计算机视觉基础：图像理解

胡建芳、谢晓华，郑伟诗

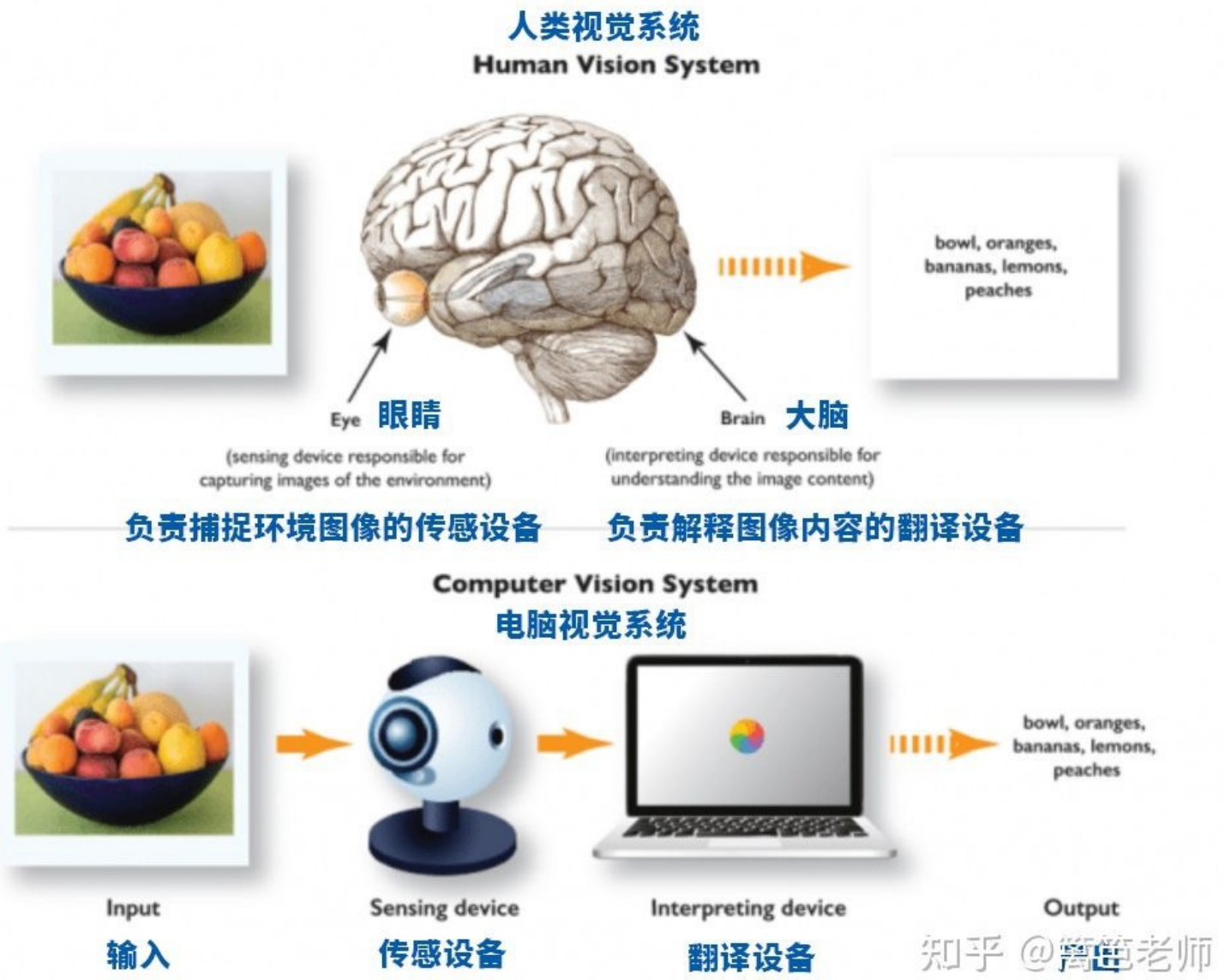
https://cse.sysu.edu.cn/content/5143

中山大学

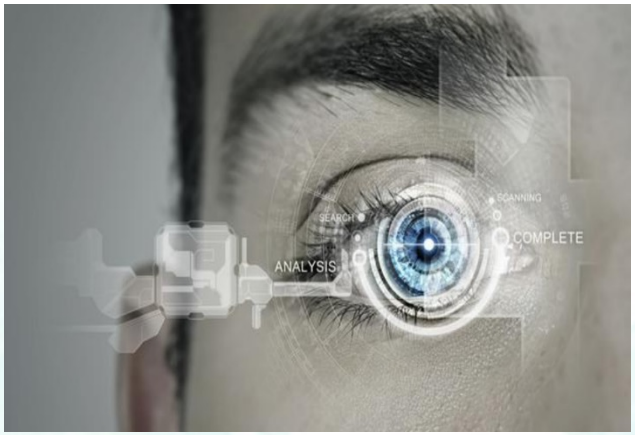
机器智能与先进计算
教育部重点实验室

第一部分：概述

什么是计算机视觉



什么是计算机视觉



至今没有公认统一的定义

人工智能的眼睛、让机器看懂世界

研究如何使人工系统从图像或多维视觉数据中“感知”的科学

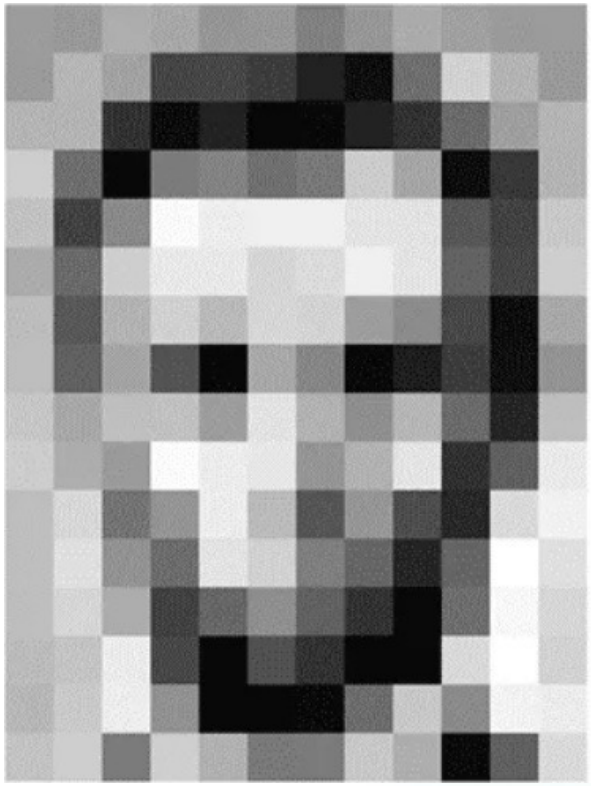
有学者把计算机视觉看成模式识别的一个分支

要区分生物视觉与计算机视觉

高度相关的学科：光学、神经生物学、信号处理、机器学习、模式识别、数学

什么是计算机视觉

人看到的



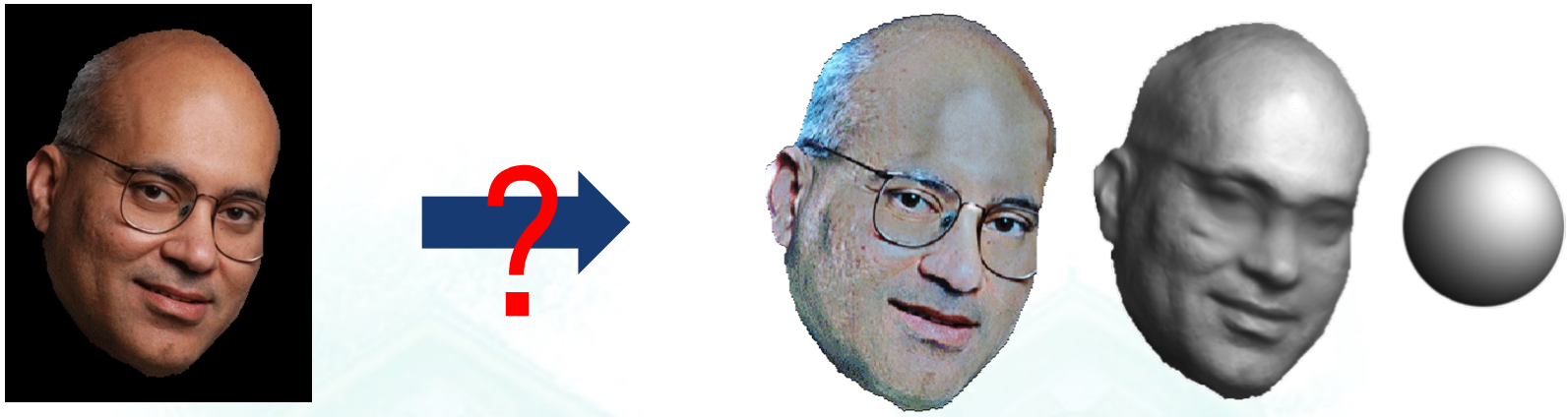
计算机看到的

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

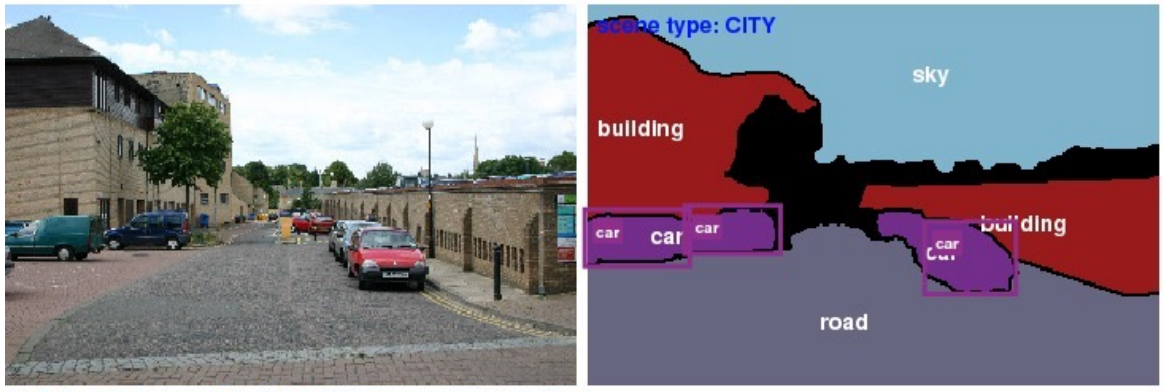
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

