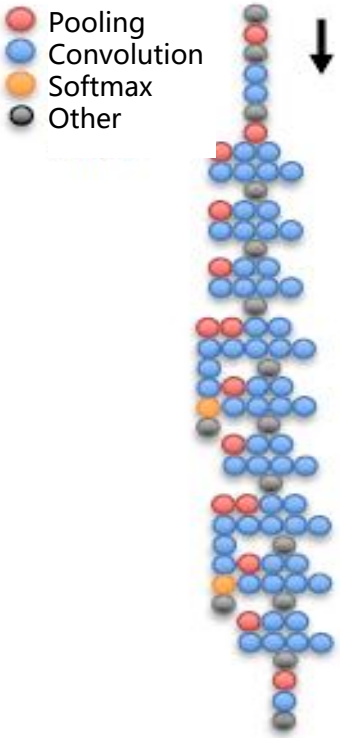


[Krizhevsky NIPS 2012]

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

### Year 2014

GoogLeNet



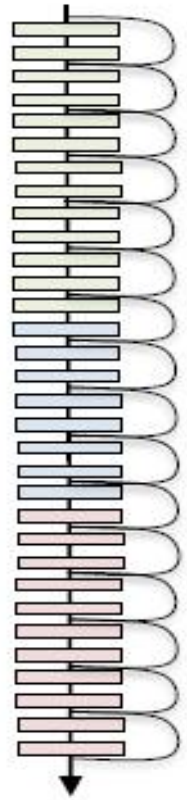

[Szegedy arxiv 2014]

VGG


[Simonyan arxiv 2014]

### Year 2015

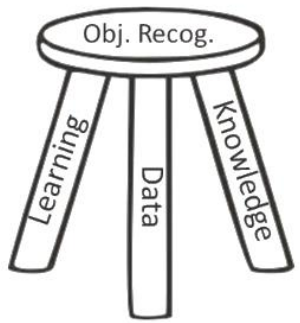
MSRA

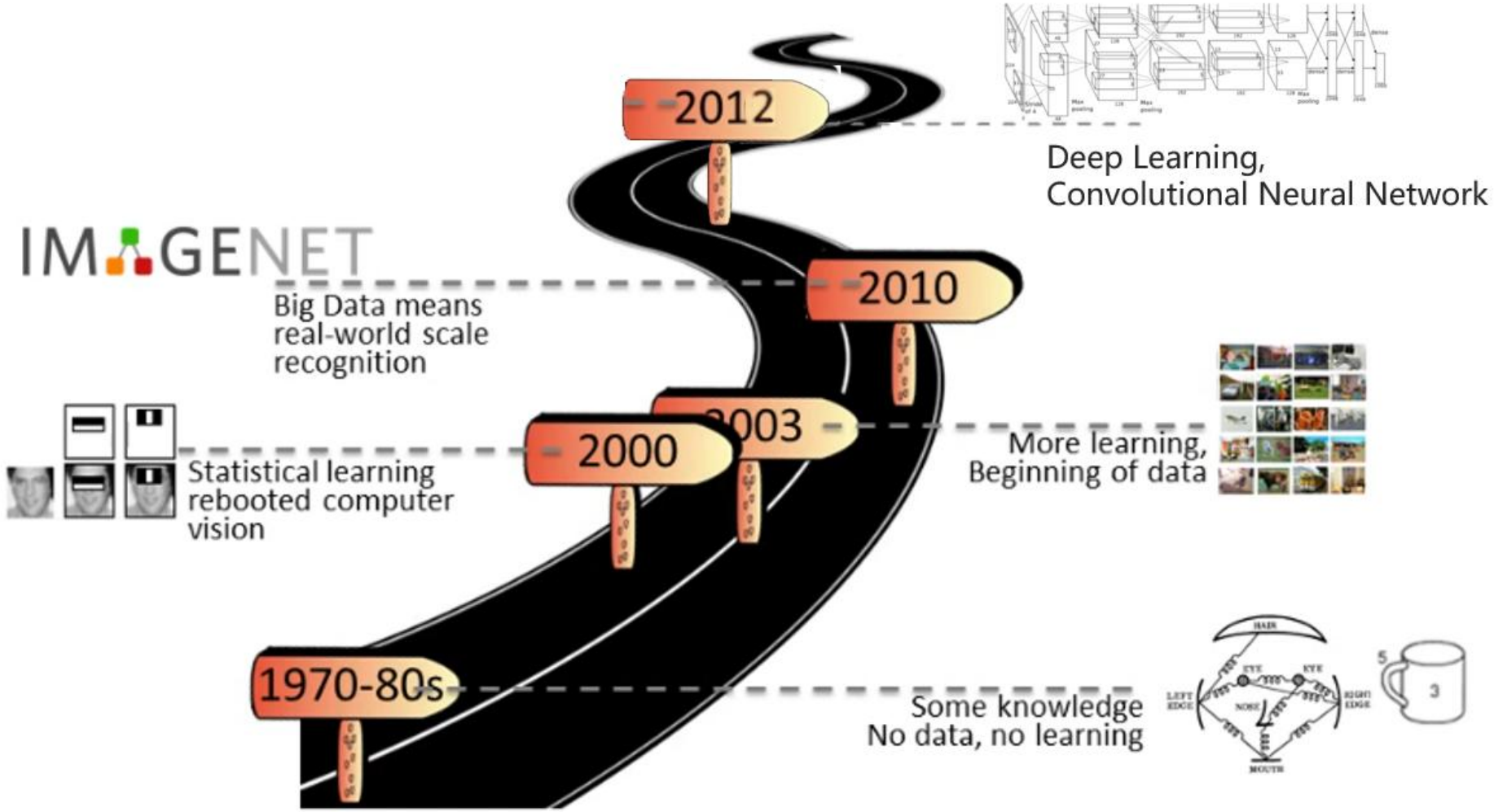


[He ICCV 2015]

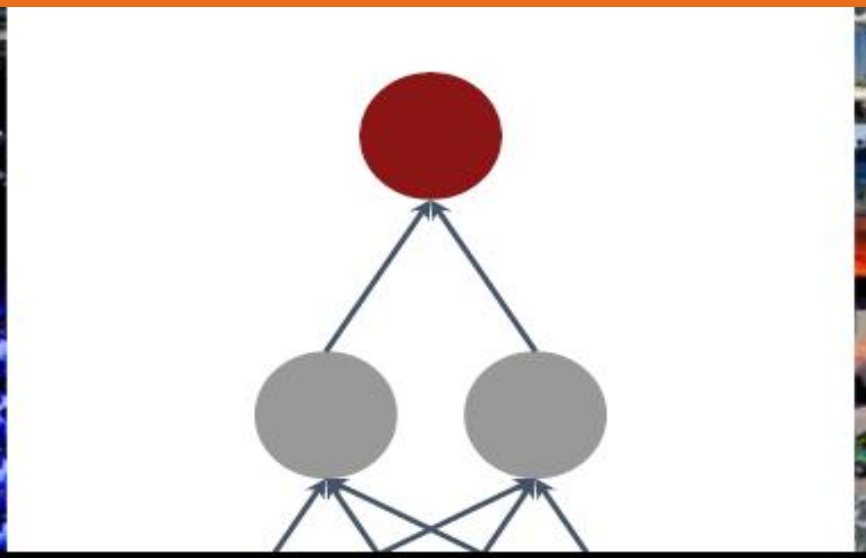
# 2.3 人类在视觉方面的认识和努力

## 对象识别的关键因素：数据、学习和知识

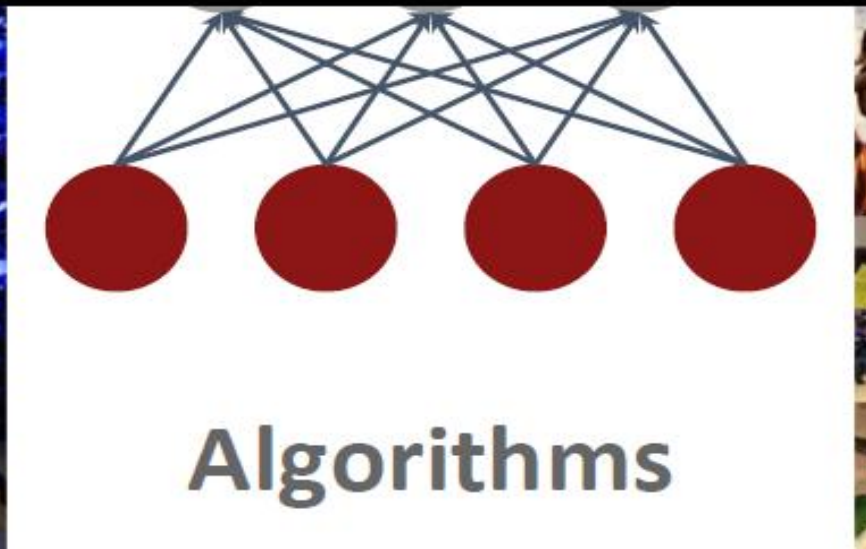


# 2.3 人类在视觉方面的认识和努力

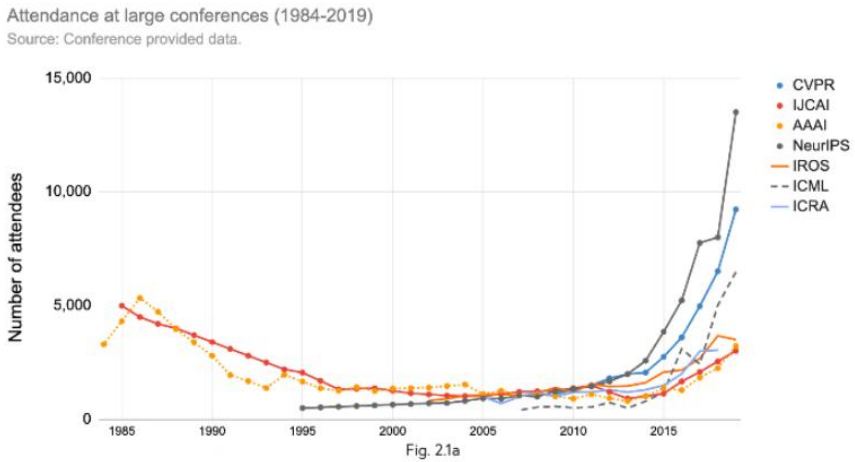


## The Deep Learning Revolution

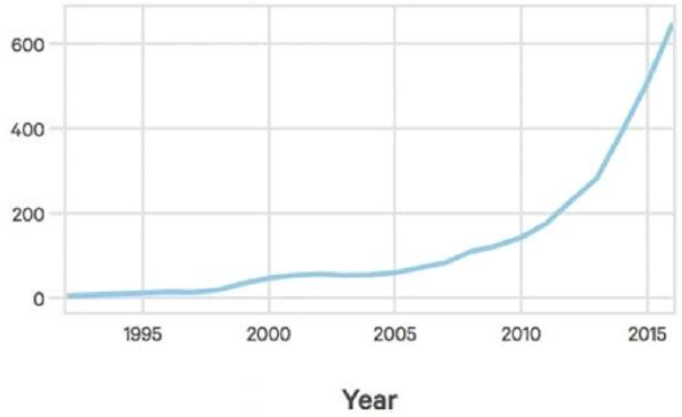

# 2.3 人类在视觉方面的认识和努力

## 人工智能的爆炸性发展和影响



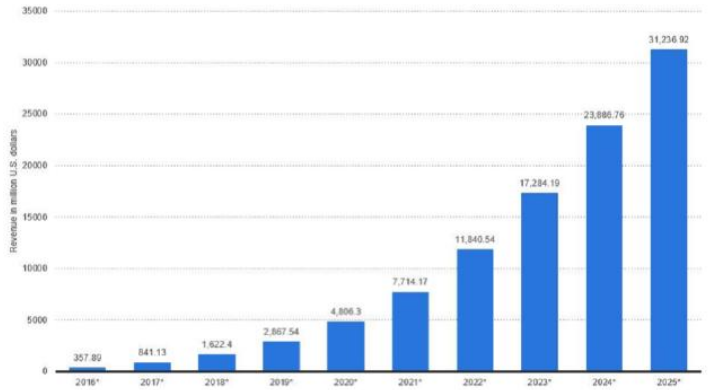

Number of attendance At AI conferences

Source: The Gradient




Startups Developing AI Systems

Source: Crunchbase, VentureSource, Sand




Enterprise Application AI Revenue

Source: Statista

# 2.3 人类在视觉方面的认识和努力

## 图像分割，2D/3D图像生成

This image is CC0 public domain

Progressive GAN, Karras 2018.

Wang et al, "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images", ECCV 2018

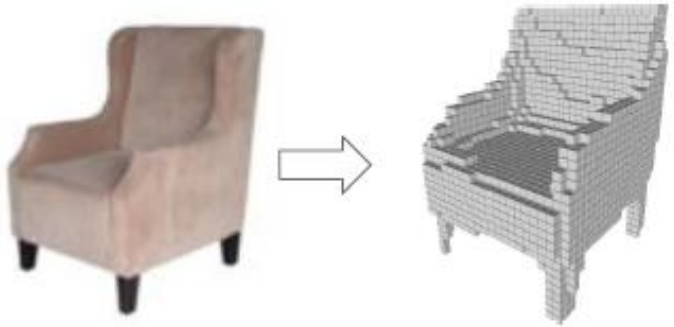
# 2.3 人类在视觉方面的认识和努力

## 时空关系场景图


Ji, Krishna et al., Action Genome: Actions as Composition of Spatio-temporal Scene Graphs, CVPR 2020

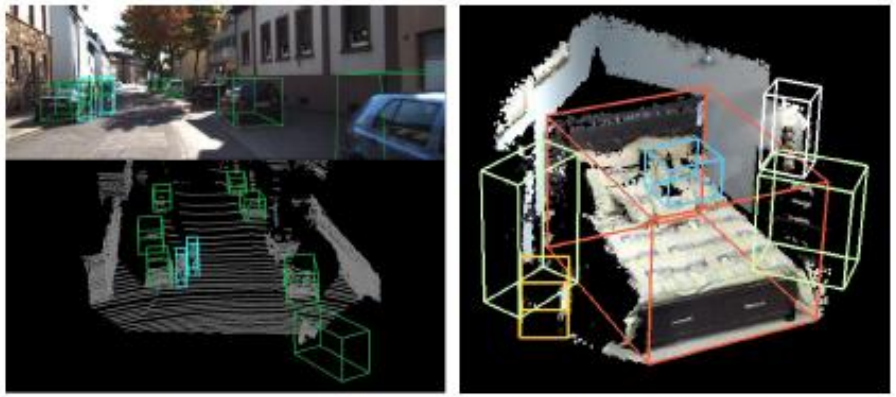
# 2.3 人类在视觉方面的认识和努力

## 3D视觉和机器人视觉

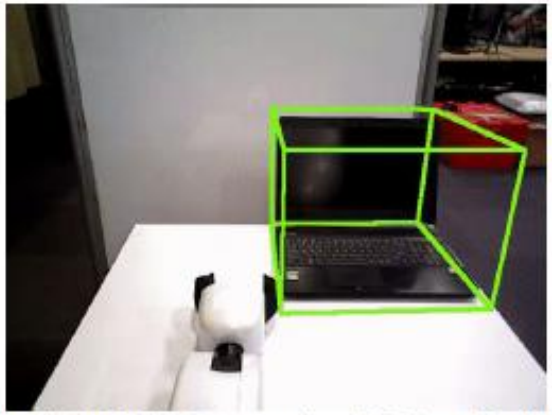


Choy et al., 3D-R2N2: Recurrent Reconstruction Neural Network (2016)


Mandlekar and Xu et al., Learning to Generalize Across Long-Horizon Tasks from Human Demonstrations (2020)



Xu et al., PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation (2018)



Wang et al., 6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints (2020)

# 2.3 人类在视觉方面的认识和努力

## 图像和语句匹配（跨模识别）

**PT = 500ms**



Some kind of game or fight. Two groups of two men? The man on the left is throwing something. Outdoors seemed like because i have an impression of grass and maybe lines on the grass? That would be why I think perhaps a game, rough game though, more like rugby than football because they pairs weren't in pads and helmets, though I did get the impression of similar clothing. maybe some trees? in the background.