从一组数字中抽象出语义信息

计算机视觉做什么



底层属性理解与重建（三维、光照、材质（反射率）、视角……）



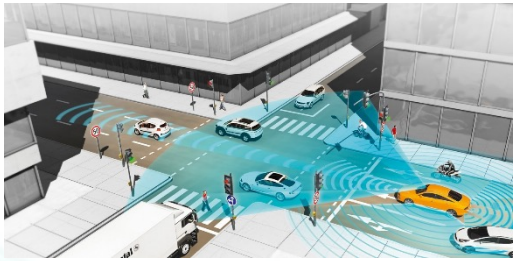
高层语义识别：目标检测、目标分割、身份识别、行为识别、属性识别、材料分类……

计算机视觉有什么用

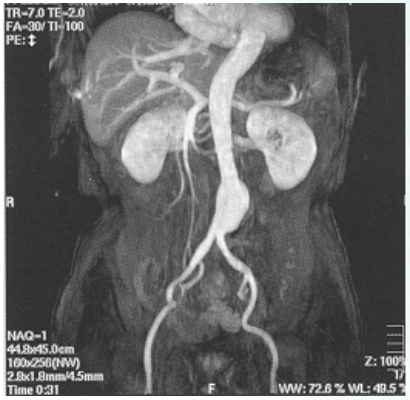
安全监控

军事应用

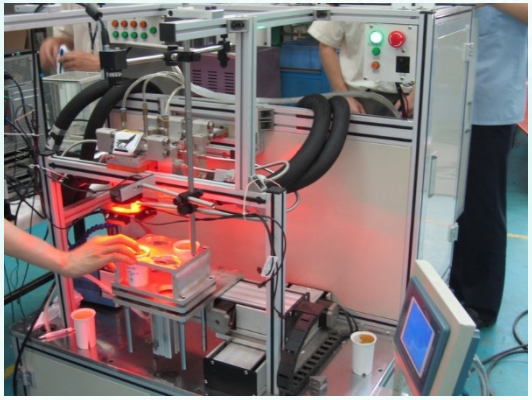
城市管理



智能交通



医学辅助

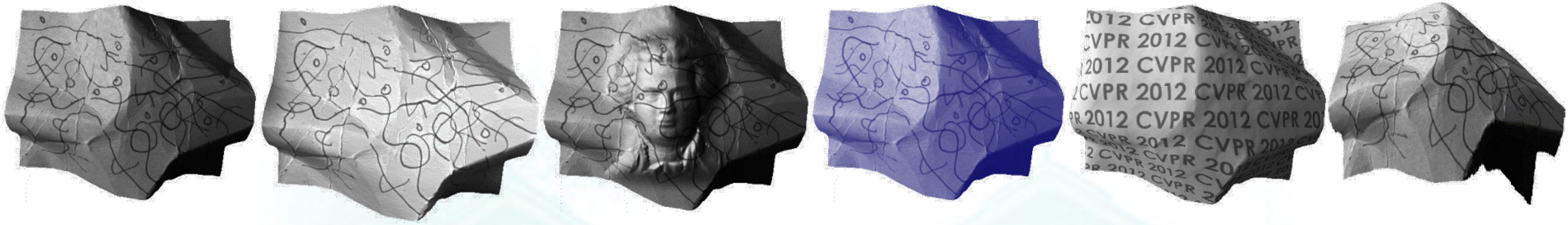


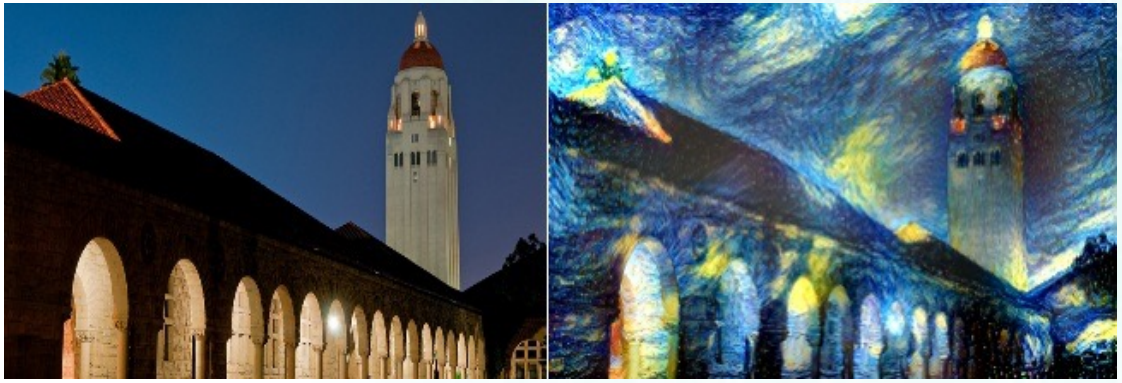
工业检测

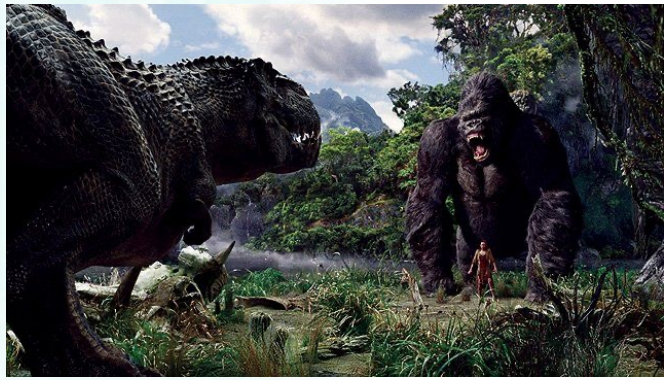
人机交互

……

计算机视觉有什么用







服务于： 数字艺术、虚拟现实、混合现实

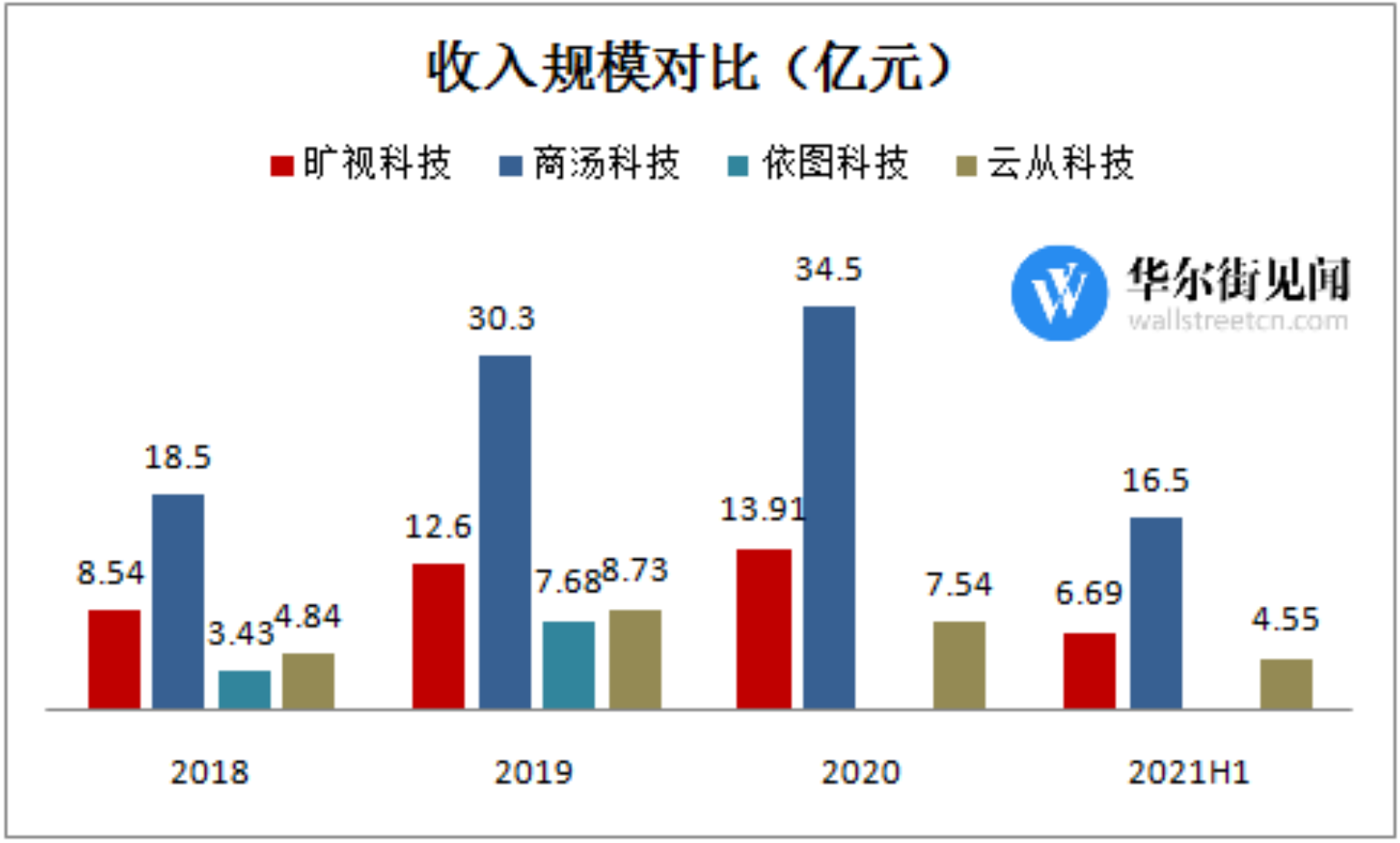
市场行情

图表 6：中国 AI 融资规模过去两年短期遇冷

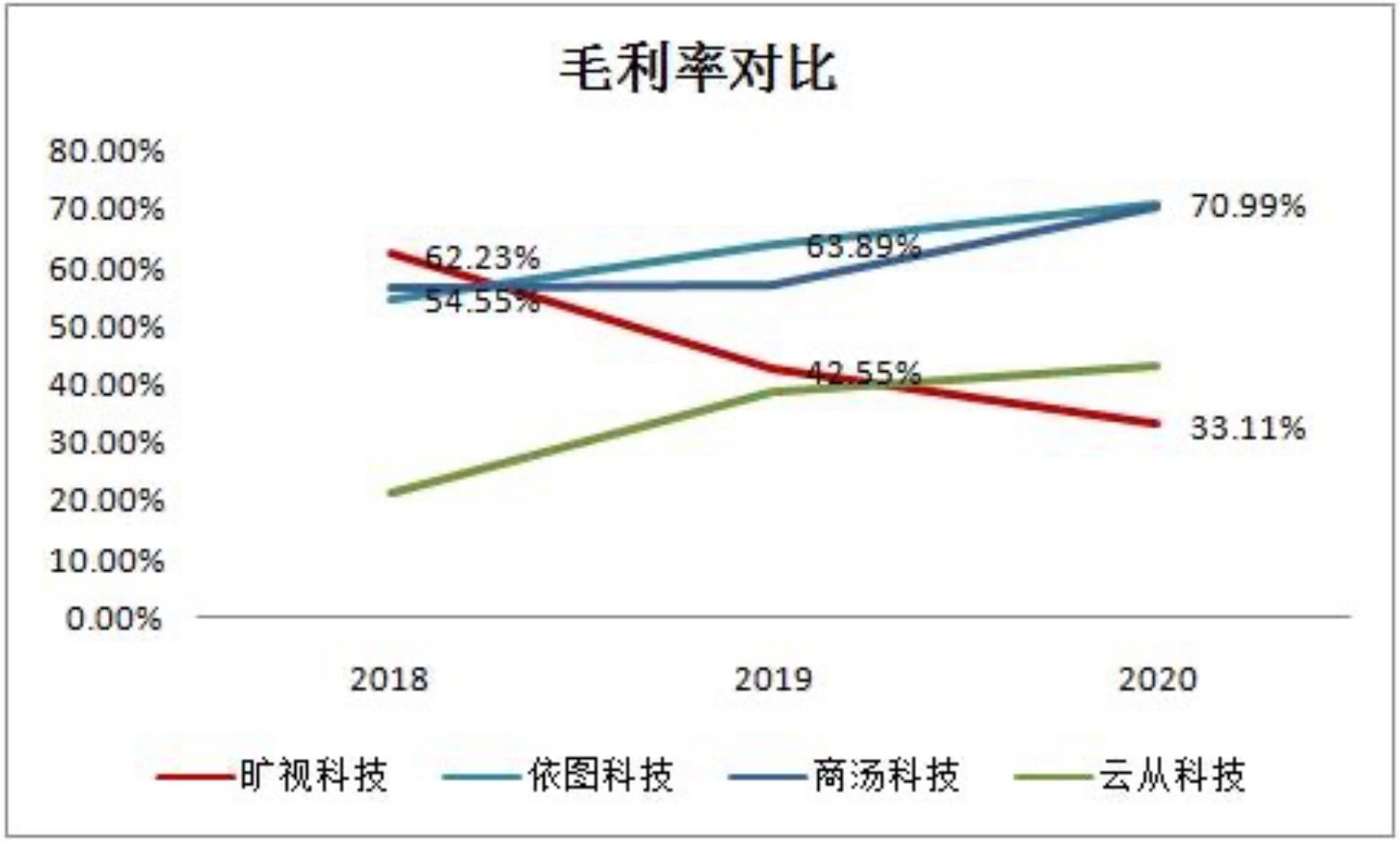