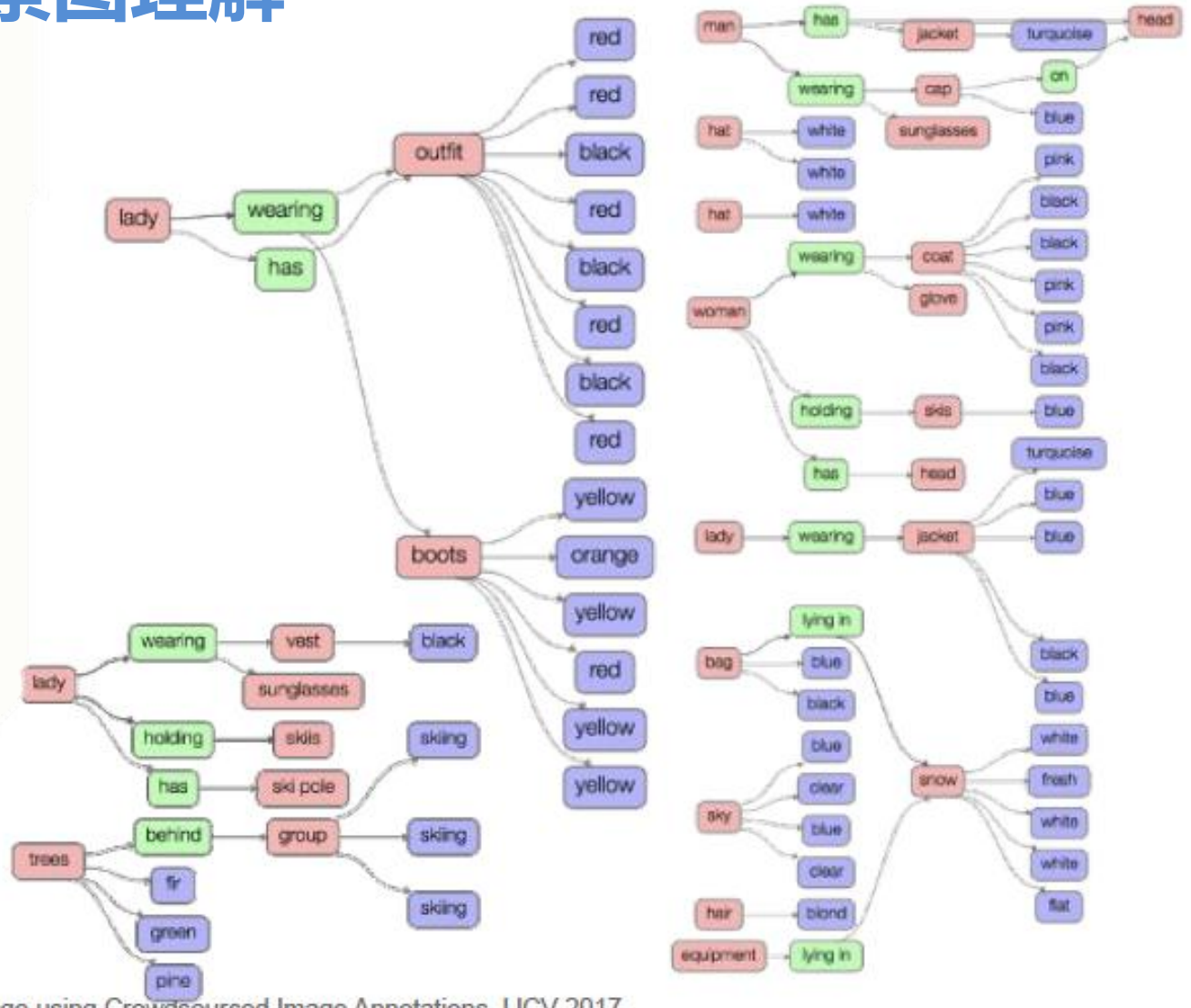
Fei-Fei, Iyer, Koch, Perona, *JoV,* 2007

# 2.3 人类在视觉方面的认识和努力

## 场景图理解

This image is CC0 public domain

[Three Ways Computer Vision Is Transforming Marketing](#)
- Forbes Technology Council




Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

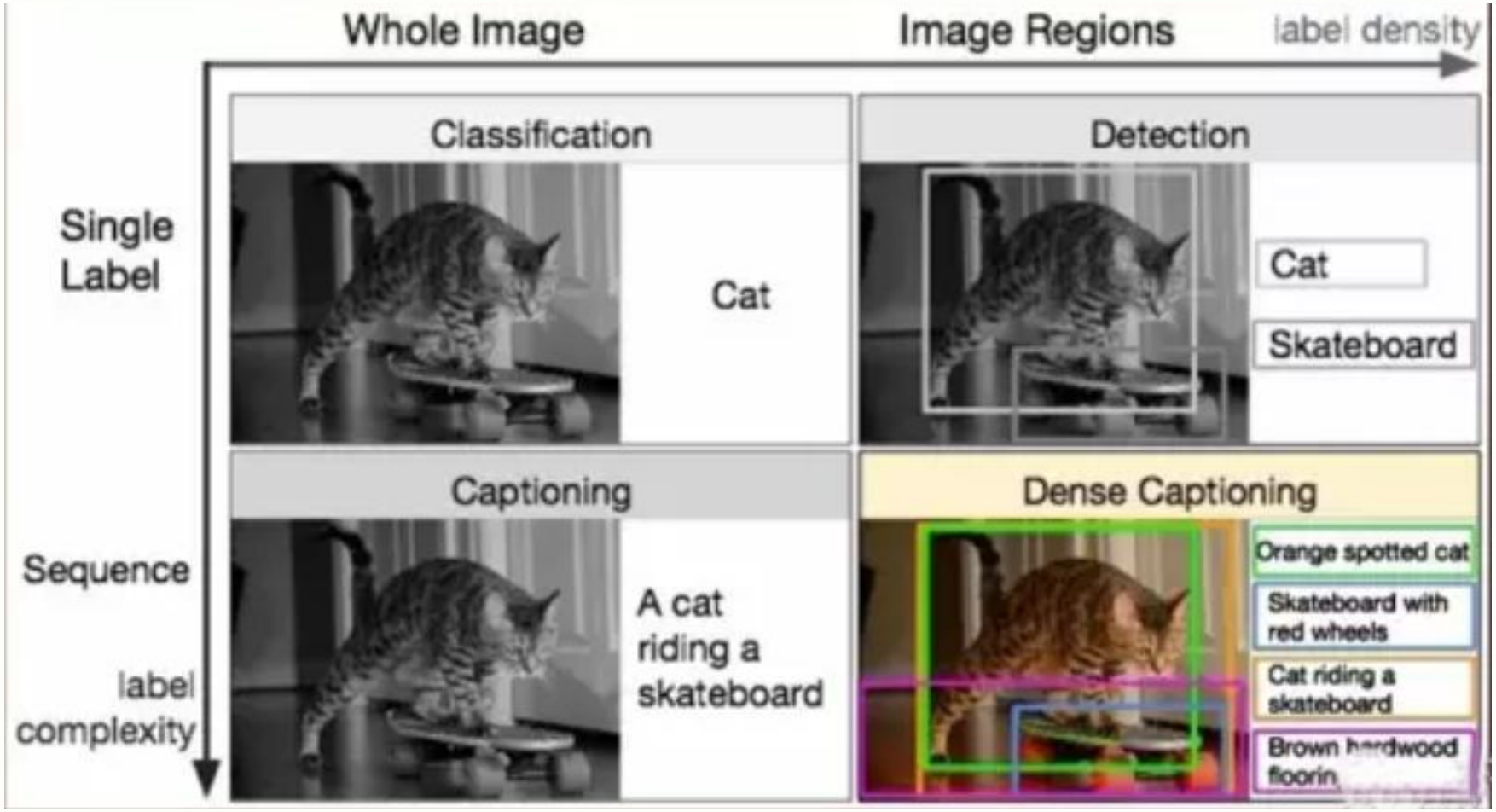
# 2.3 人类在视觉方面的认识和努力

## 场景图理解：视觉问答VQA



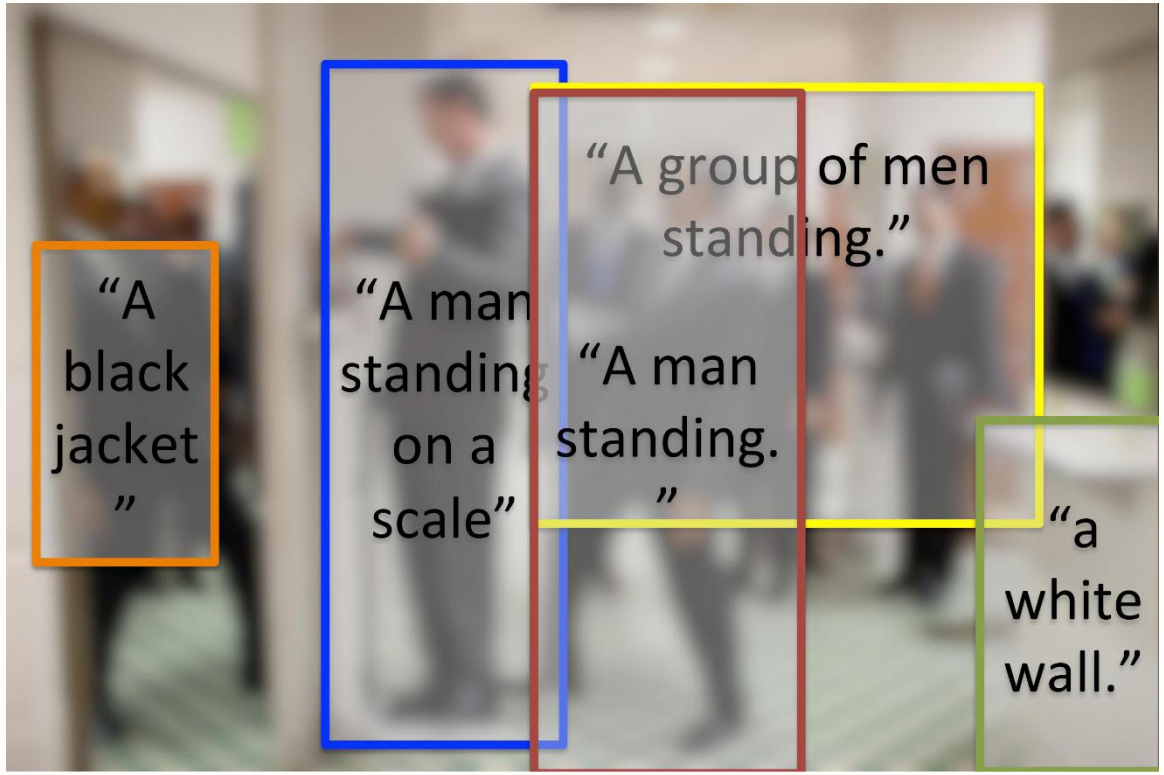
# 2.3 人类在视觉方面的认识和努力

## 从分类到密集图像字幕 (Image Caption)



# 2.3 人类在视觉方面的认识和努力

## 基于区域的密集图像字幕



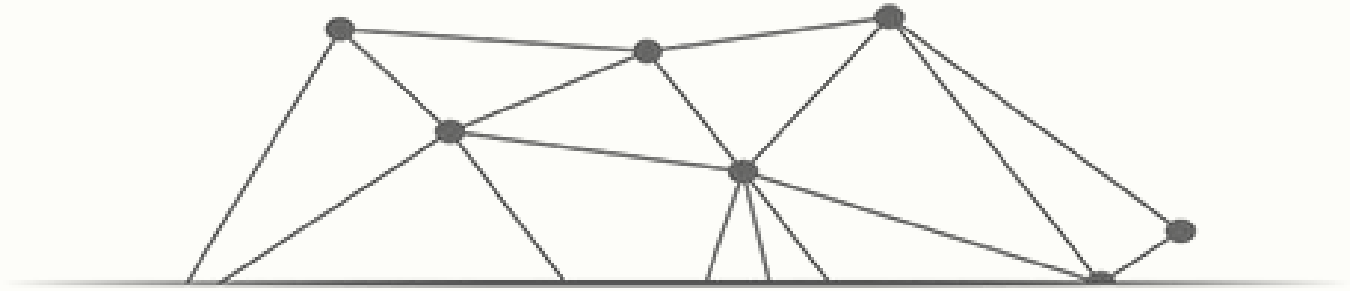

“A group people in a room.”