来源：IT 橘子，深圳市人工智能行业协会，国金证券研究所

市场情况

市场行情


收入规模对比（亿元）
旷视科技
商汤科技
依图科技
云从科技
华尔街见闻
wallstreetcn.com
8.54
18.5
3.43
4.84
12.6
30.3
7.68
8.73
13.91
34.5
7.54
6.69
16.5
4.55
2018
2019
2020
2021H1

市场行情


毛利率对比
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
62.23%
54.55%
63.89%
70.99%
42.55%
33.11%
2018
2019
2020
旷视科技
依图科技
商汤科技
云从科技

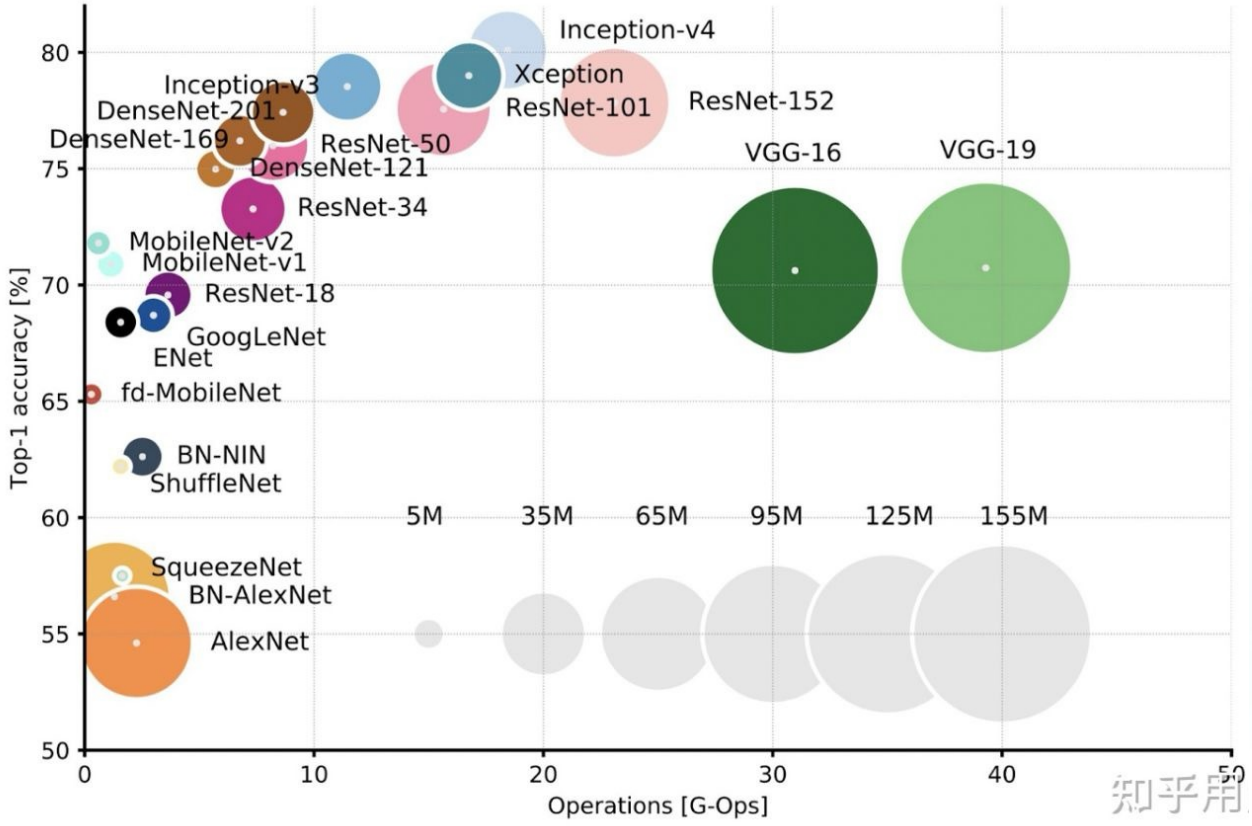
市场情况

市场行情


净亏损对比（亿元）
旷视科技
商汤科技
依图科技
云从科技
2018
2019
2020
2021H1
-28
-34.44
-11.61
-1.81
-66.43
-49.68
-36.42
-17.08
-33.26
-121.58
-6.9
-18.58
-37.13
-3.24
华尔街见闻
wallstreetcn.com

发展趋势

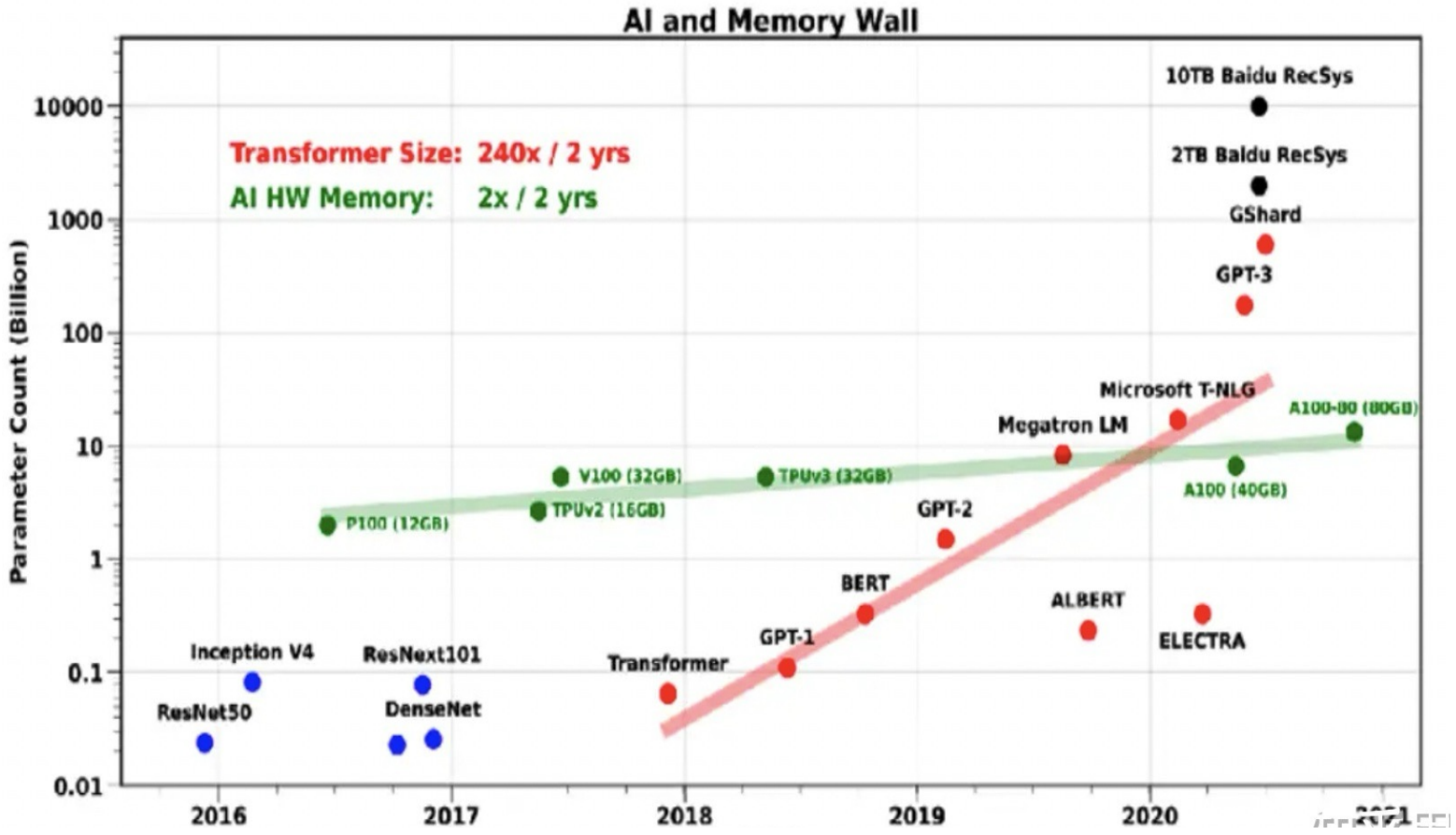
深度学习初期模型越来越大



大模型

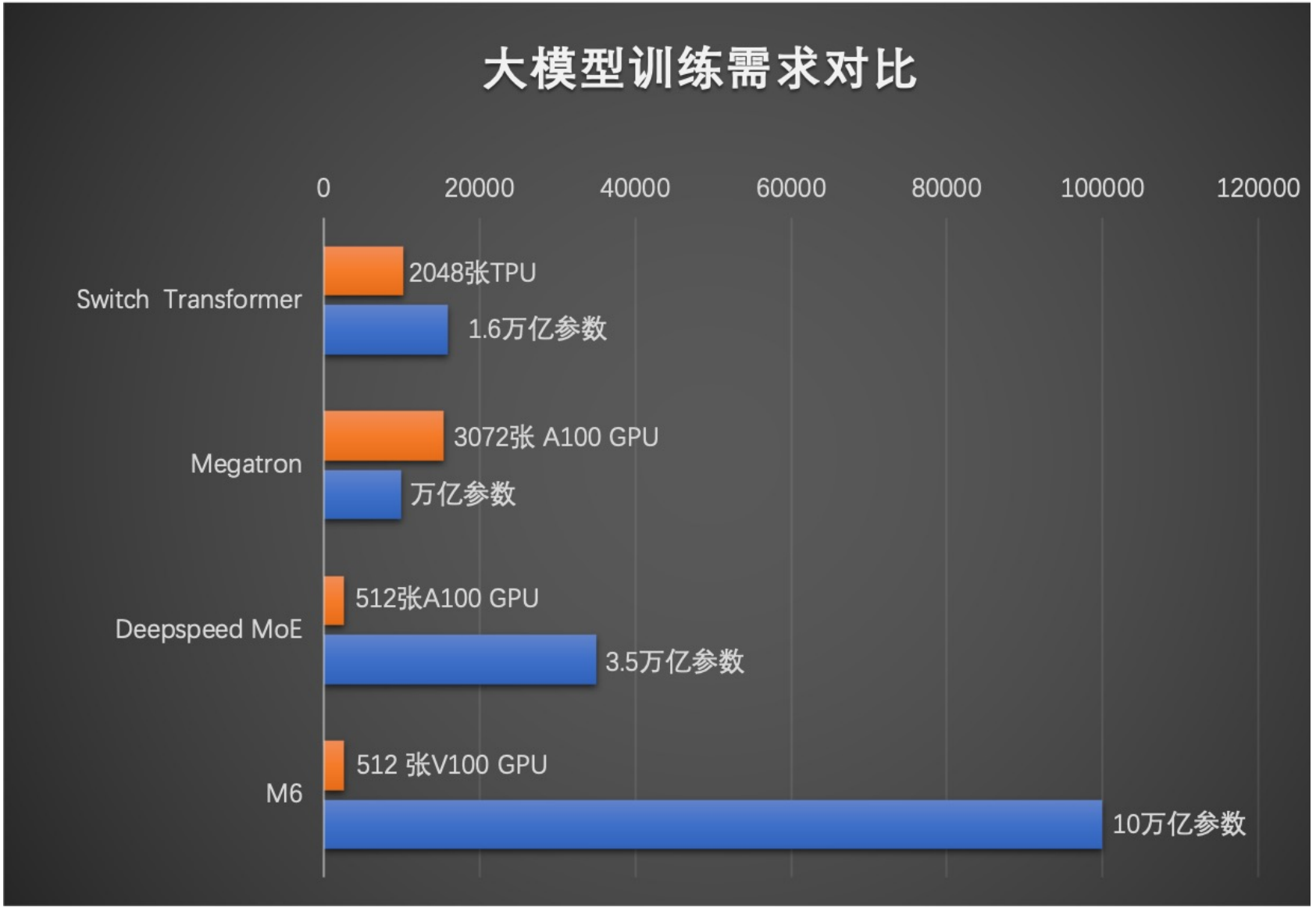
发展趋势

深度学习近年来超大模型涌现



大模型

发展趋势

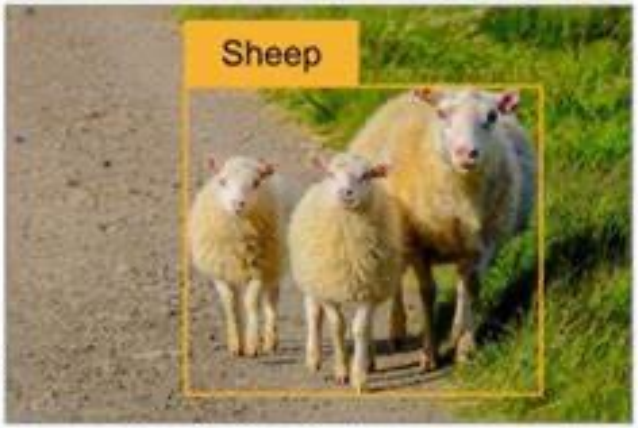

大模型训练需求对比
0
20000
40000
60000
80000
100000
120000
Switch Transformer
2048张TPU
1.6万亿参数
Megatron
3072张 A100 GPU
万亿参数
Deepspeed MoE
512张A100 GPU
3.5万亿参数
M6
512 张V100 GPU
10万亿参数