# 2.3 人类在视觉方面的认识和努力

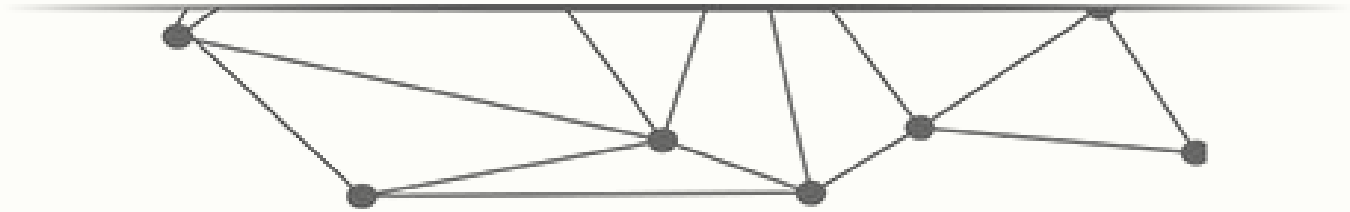
## 按照给定的描述查找区域 (Image Caption)


# 2.3 人类在视觉方面的认识和努力

## 计算机视觉的各种应用



# 课堂互动 13.1.2



Part 03

# 基于深度学习的视频内容理解

/ 无处不在的视频

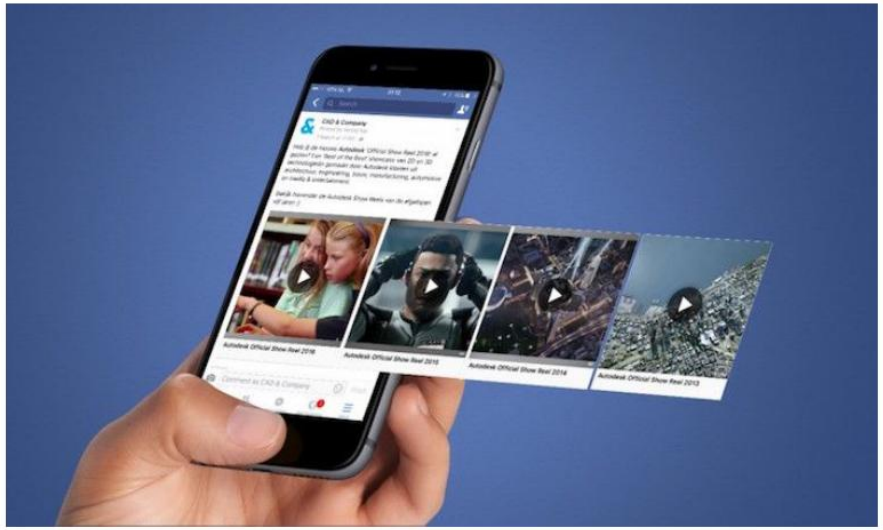
/ 视频内容理解在网络安全领域的应用

/ 视频数据的特点

/ 视频数据发展现状

/ 为什么需要视频数据的智能分析

# 3.1 无处不在的视频

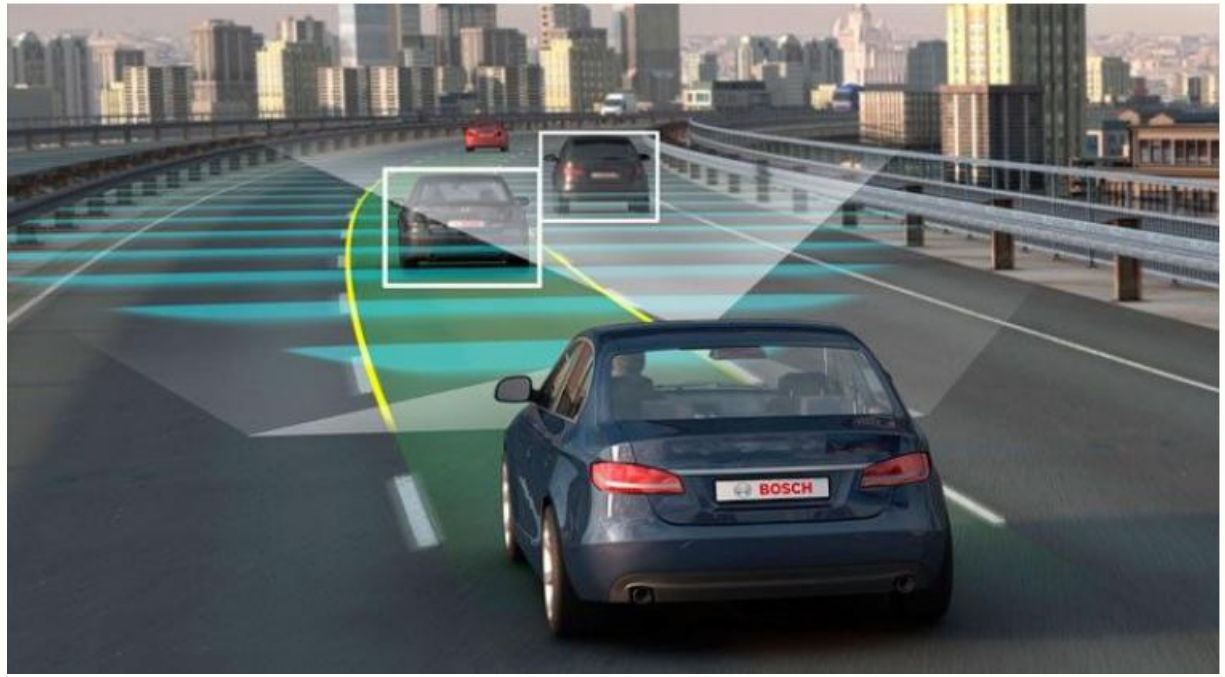




# 3.1 无处不在的视频

## 机器人/工业机械臂

# 3.1 无处不在的视频

## 自动驾驶 和 智能交通监控




# 3.1 无处不在的视频

## 智慧教育和智慧医疗


远程联合会诊

智慧教育

# 3.1 无处不在的视频

## 集体活动理解

# 3.2 视频内容理解在网络安全领域的应用

## 特定人物识别


达赖


习近平


薄熙来

蔡英文

# 3.2 视频内容理解在网络安全领域的应用

## 网络非法信息检测


关键字：六四 游行


关键字：六四 中共


关键字：法轮功


关键字：中共腐败

# 3.2 视频内容理解在网络安全领域的应用

## 网络非法信息检测



**违 禁**
**电视台**


# 3.2 视频内容理解在网络安全领域的应用

作为新时代的大学生，我们有责任和义务利用所学知识，坚决打击一切反党、反国家的言论和个人，为人民创造一片洁净的网络空间。

# 3.3 视频数据的特点

## 数据量大、占比高

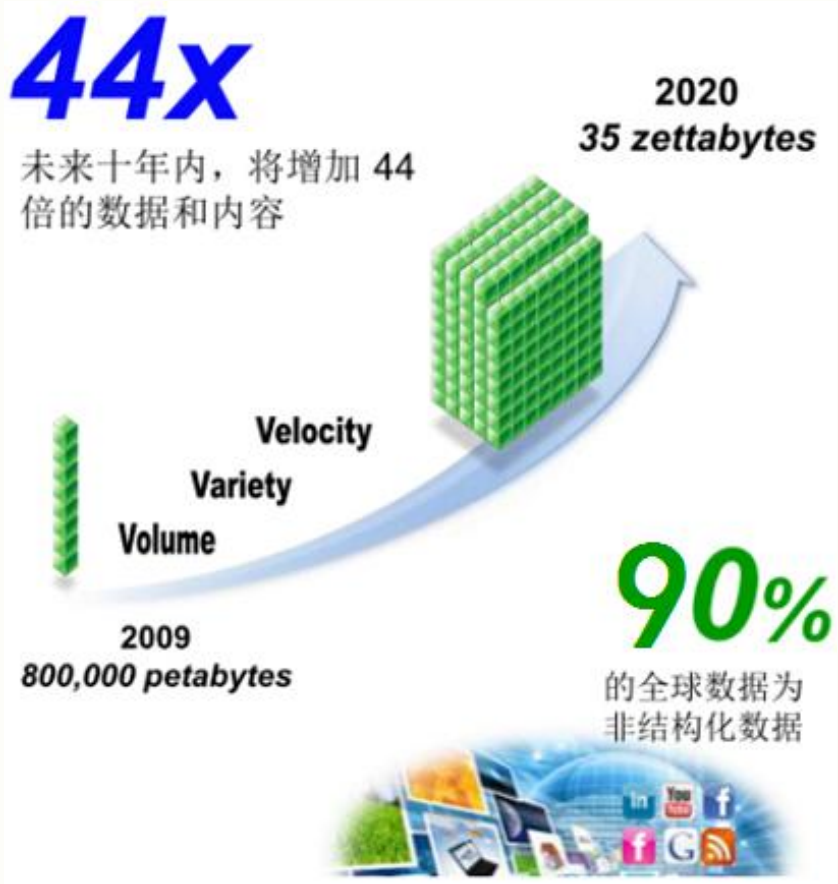
- 每分钟有超过**500个小时**视频上传到 YouTube
- 每天发布视频数量**3400万**，日观看量超过**10亿**。
- 腾讯视频用户总量都**12.56亿**，月活跃用户**8亿**。**日产生视频超过7万小时**
- 全球宽带市场规模421.1亿美元，网络视听收入**5642.81亿**，其中短视频、网络直播**4282.52亿**

数据统计：2024年

# 3.3 视频数据的特点

## 增速快

- 2012年全球互联网视频图像数据达到2.52ZB，2020年将达到31.5ZB
- YouTube 频道数量年比增长 50%
- NETFLIX 原创视频数量平均每年增长185%



# 3.3 视频数据的特点

## 传播快

- 截至2015年10月，就有十支影片达到10亿点阅率。截至2016年12月，已有44支影片超过10亿点阅率
- 采用了P2P技术后，视频下载速度极大提高

| 排名 | 影片名称 | 点阅率（亿）截至2017.1.1 | 上传日期 |
|---|---|---|---|
| 1 | 江南 Style | 27.245 | 2012.7.15 |
| 2 | See You Again | 23.106 | 2015.4.6 |
| 3 | Sorry | 21.275 | 2015.10.22 |

江南style，点阅量超27亿（截止到16年底）

# 3.3 视频数据的特点

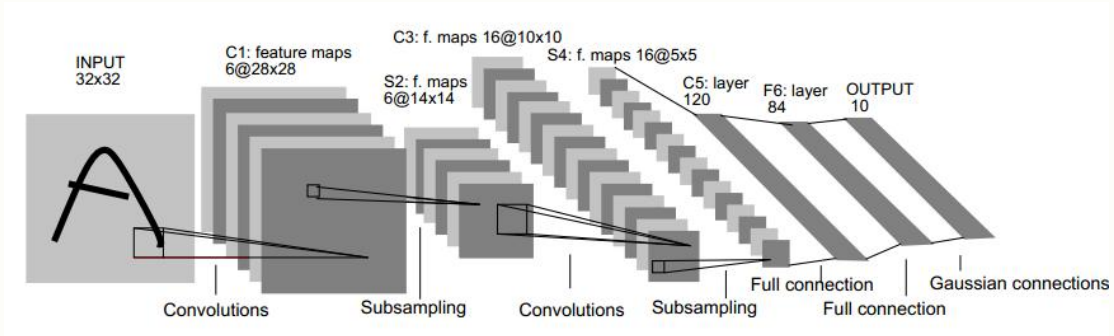
## 实时性、鲁棒性、高度集成

- 高度集成在CPU或者相机或者独立模块
- 实时性和鲁棒性的权衡


信工所S-Lab团队凭借此项工作荣获2016CCF大数据与计算智能大赛综合特等奖

# 3.4 视频技术发展现状

深度神经网络



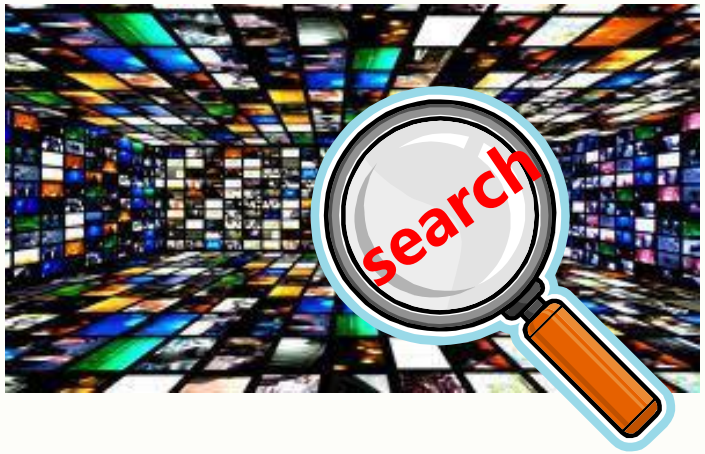

云计算

- 视频技术，在国内外已经有近10年的发展与应用
- 国际上比较著名的专业智能视频分析厂商有VCA Technology、IOImage、ObjectVideo、Bosch、Axis，另外IBM、Sony、松下、PELCO、霍尼韦尔、西门子等公司在该领域也有相当有影响力的整体解决方案产品
- 国内的智能视频分析解决方案厂商主要有海康威视、大华、博康等

# 3.4 视频技术发展现状

## 视频大数据时代-数据爆炸性增长

- 监控摄像头在公共场所中的广泛应用
  - ✓ 摄像机数目每年在以20%加速增长
  - ✓ 2015年中国有3000万摄像头用于安全监控，到2021年已增至5.8亿台




**视频内容分析与计算**

- 结构化分析与表达
- 目标提取与分类
- 目标的快速检索

# 3.4 视频技术发展现状

## 高清时代的到来–清晰度越来越高

# 3.5 为什么需要视频数据的智能分析

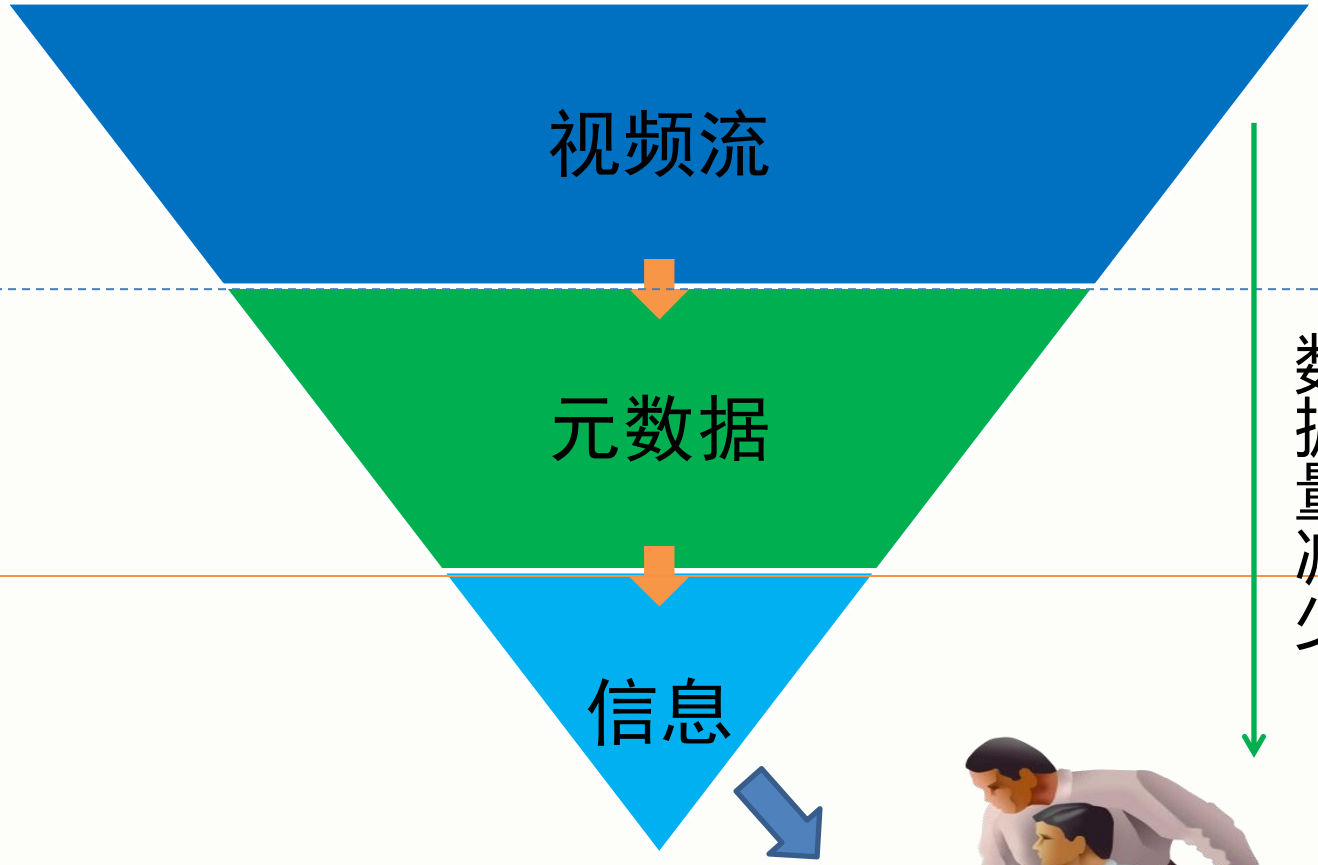
“在传统的闭路电视监控模式下，保安人员需要监视**太多的视频画面，远远超出人的接受能力**，导致实际监控效果低下。

实验结果表明，在盯着视频画面**仅仅22分钟**之后，人眼将对视频画面里**95%以上**的活动信息**视而不见**。因此我们需要智能视觉监控来改善监控效果，同时减轻保安人员的负担。”

—IMS Research

# 3.5 为什么需要视频数据的智能分析

## 从信息角度看：三个阶段




知识

智慧

# 3.5 为什么需要视频数据的智能分析

## 从认知角度看：给监控装上“大脑”

人的视觉系统

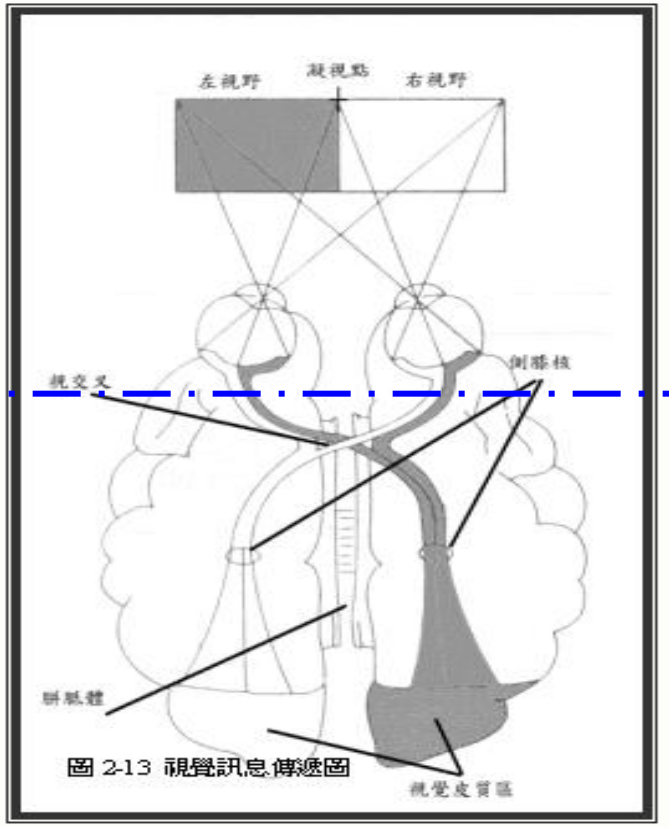
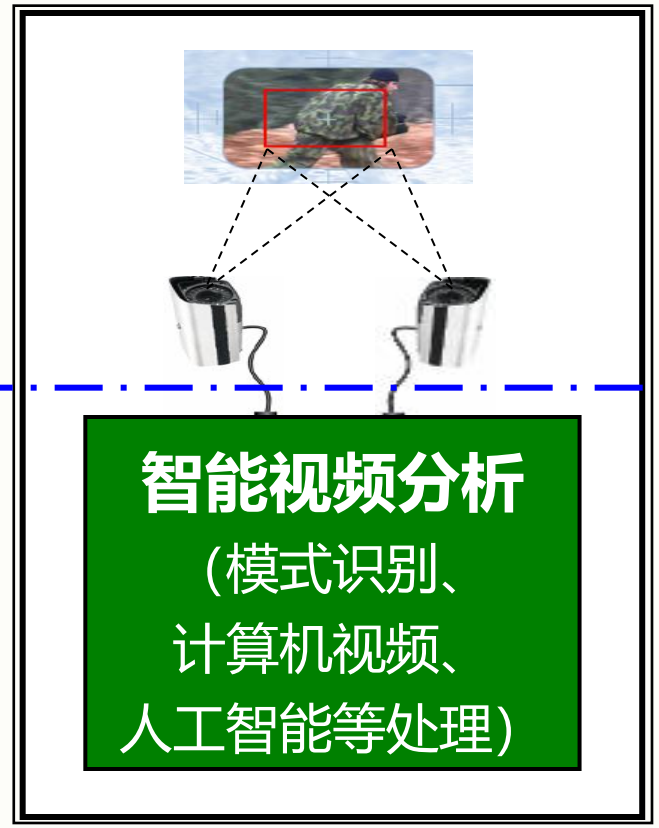


圖 2-13 視覺訊息傳遞圖

智能视频监控




眼睛

**大脑**

摄像机

**智能视频分析单元**

# 3.5 为什么需要视频数据的智能分析

## 从发展角度看：从看清到理解


**看明白**

- **专家系统**
  - 理想目标：图像智能化
  - 让视频更有价值

**看清楚**

- **全天候系统**
  - 次基本需求，提高用户视觉体验
  - 实现24小时应用、全天候应用
  - 作为看明白的基础

**看得见**

- **基本系统**
  - 最基本需求，解决传输和存储的难题
  - 实时处理
  - 事后取证

# 3.5 为什么需要视频数据的智能分析

## 从视频分析技术看：从规则式到大数据学

大数据时代

- 数据挖掘
- 深度学习

数据驱动模式，即数据为王的时代，正符合视频大数据分析和挖掘的时代潮流

机器学习
模式识别

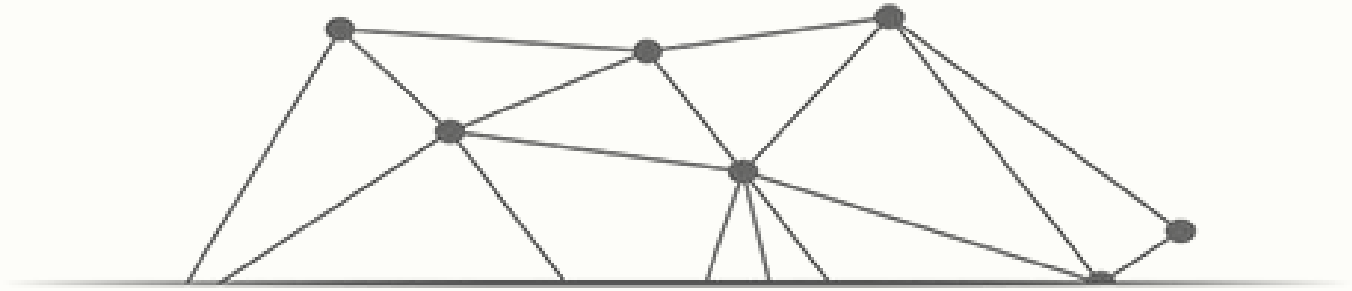
- 目标检测
- 车牌识别
- 人脸识别
- 文本识别
- 视频检索

标注样本，归一化，特征选择，然后分类，问题是要求数据严格，标注严格，参数与阈值较多，算法受应用环境限制。

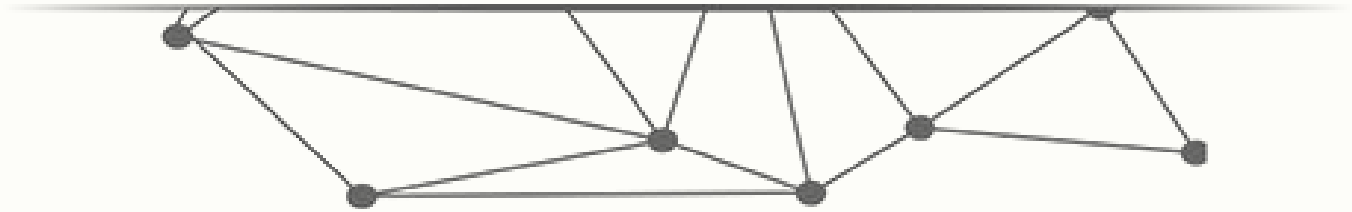
规则式
建模

- 视频压缩
- 背景建模
- 目标跟踪
- 轨迹判断
- 视频摘要

规则式主要靠模板匹配模型，模板的好坏直接影响性能，比如omega头肩模型，混合高斯模型等.



# 课堂互动 13.1.3



# 3.5 视频理解的难点与挑战

- **昂贵的计算代价**

  视频的尺度 >>图像数据集

- **数据集质量较低**

  分辨率不足，动态模糊，遮挡

- **需要大量的训练数据，但是通常不足**

# 3.6 常见的视频处理方法

## 传统模型

- **特征：**
- 局部特征：HoG（方向梯度直方图）+ HOF（光流直方图）
- 运动轨迹：
  - 行为边界直方图 (MBH)
  - 密集轨迹(dese trajectories)：性能较好，但计算复杂性高

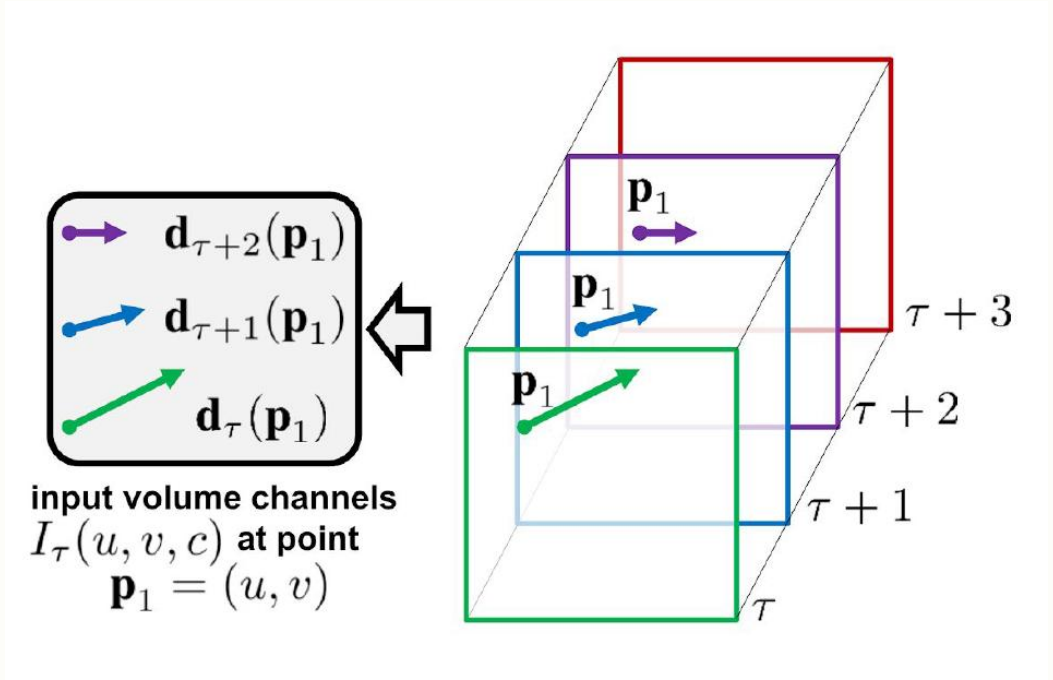
- **特征融合方法:**
- 视觉词袋模型 (Bag of Visual Words) （Ref）
- 费舍尔向量(Fisher verctors) （Ref）