第二部分：图像理解与识别

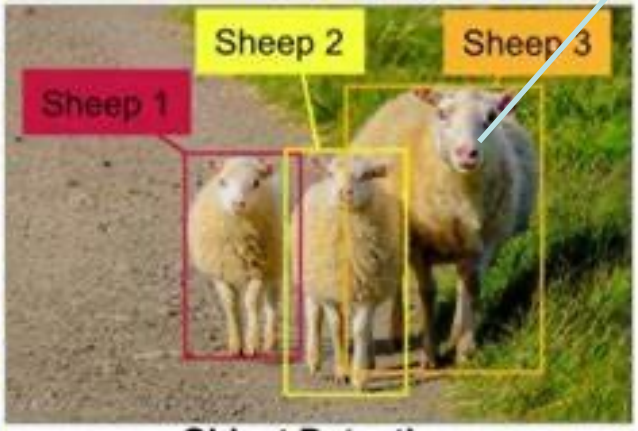
图像识别的基本任务

- 分类、检测、分割

——不同级别：图像、样例、部件、像素

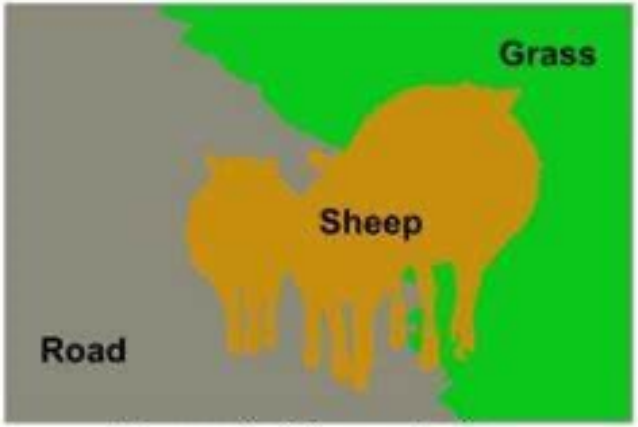


Classification + Localization

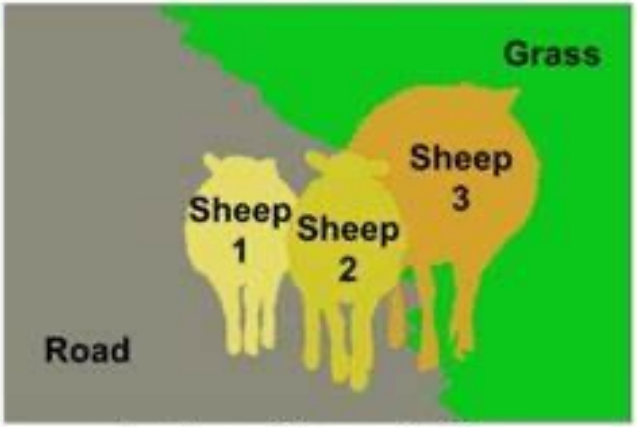


Object Detection

年龄：2
性别：母
姿态：站
身份：玛丽



Semantic Segmentation



Instance Segmentation

关键：
特征表达
分类方法

图像识别的基本任务

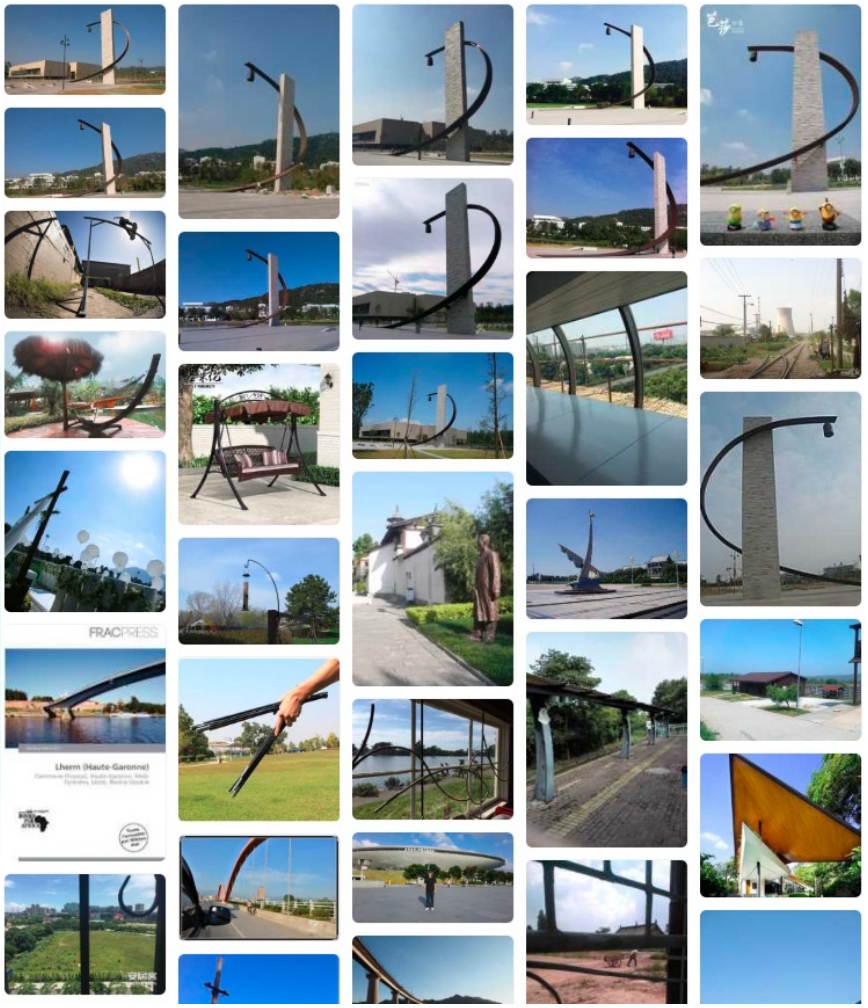
- 物体检测、显著物体检测、非显著物体检测、显著性检测、显著物体分割……

图像识别的基本任务

- 检索

——性能如何评估？

输入



百度检索结果

图像识别的传统流程


Low-level sensing
Pre-processing
Feature extract.
Feature selection
Inference: prediction, recognition

计算机视觉算法怎么做：特征提取

- 基于手工设计的特征：

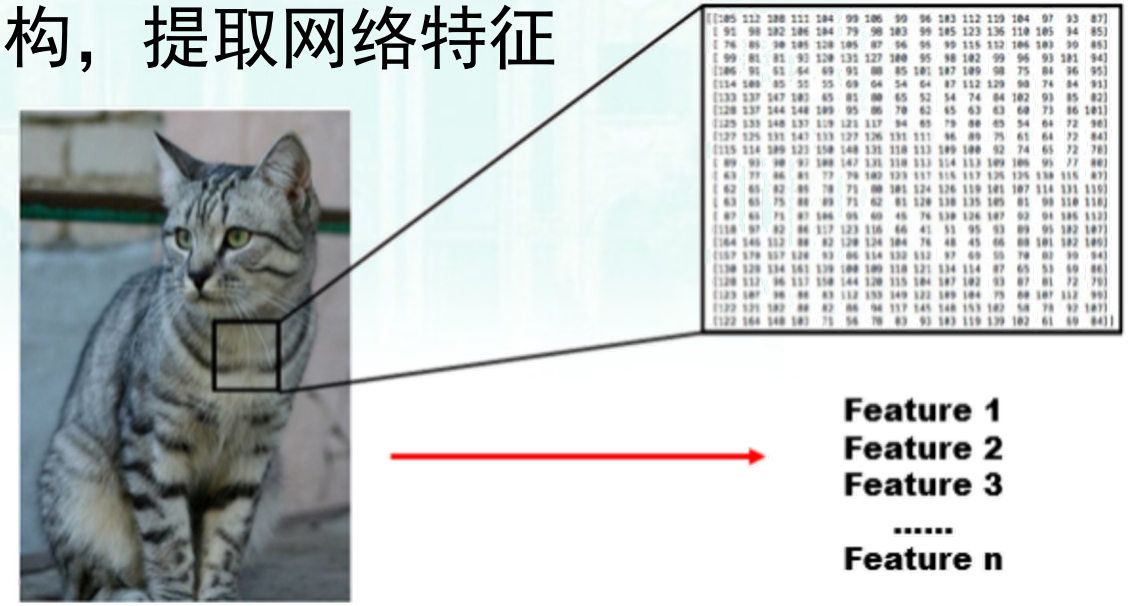
  根据人工经验，设计特征计算方法，需要较强的工程技术能力。

  手工设计很难，几十年来也就是那么几套方案

- 基于深度学习的特征：

  手工设计网络结构，以学习的方式，提取视频特征

  自动学习网络结构，提取网络特征



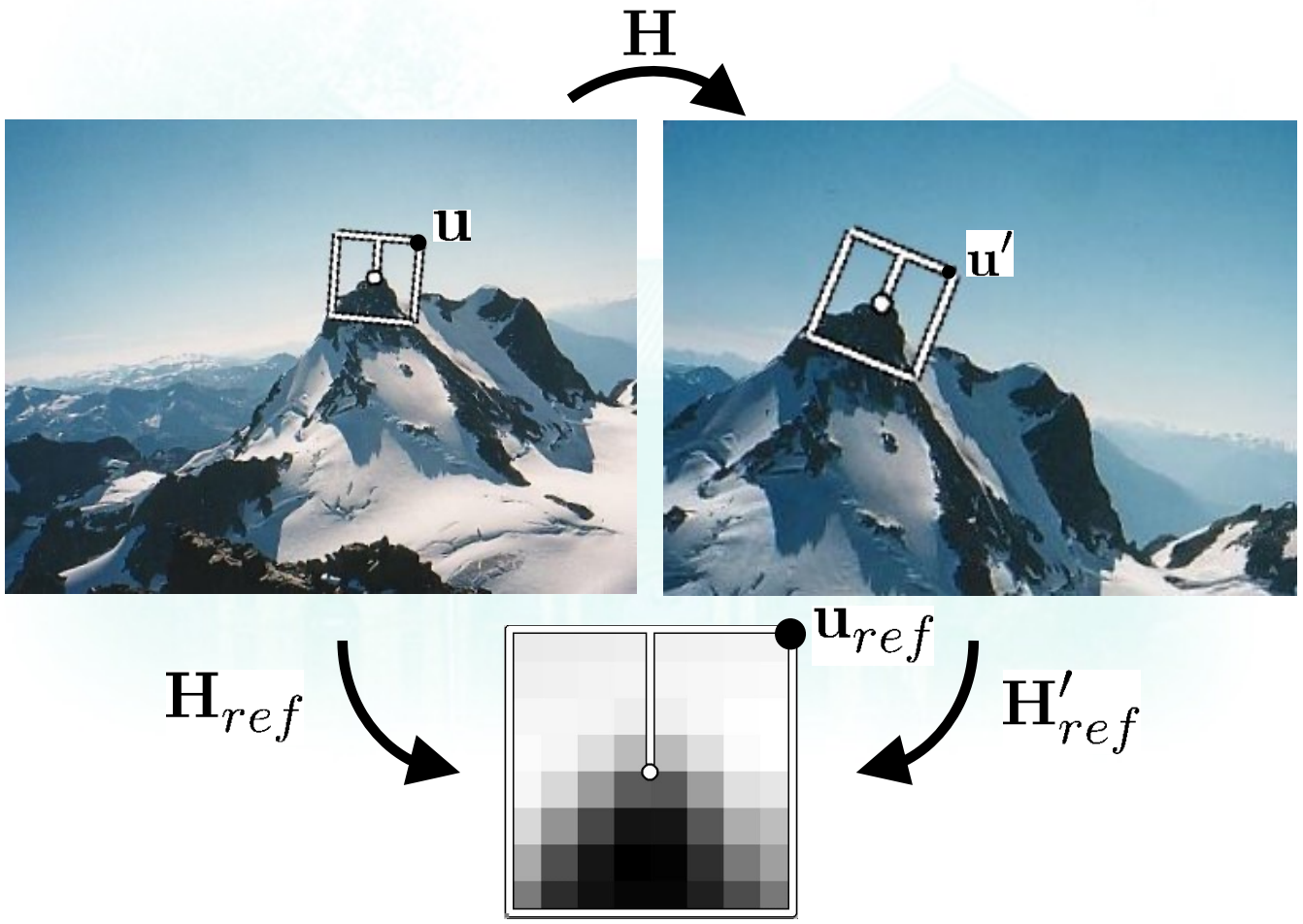
图像特征

- **颜色**：颜色直方图、颜色矩、颜色聚合向量、颜色相关图
- **纹理**：灰度共生矩阵、能量谱、纹理基元、随机场、小波、Gabor、 LBP、HOG、 SIFT、SURF、ORB、BRIEF
- **形状**：几何参数（如面积、周长、圆度、偏心率、主轴方向、代数不变矩等）、傅里叶形状描述符、有限元法、Shape Context
- **深度学习特征**
- **空间关系**

- **特征的鲁棒性**：旋转、尺度、平移、光照变化、仿射、视角、噪声
- **高级鲁棒性（与任务相关）**：跨年龄不变、遮挡不变、形变不变
- **相关研究**：关键点检测、特征编码（面向匹配）、度量学习