# 3.6 常见的视频处理方法

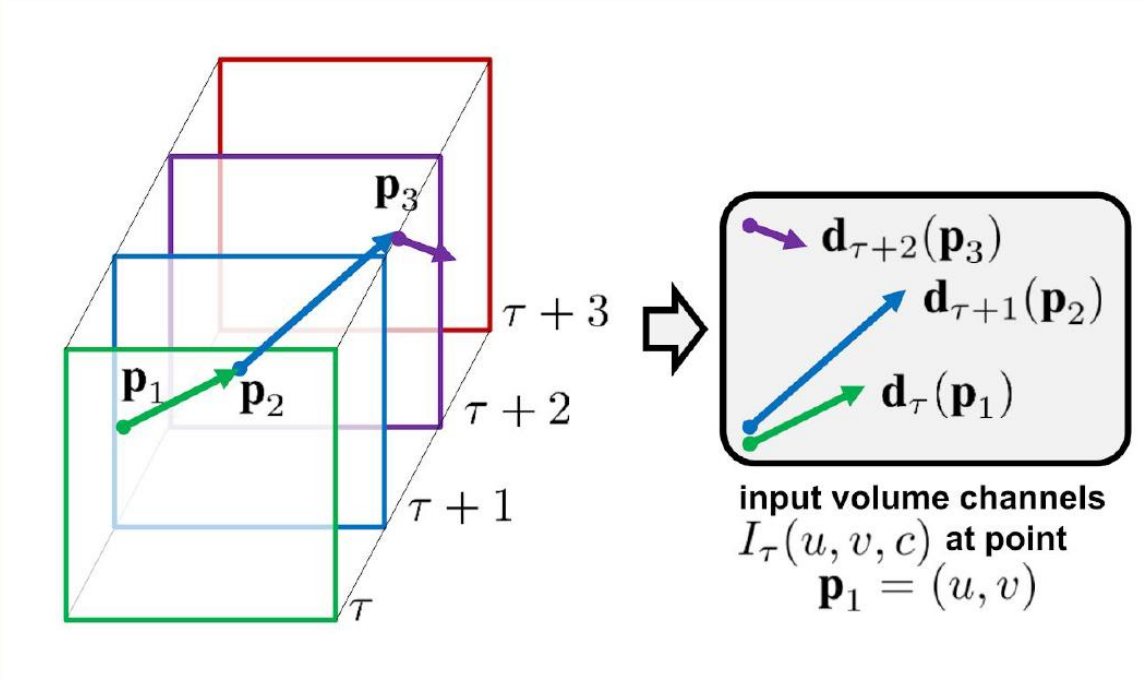
## 传统模型

- **运动表达：光流法vs轨迹跟踪**

1. 光流




2. 轨迹跟踪

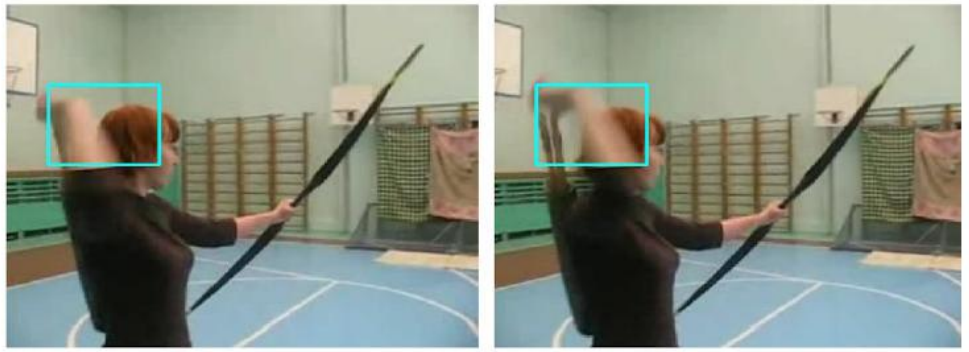


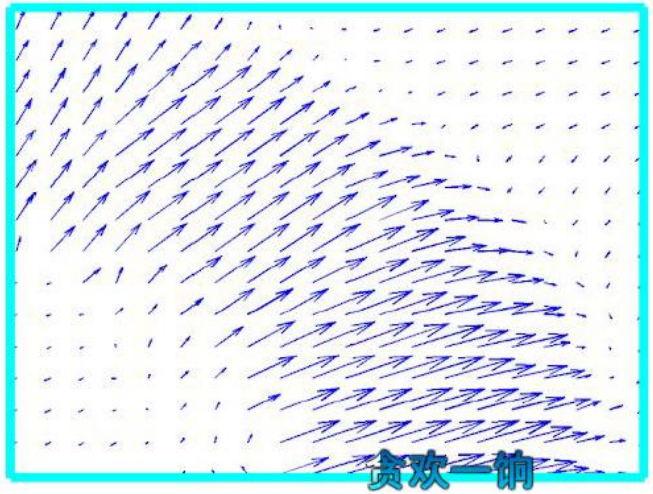

Simonyan and Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

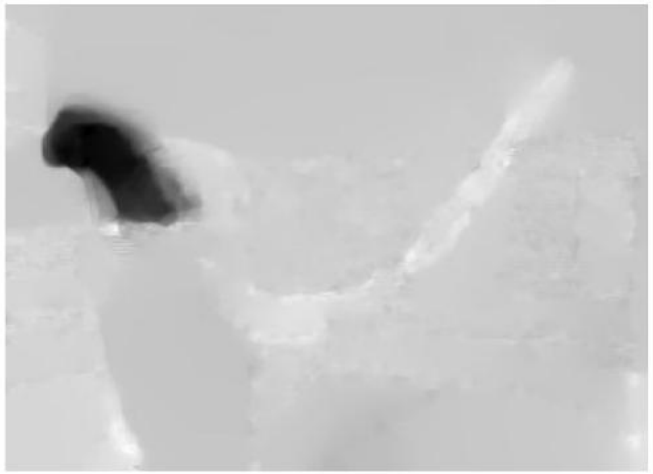
# 3.6 常见的视频处理方法

## 传统模型

- 运动表达：光流法





# 3.6 常见的视频处理方法

## 基于深度学习的方法

## Large-scale Video Classification with Convolutional Neural Networks

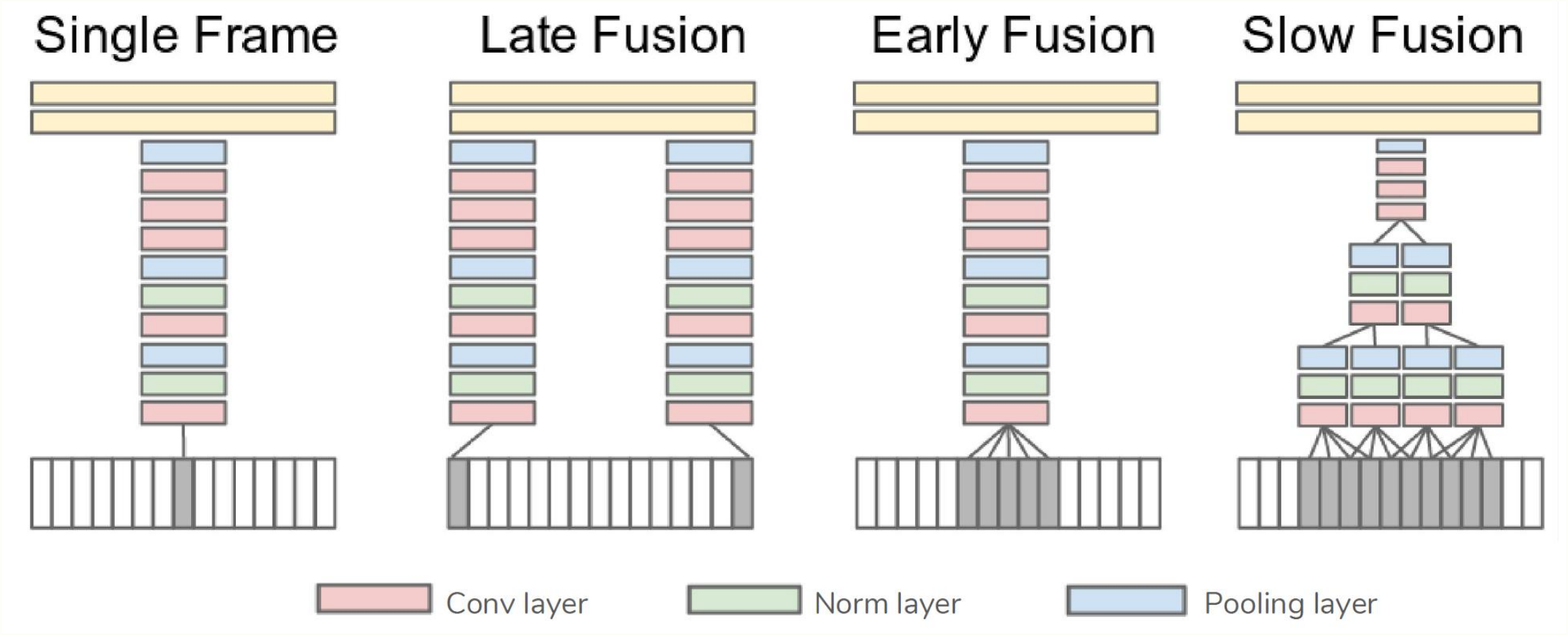
两个问题？

- 从**建模**的角度看：什么样的体系结构能够更好地捕捉时间模式
- 从**计算**的角度看：在不影响精度的前提下如何减少计算开销

# 3.6 常见的视频处理方法

## 基于深度学习的方法

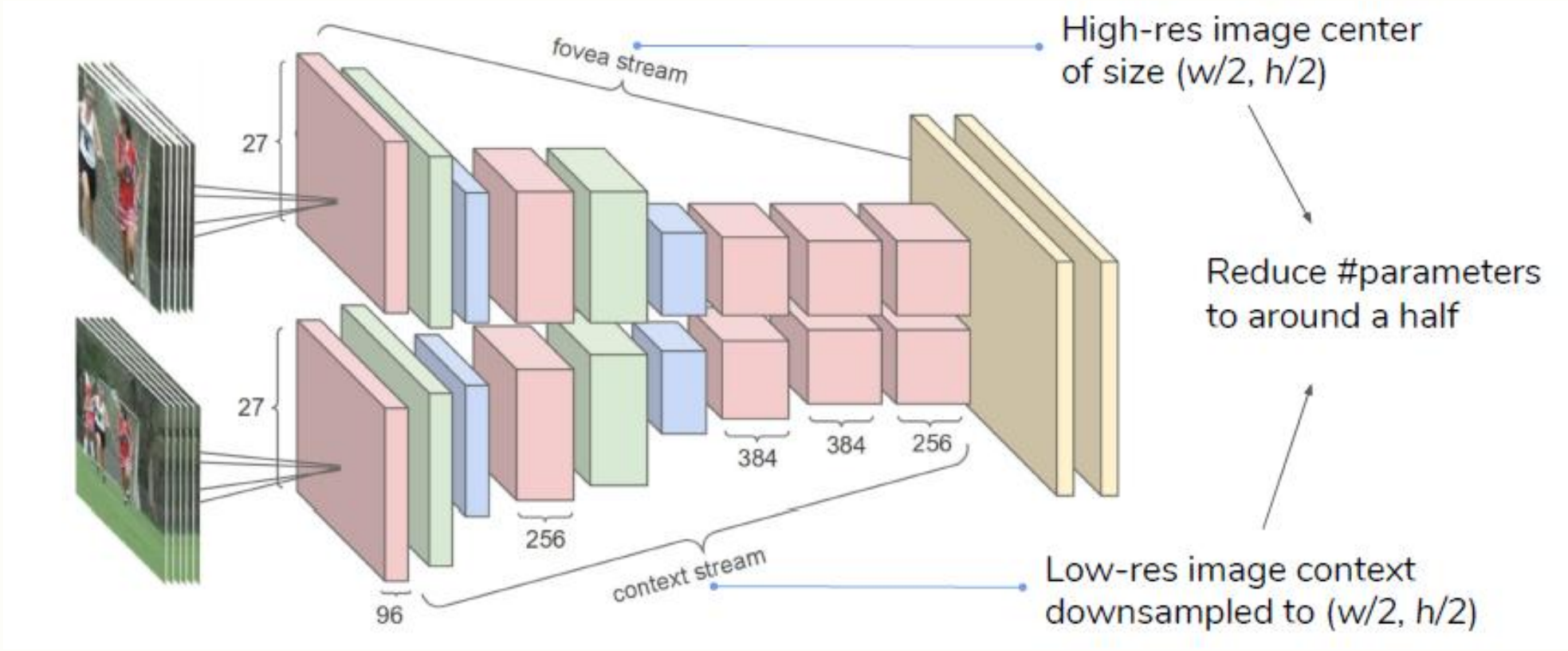
体系结构：以不同的方式融合多个不同的帧



# 3.6 常见的视频处理方法

## 基于深度学习的方法

计算开销：通过降低空间维度减少空间复杂性
-> 多分辨率：低分辨率流 + 高分辨率流



# 3.6 常见的视频处理方法

## 基于深度学习的方法

- **CNN+RNN: Videos as Sequences**

**口 前期工作：** 多帧特征是局部时间

**口 假设：** **全局描述**是有益的

**口 设计选型**

✓ 形态：　1) RGB　2) 光流　3) RGB + 光流

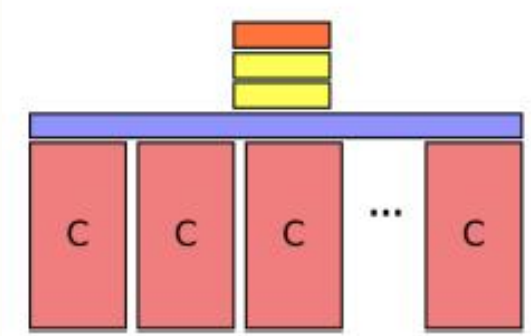
✓ 特征：　1) 手工特征　2) 抽取CNN特征

✓ 时间聚合：　1) 时域池化　2) RNN （e.g. LSTM, GRU）

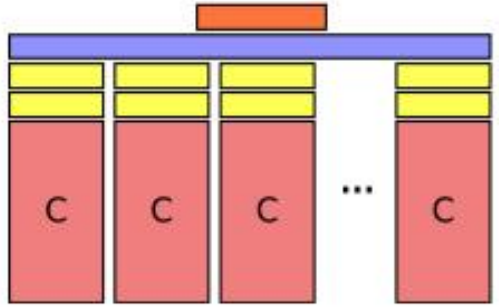
# 3.6 常见的视频处理方法

## 基于深度学习的方法

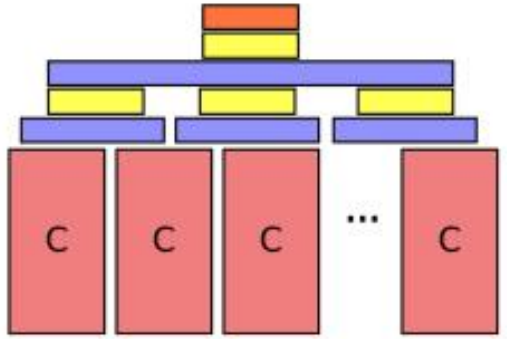
## Beyond Short Snippets: Deep Networks for Video Classification



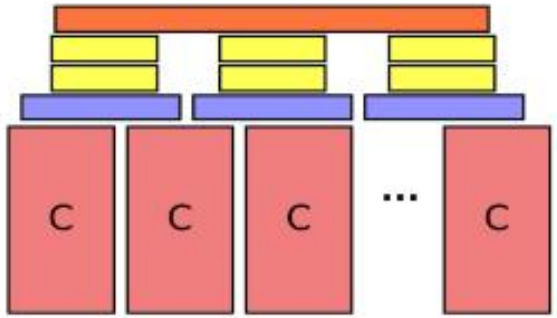

1) Conv Pooling



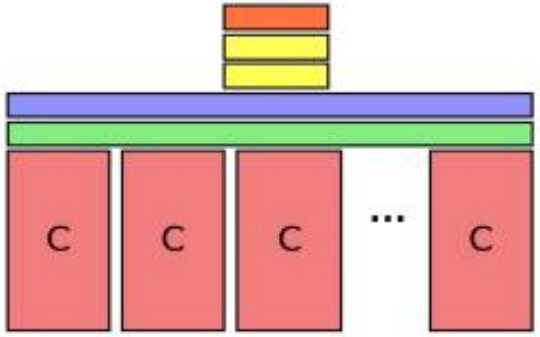

2) Late Pooling




3) Slow Pooling




4) Local Pooling

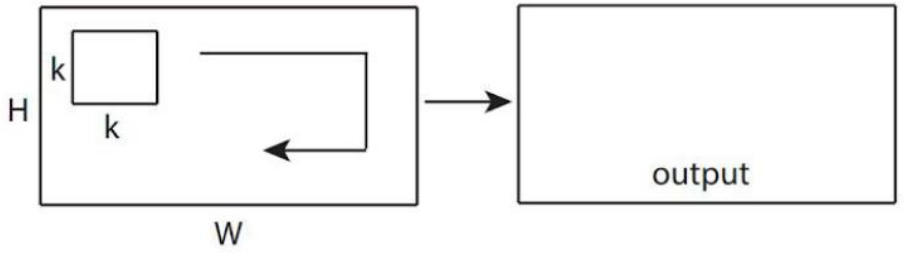



5) Time-domain convolution

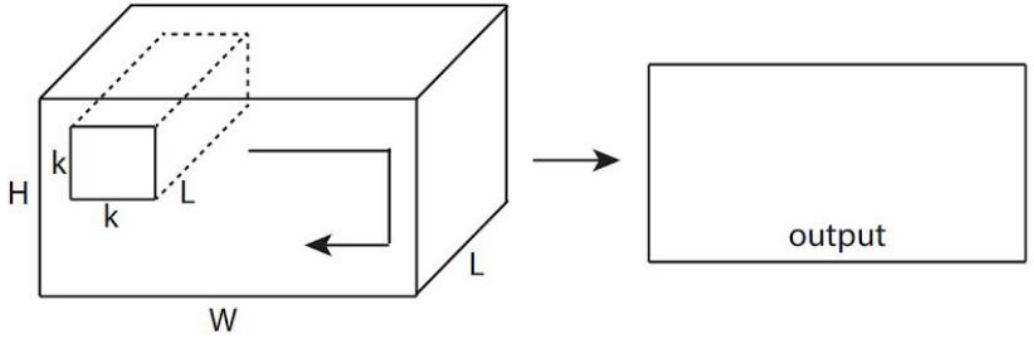
# 3.6 常见的视频处理方法

## 2D 和 3D 卷积

**前期工作**：2D卷积折叠时间

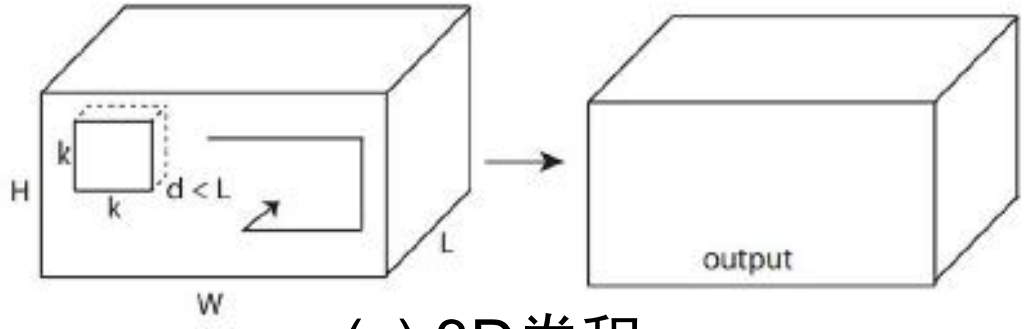



(a) 2D卷积




(b) 多帧上的2D卷积

**建议**：3D卷积 - 》 直接学习编码时间信息的特征

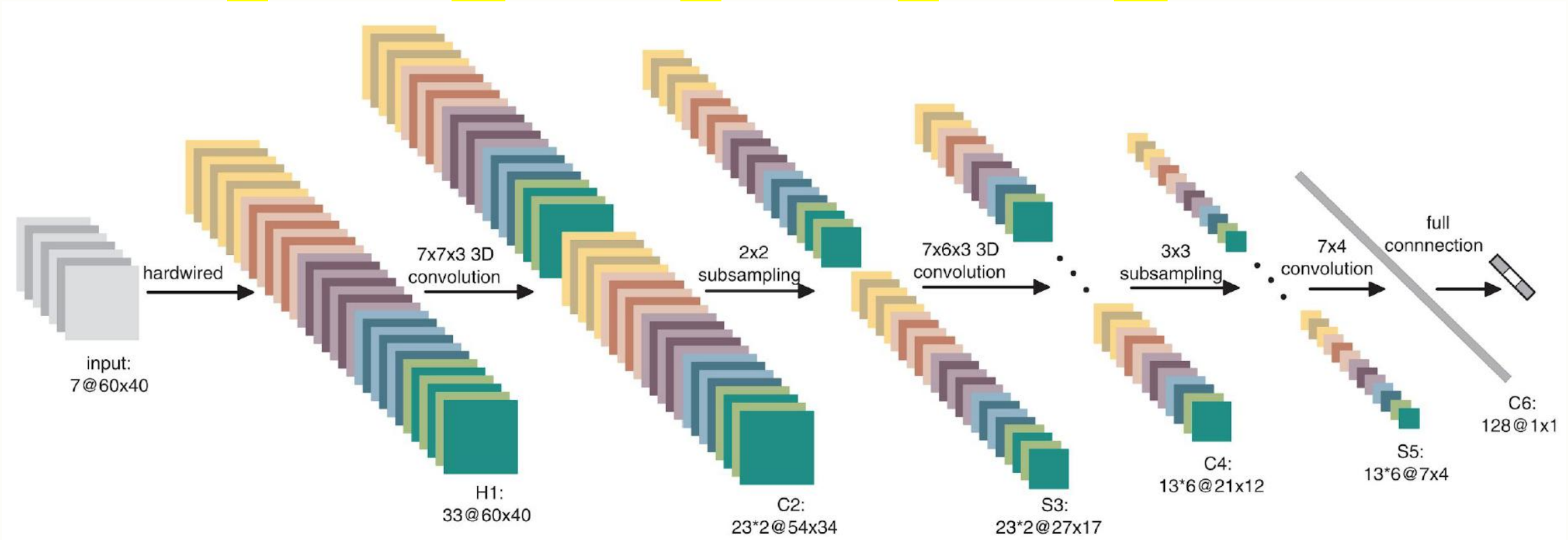



(c) 3D卷积

# 3.6 常见的视频处理方法

## 2D 和 3D 卷积

多通道输入： 1) 灰度图； 2) 梯度 x ； 3) 梯度 y； 4) 光流 x； 5) 光流 y



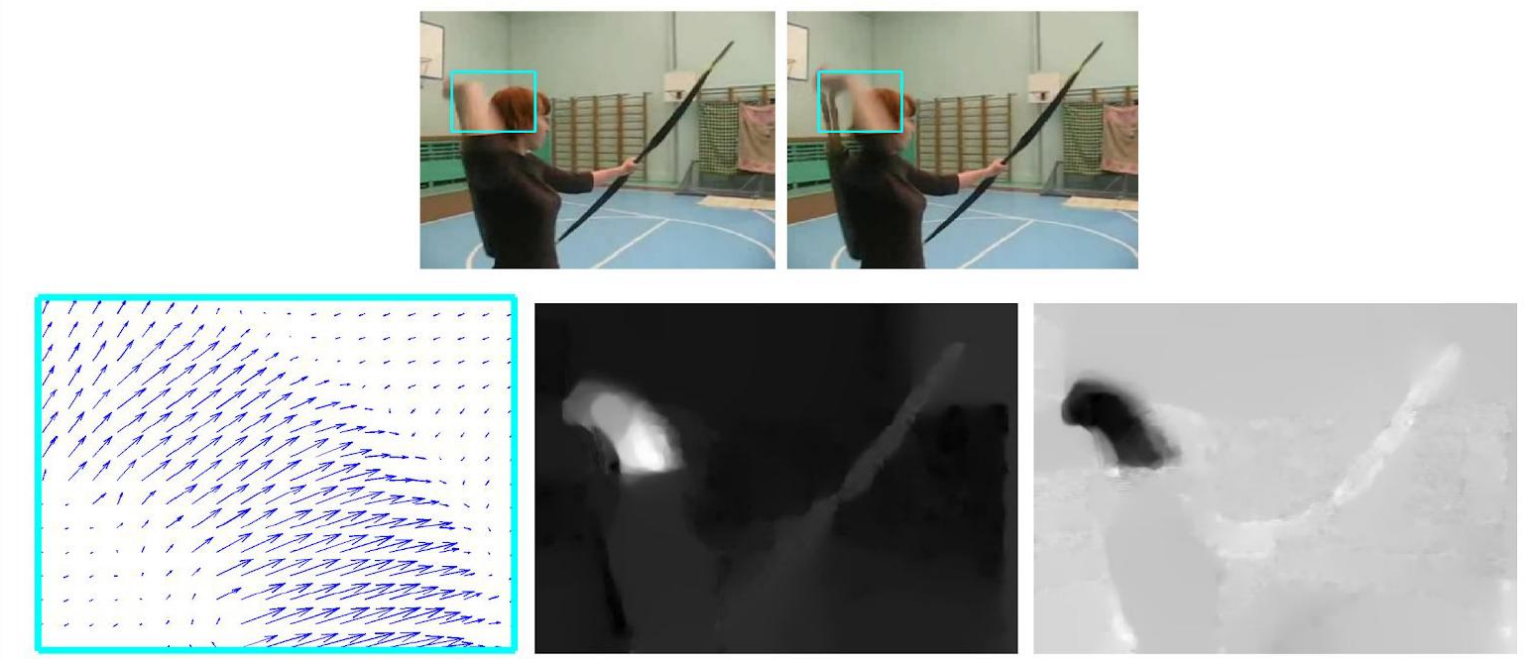
# 3.6 常见的视频处理方法

## 双流模型

## 视频(Video) = 外观(Appearance) + 行为(Motion)

**互补信息**：

- 单帧：静态的外观信息
- 多帧：光流信息（PS: 光流，以像素位移形成的运动信息)