SIFT（Scale Invariant Feature Transform）

- 一种基于尺度空间的、对图像缩放、旋转甚至仿射变换保持不变性的图像局部特征描述算子。



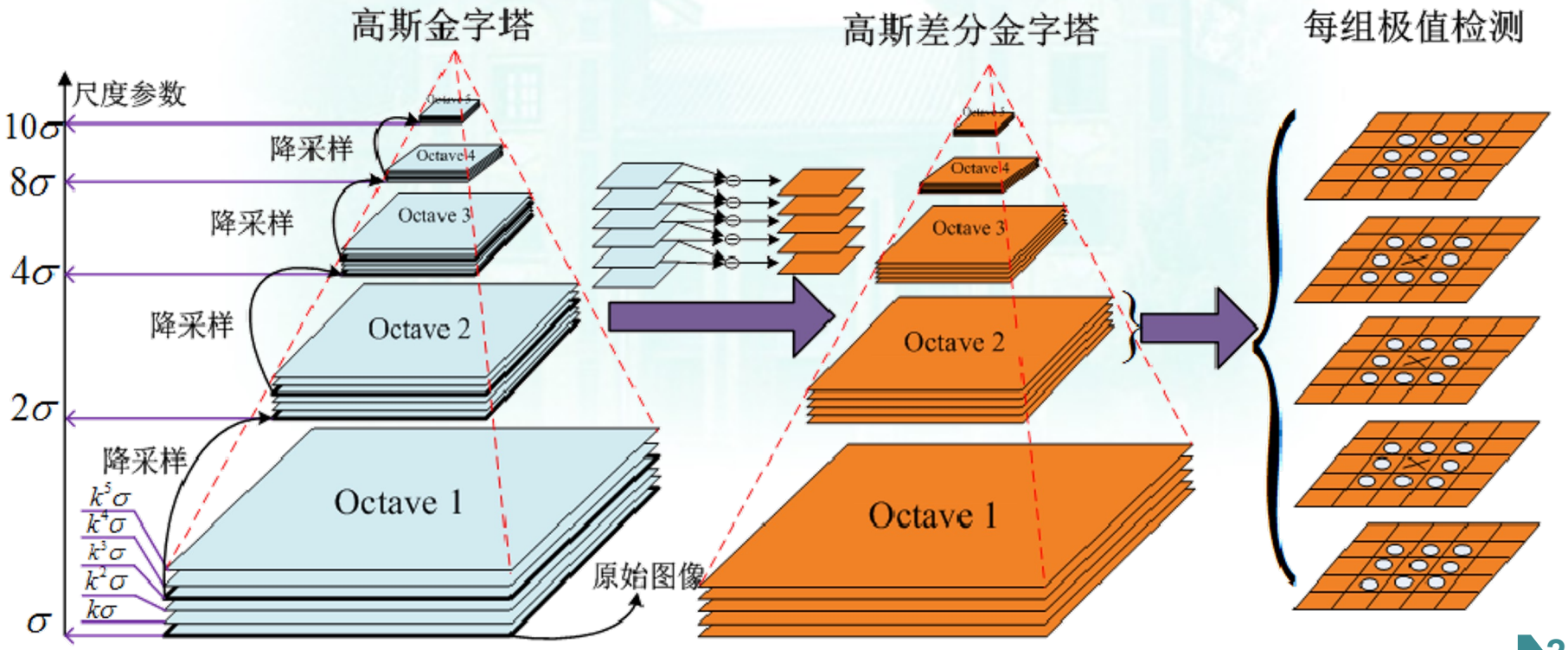
SIFT（Scale Invariant Feature Transform）

- 尺度空间极值点检测（差分高斯代替拉普拉斯）

高斯尺度空间：$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y)$

高斯差分空间：$D(x,y,\sigma)=[G(x,y,k\sigma)-G(x,y,\sigma)]*I(x,y)$

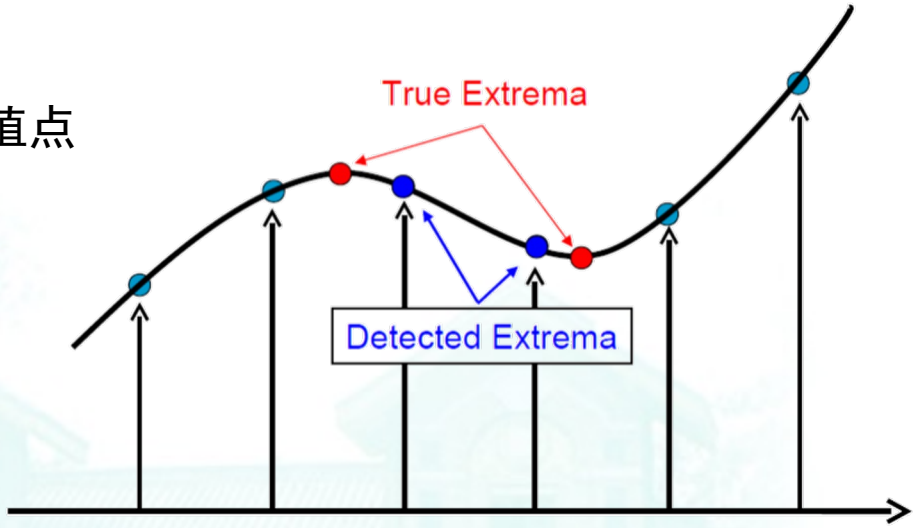
$=L(x,y,k\sigma)-L(x,y,\sigma)$

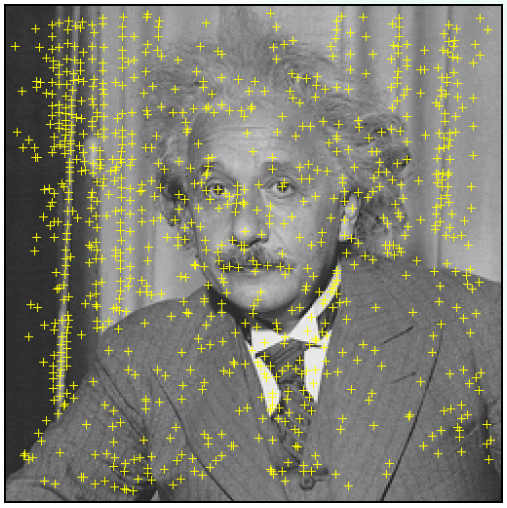


SIFT（Scale Invariant Feature Transform）

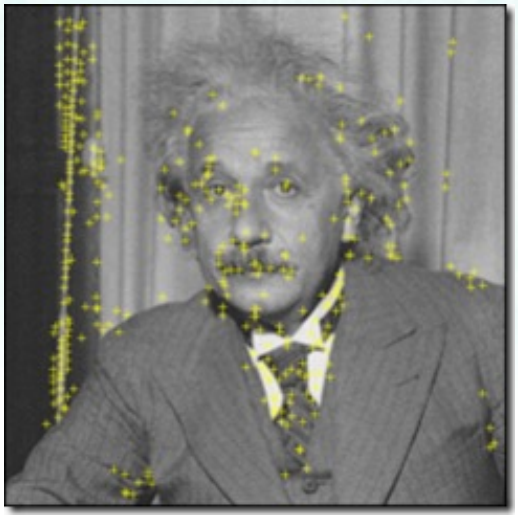
❑ 关键点定位

- 插值得到原始分辨率空间极值点
——更改偏移量大的点
- 剔除低对比度的点
- 剔除不稳定的边缘点

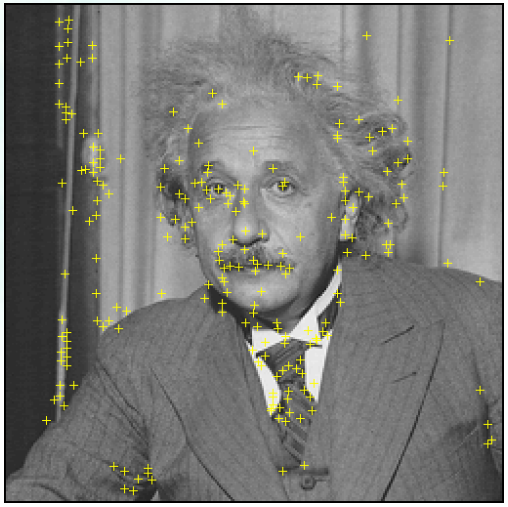




原始极值点



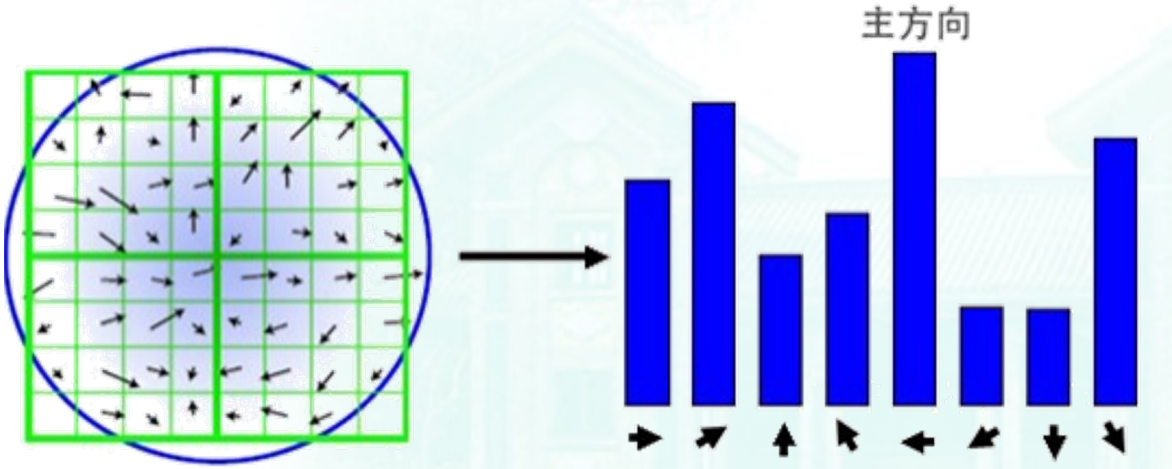
剔除低对比度点

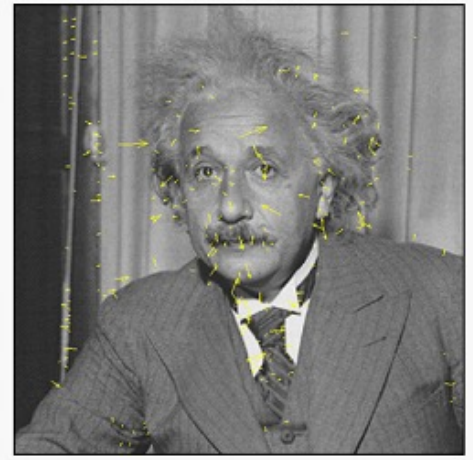


剔除边缘点

SIFT（Scale Invariant Feature Transform）

- 主方向指派





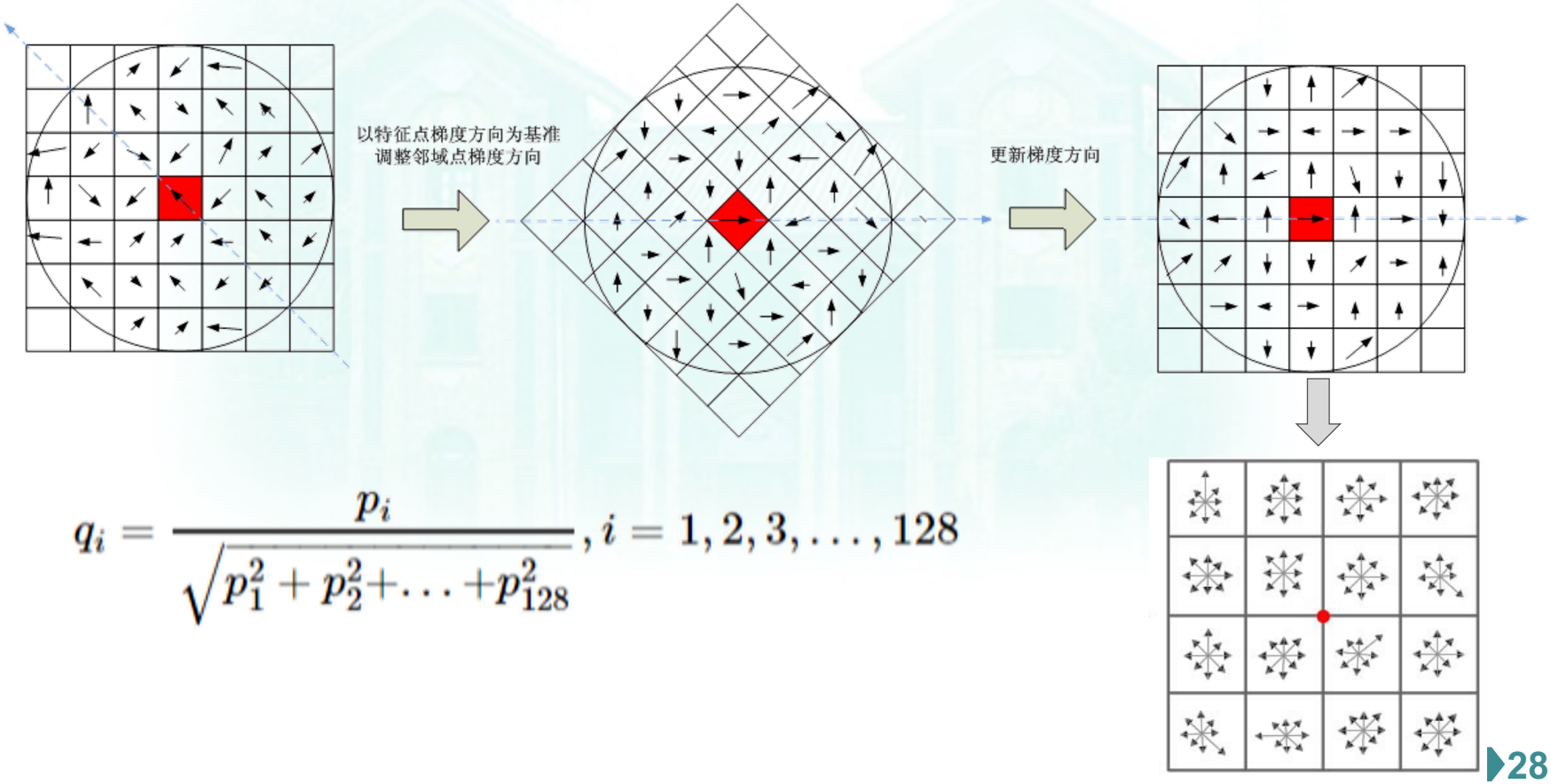
最终SIFT特征点
（位置、尺度、方向）

为了增强匹配的鲁棒性，可保留峰值大于主方向峰值80%的方向作为该关键点的辅方向

SIFT（Scale Invariant Feature Transform）

❑ 生成描绘子

- 校正旋转主方向，确保旋转不变性；
- 邻域块每块同级8个方向的直方图；
- 形成一个8x16=128维的特征向量（高斯权体现每个领域点的贡献）；
- 归一化+阈值处理，去除光照的影响。



$$q_i = \frac{p_i}{\sqrt{p_1^2 + p_2^2 + \ldots + p_{128}^2}}, i = 1, 2, 3, \ldots, 128$$

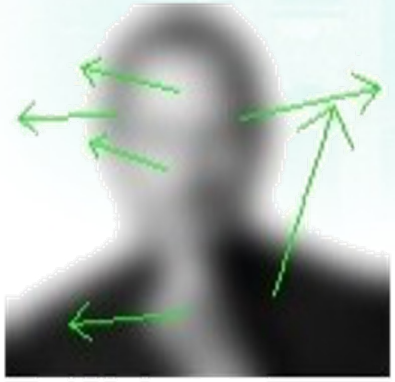
SIFT（Scale Invariant Feature Transform）

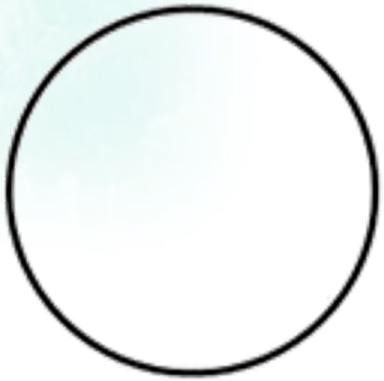
❑ 成功应用

- 图像比对
- 目标检测
- 特征点匹配

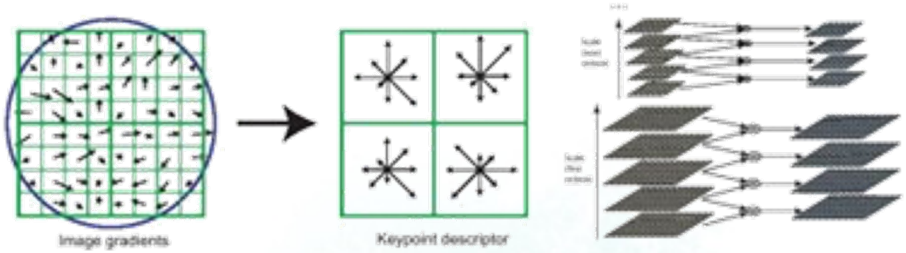
❑ 失败情况

- 模糊图像
- 光滑边缘
- 圆

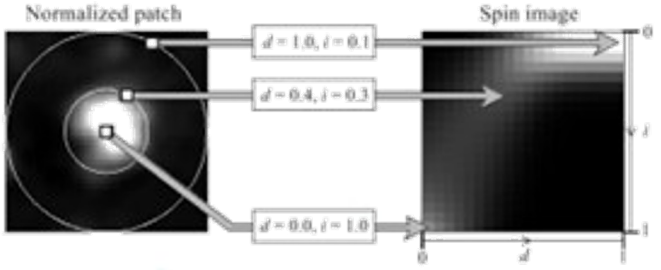




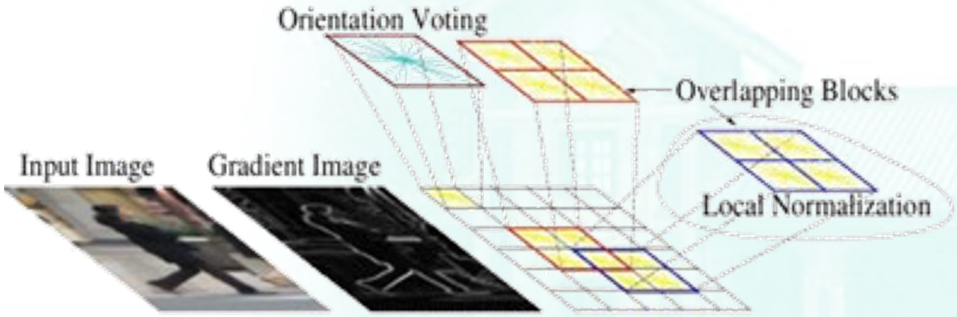
更多的特征描绘子



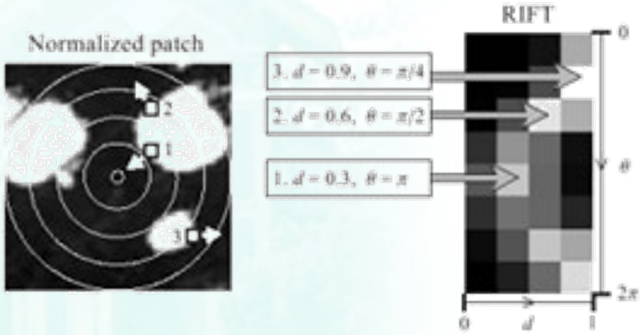
SIFT



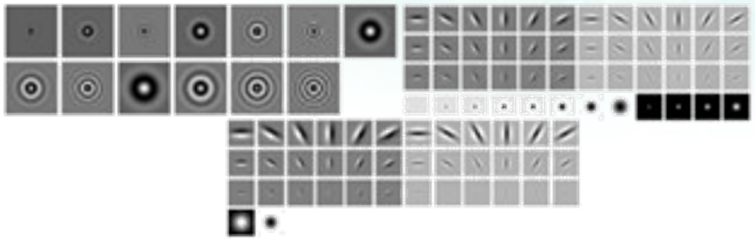
Spin image



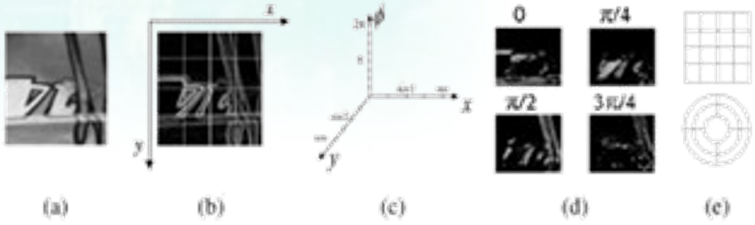
HoG



RIFT



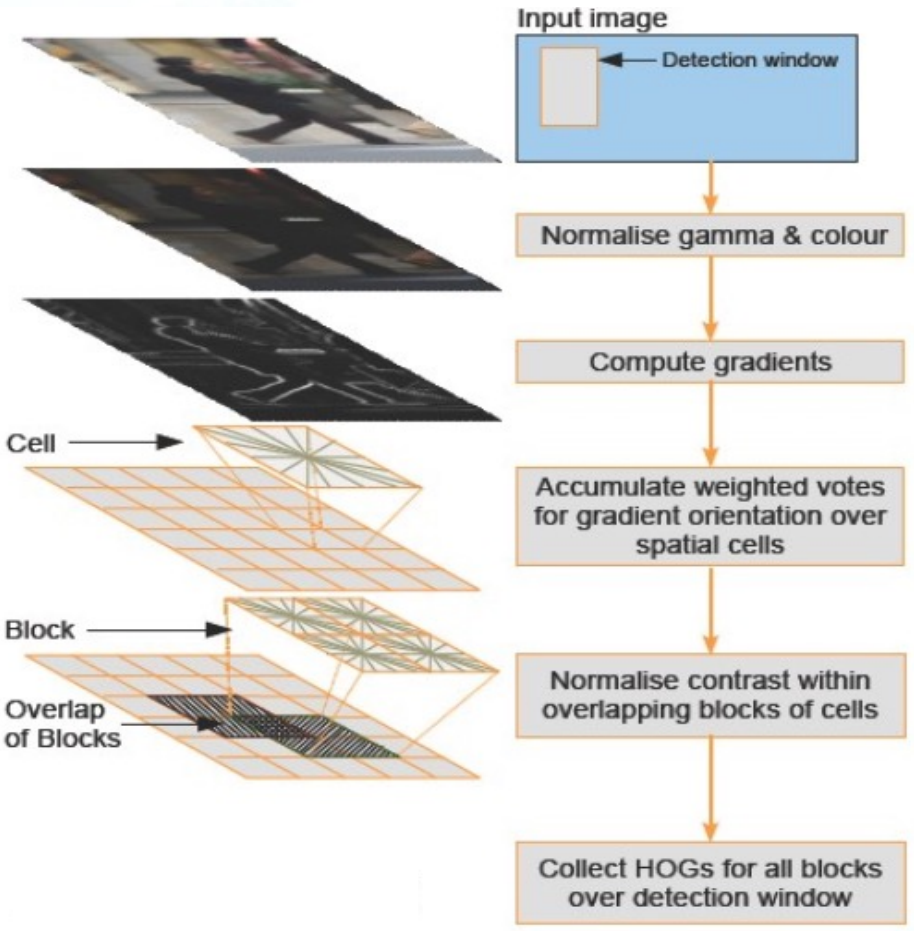
Textons



GLOH

HOG（Histogram of Oriented Gradient）