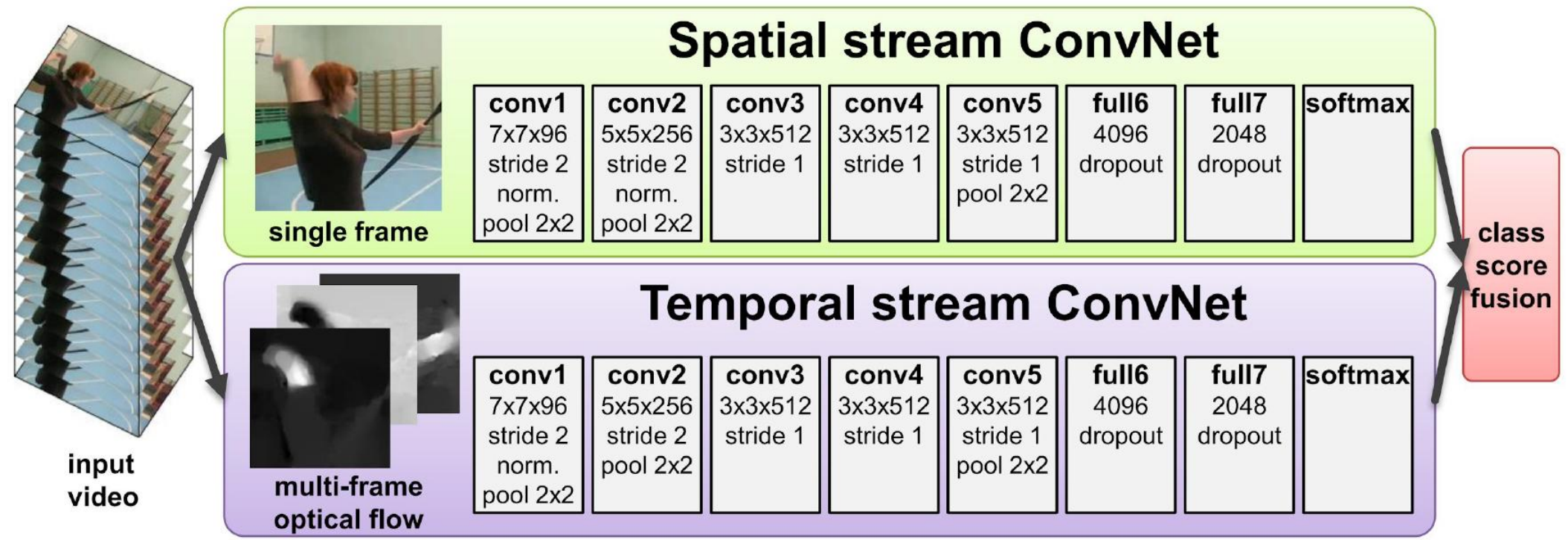


# 3.6 常见的视频处理方法

## 双流模型

## Two-Stream Convolutional Networks for Action Recognition in Videos

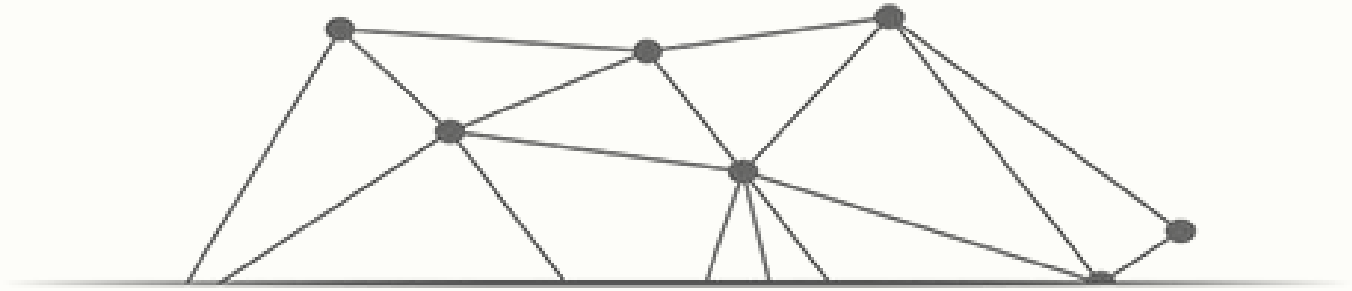
**运动（多帧）与静态外观（单帧）分离**



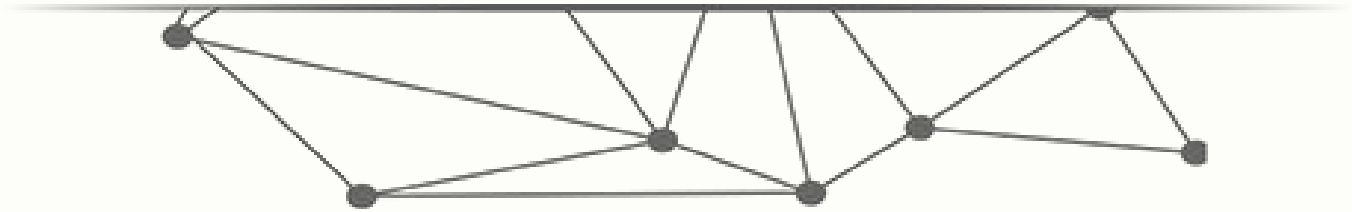
# 3.6 常见的视频处理方法

## 视频理解模型总结

- CNN+RNN：将视频理解看作是一个序列模型
- 3D卷积：在CNN中嵌入空间维度
- 双流模型：运动显式模型



# 课堂互动 13.1.4



Part 04

# 面向海量视频的视觉计算与识别

/ 复杂场景下结构保持的实时视频摘要

/ 基于多特征协同学习的高效目标检测方法

/ 多模态深度编码的目标检索

/ 基于全卷积神经网络的车型分类

/ 基于多部件深度卷积网络的人脸识别

# 4.1 复杂场景下结构保持的实时视频摘要

在安防监控中，每一起案件，调取的监控视频都超过100T，以一部电影500兆来计算，民警所看的视频量相当于20万部电影。



浓缩前视频2小时17分钟，浓缩后视频11分钟，并在浓缩后的视频中标注目标物出现的时间。因此极大的减少了办案人员确定疑犯的时间，提高了工作效率。

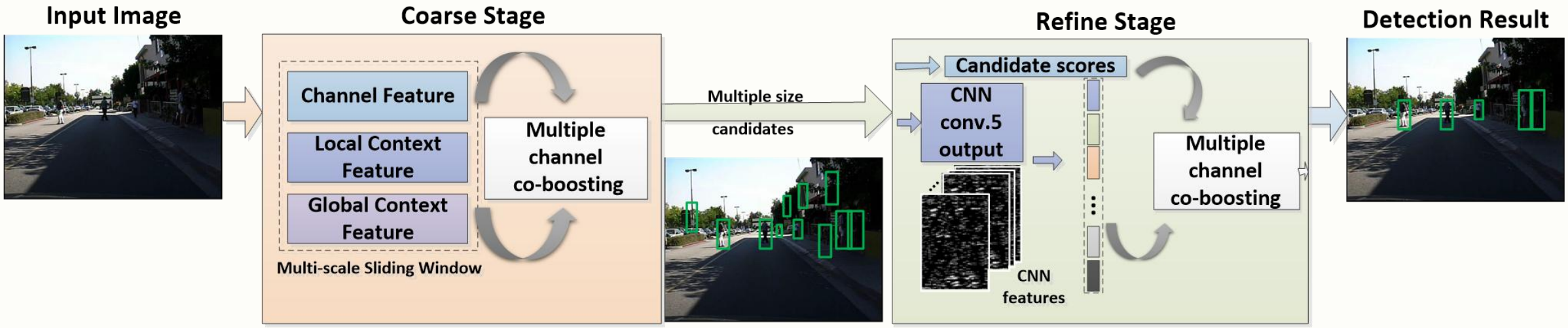
**视频摘要**

- 空间和时间两个维度去除冗余
- 重新组合原始视频中所有目标的时空序列 到一个新的视频中展示

# 4.2 基于多特征协同学习的高效目标检测方法

Boosting+多通道特征----高召回率

Boosting+深度卷积特征-------高准确率



# 4.2 基于多特征协同学习的高效目标检测方法

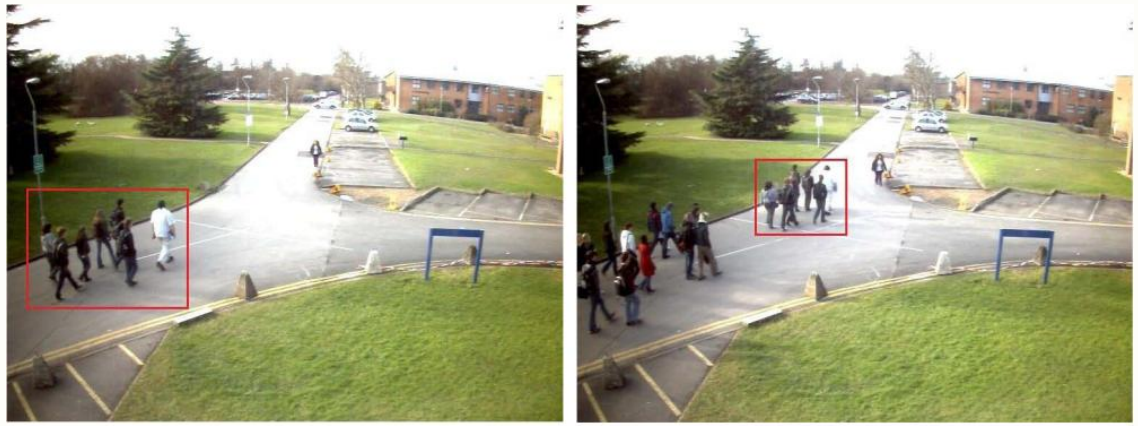
## 行人检测与跟踪

- **基于群组上下文关系的行人跟踪与计数**
  - ✓ 短暂发生的严重遮挡
  - ✓ 目标尺度变化

**整合空间与时间信息**

**群组（群组上下文）**

Frame 140　　Frame 144

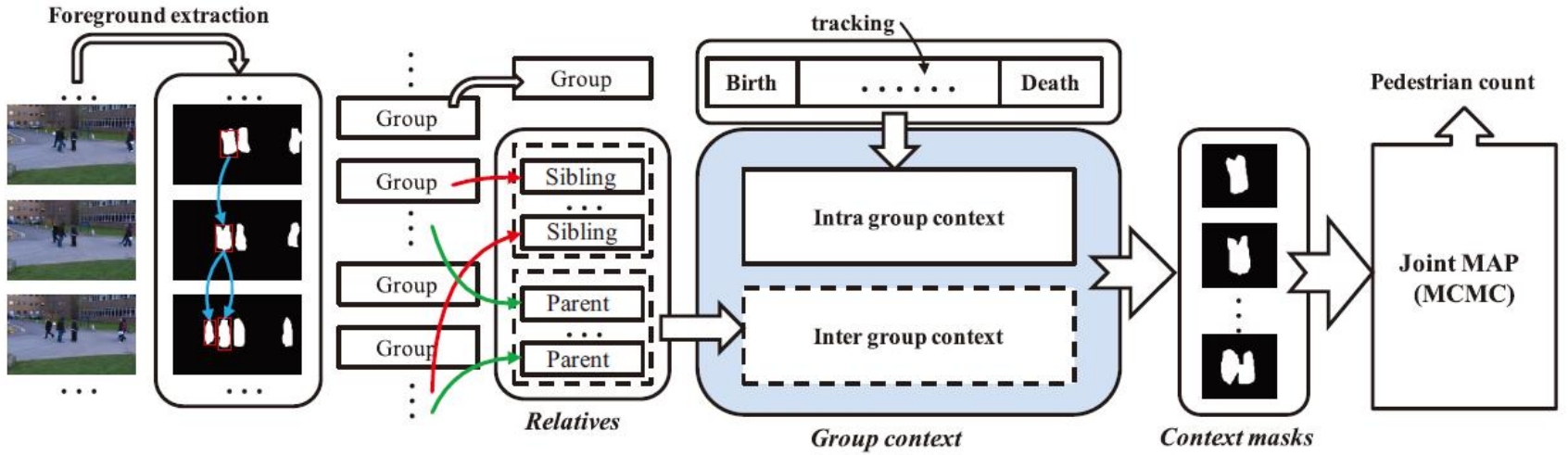


Frame 010　　Frame 042

# 4.2 基于多特征协同学习的高效目标检测方法

## 行人检测与跟踪

- **基于群组上下文关系的行人跟踪与计数**
  - **基本思想：**
    - ✓ 提出了一种基于群组上下文关系的行人计数方法。
    - ✓ 通过群组相关性矩阵来生成群组上下文，准确刻画了群组之间和群组内部时空关系，有效提高了计数精度。

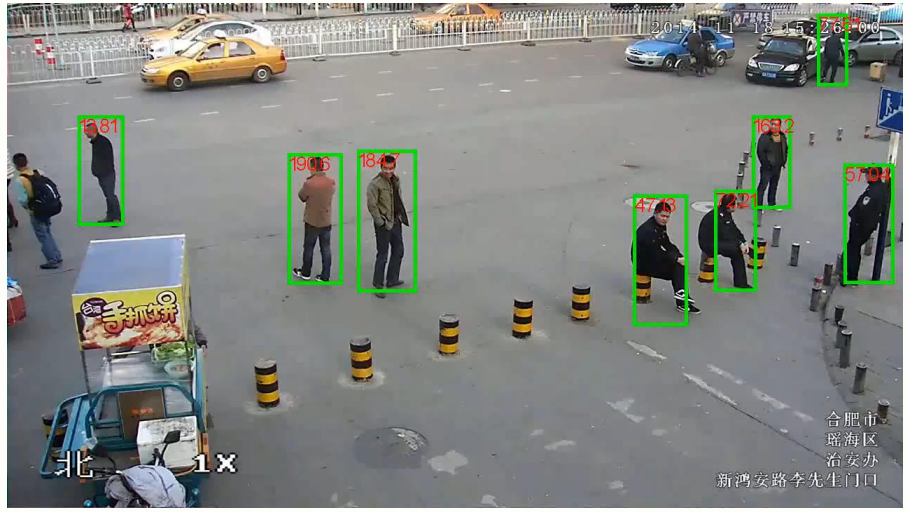



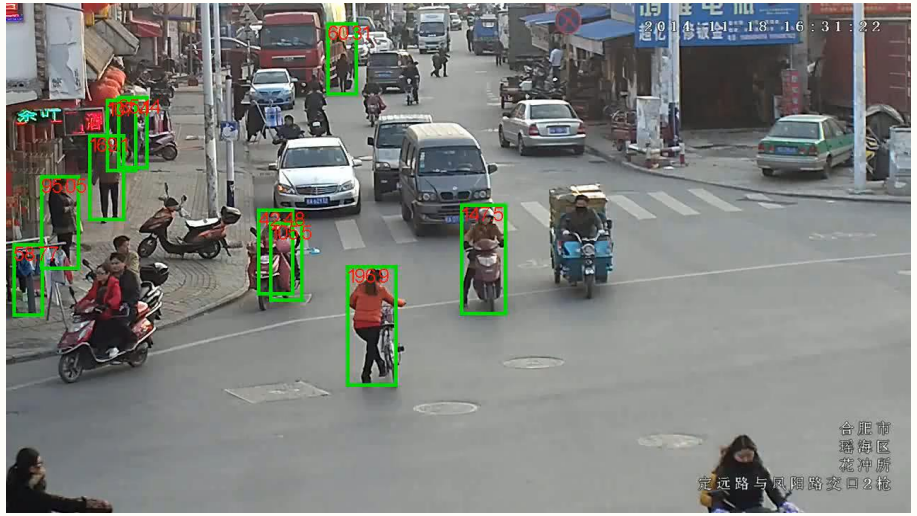
# 4.2 基于多特征协同学习的高效目标检测方法

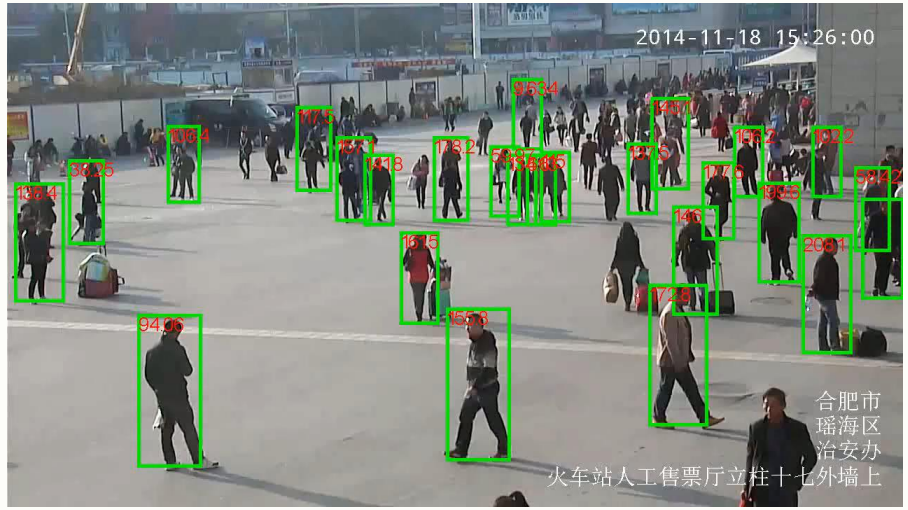
## 行人检测与跟踪

# 4.2 基于多特征协同学习的高效目标检测方法

## 行人检测与跟踪

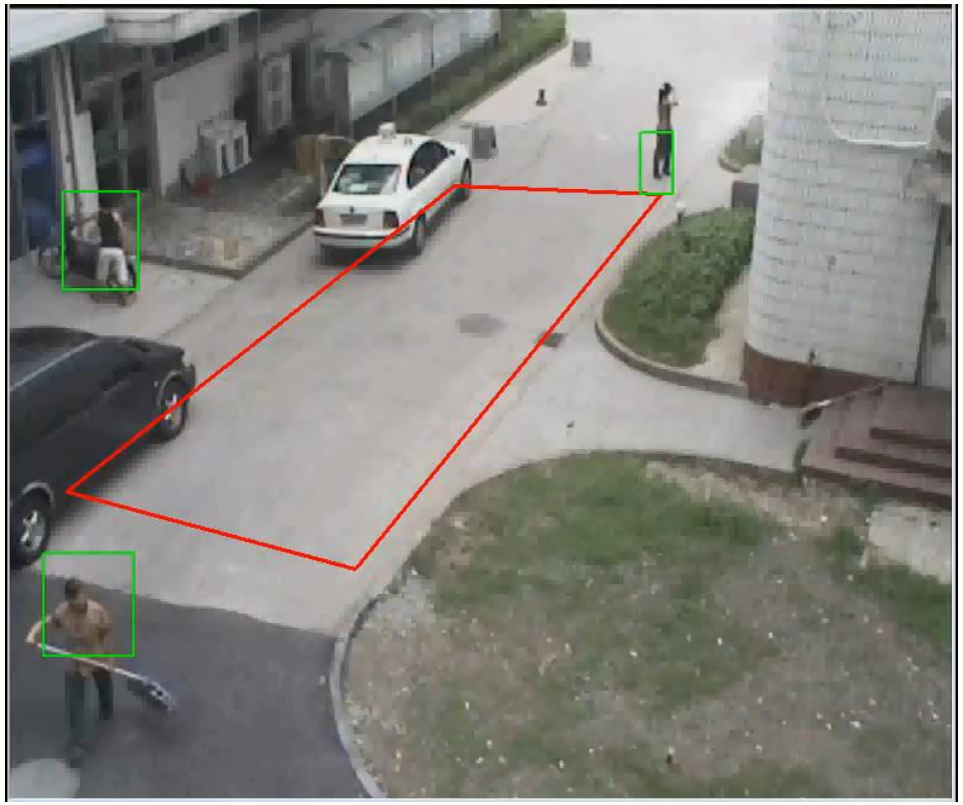



# 4.2 基于多特征协同学习的高效目标检测方法

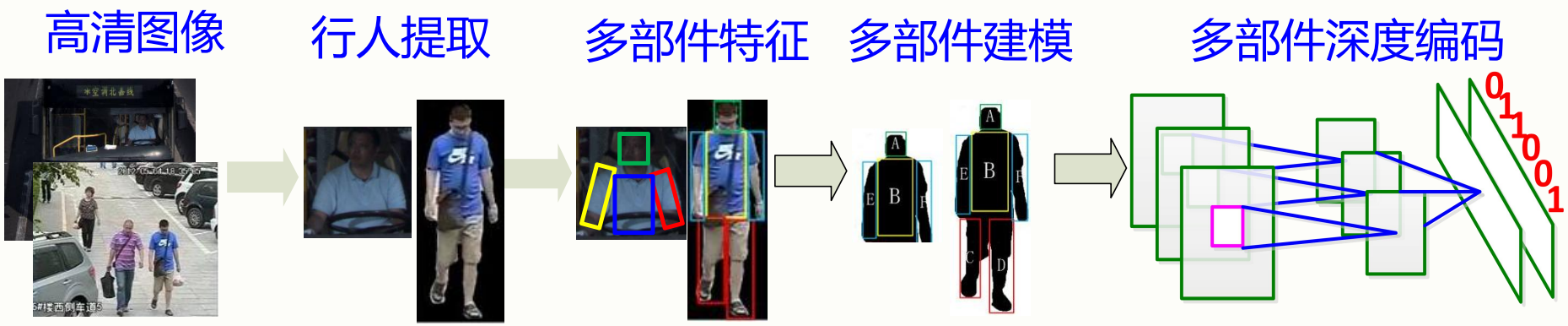
## 枪球联动



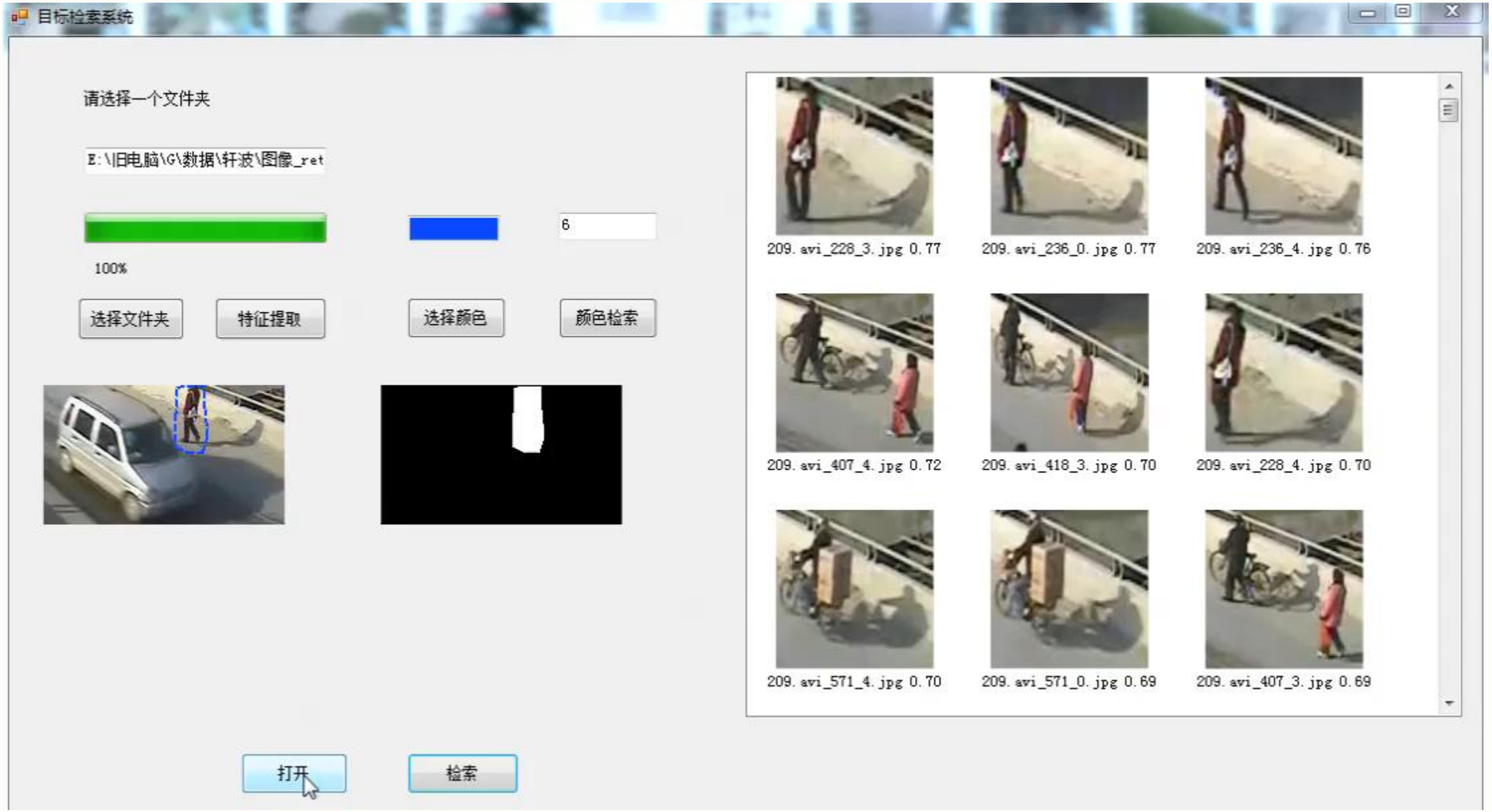
# 4.3 多模态深度编码的目标检索

**视频目标检索的困难与挑战：**

- ✓ 摄像机视角
- ✓ 尺度的变化
- ✓ 遮挡问题
- ✓ 光照的变化
- ✓ 掩膜不完整

# 4.3 多模态深度编码的目标检索



# 4.3 多模态深度编码的目标检索

## 面向大范围场景的人群密度估计

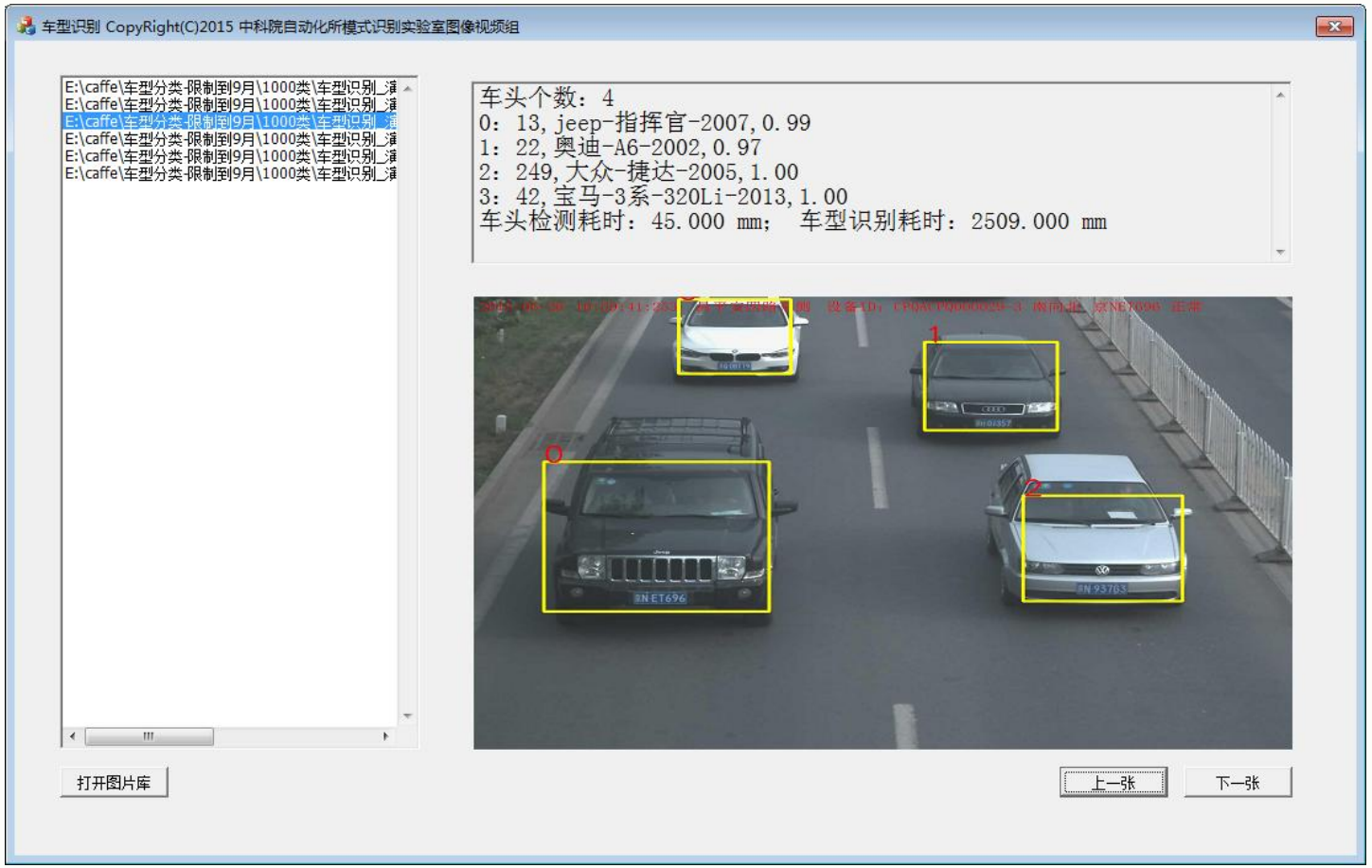
深度卷积编解码网络

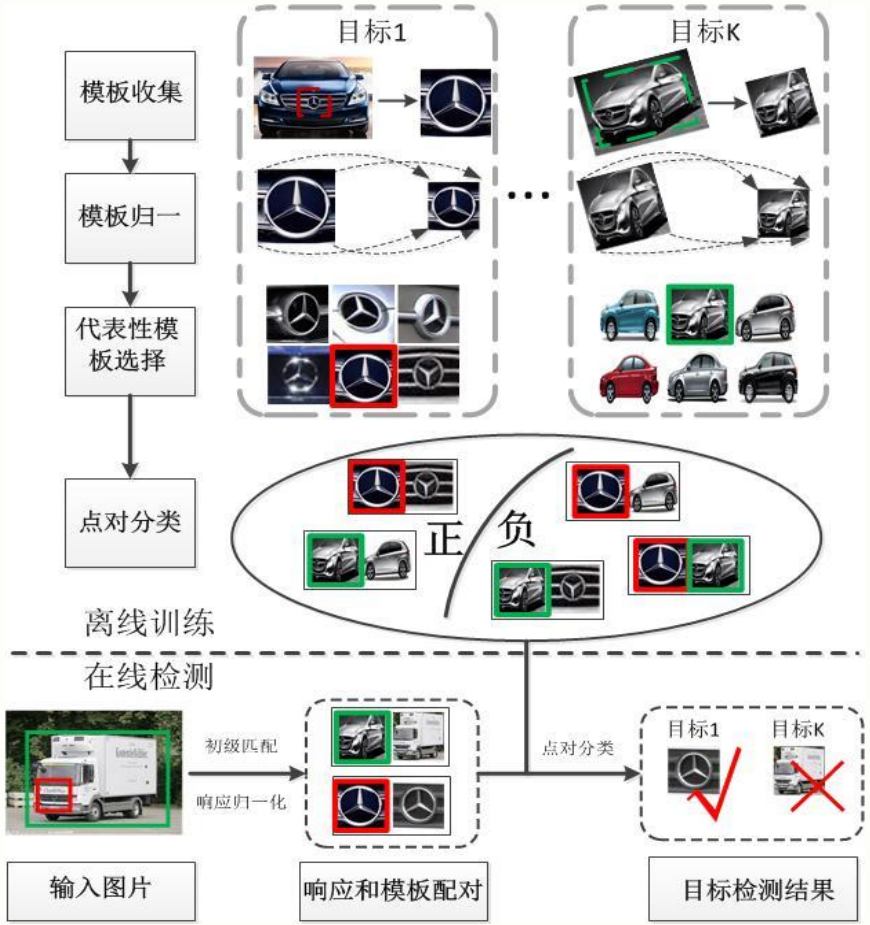
输入图像



场景语义解析


# 4.4 基于全卷积深度网络的车型识别



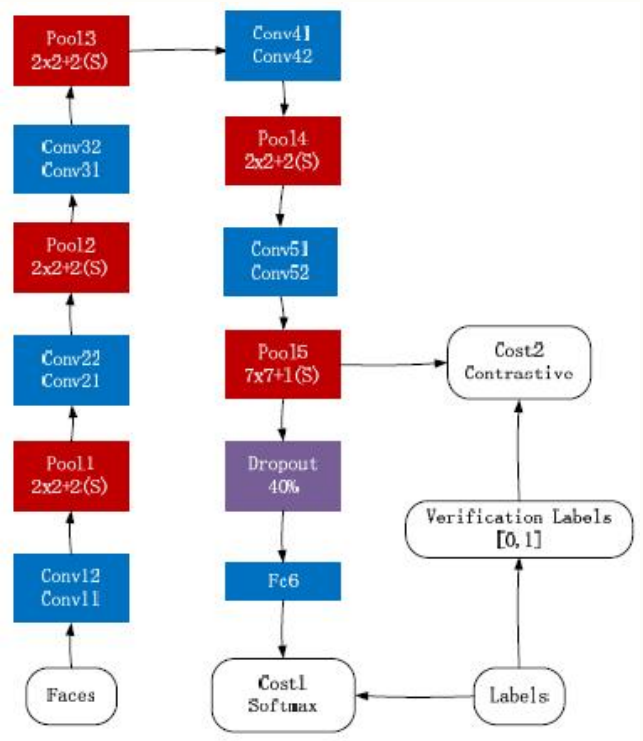




End-to-end 车型和车标识别
1800类 精度98%, CASIA2015

# 4.5 基于多部件深层卷积网络的人脸识别

高精度人脸识别：**多模型模型+海量数据+计算资源**

- 采用very deep网络结构，LFW上单网络人脸识别准确率达到**99.3%**（**512**维特征）。
- 二代证vs现场照片，FAR=0.001处，人脸验证准确率达到95%以上。


Jingqiao Wang，CASIA2014

# 欧老师的联系方式

读万卷书 行万里路 只为最好的修炼


QQ：14777591 (宇宙骑士)

Email：ouxinyu@alumni.hust.edu.cn

Website: http://ouxinyu.cn

Tel：18687840023

地址：安宁校区 诚远楼201

南院 智能应用研究院A306-2