- 通过计算和统计图像局部区域的梯度方向直方图来构成特征，关注对象的结构或形状

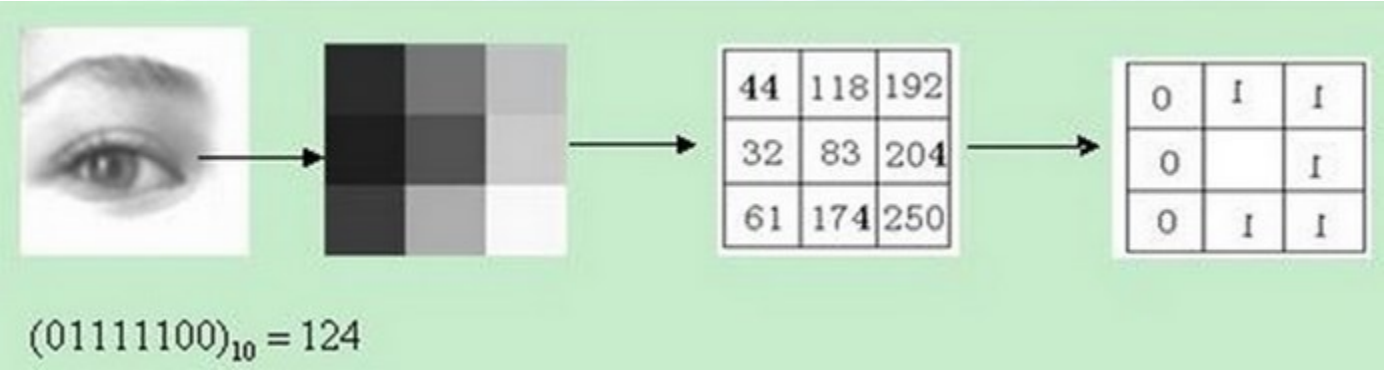


LBP（Local Binary Pattern）

- 描述图像局部纹理的特征

原始的LBP算子定义为在3*3的窗口内，以窗口中心像素为阈值，将相邻的8个像素的灰度值与其进行比较，若周围像素值大于中心像素值，则该像素点的位置被标记为1，否则为0。

3*3邻域内的8个点经比较可产生8位二进制数（通常转换为十进制数即LBP码，共256种），即得到该窗口中心像素点的LBP值，并用这个值来反映该区域的纹理信息。



LBP（Local Binary Pattern）

- 改进1：圆形LBP算子

将 3×3邻域扩展到任意邻域，并用圆形邻域代替了正方形邻域

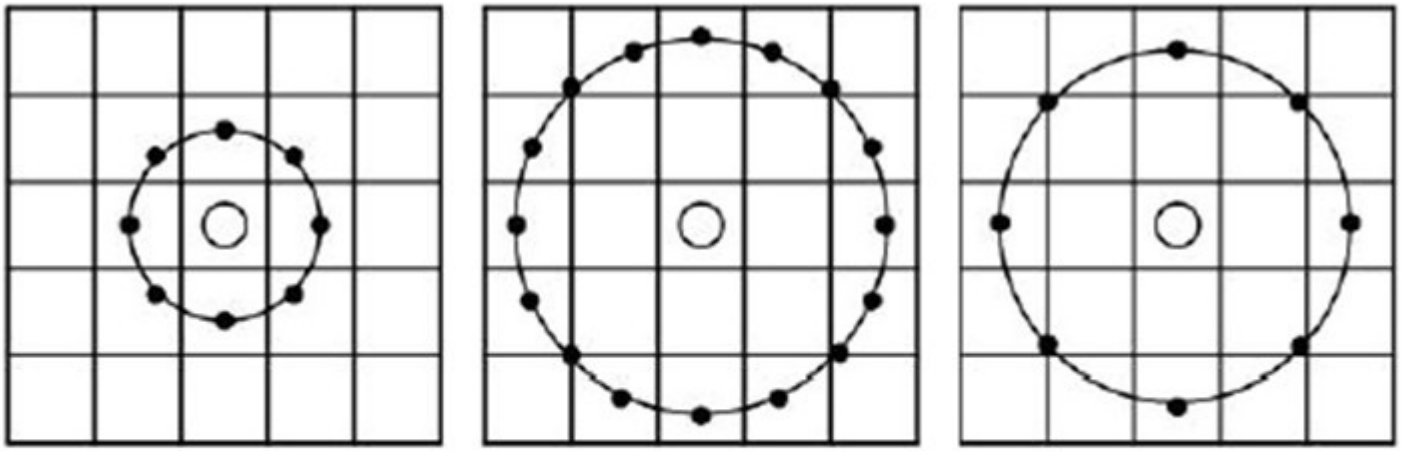
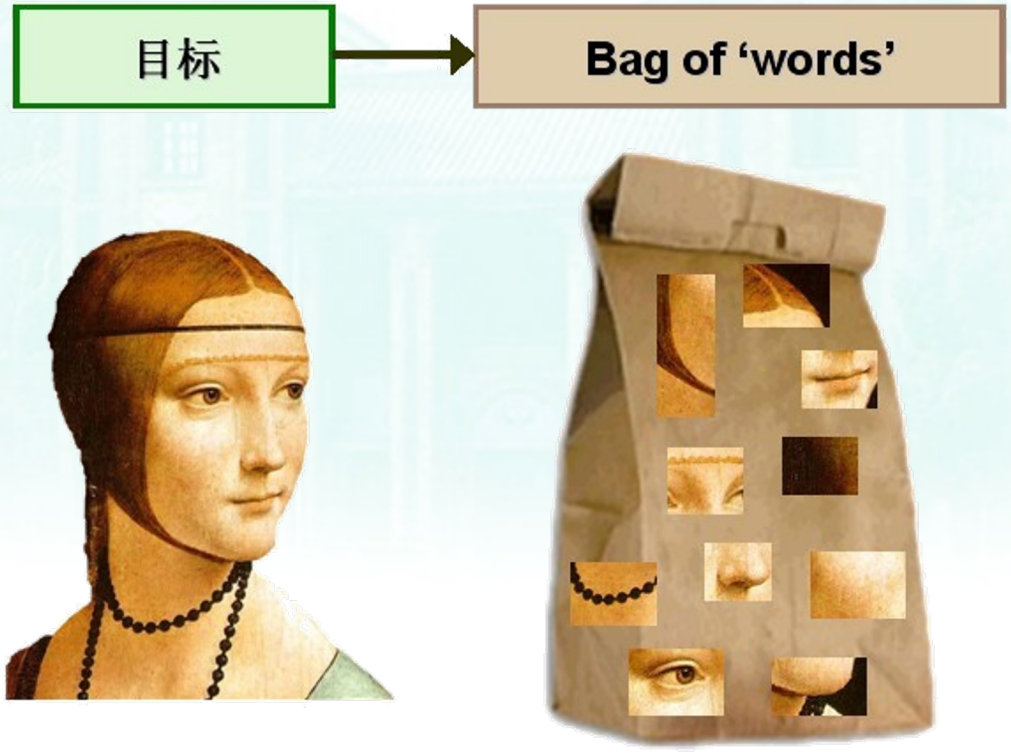


图2.2 几种LBP算子

特征编码

- BOW（Bag-of-Visual-Words）
- FV（Fisher Vector）
- VLAD（vector of locally aggregated descriptors）

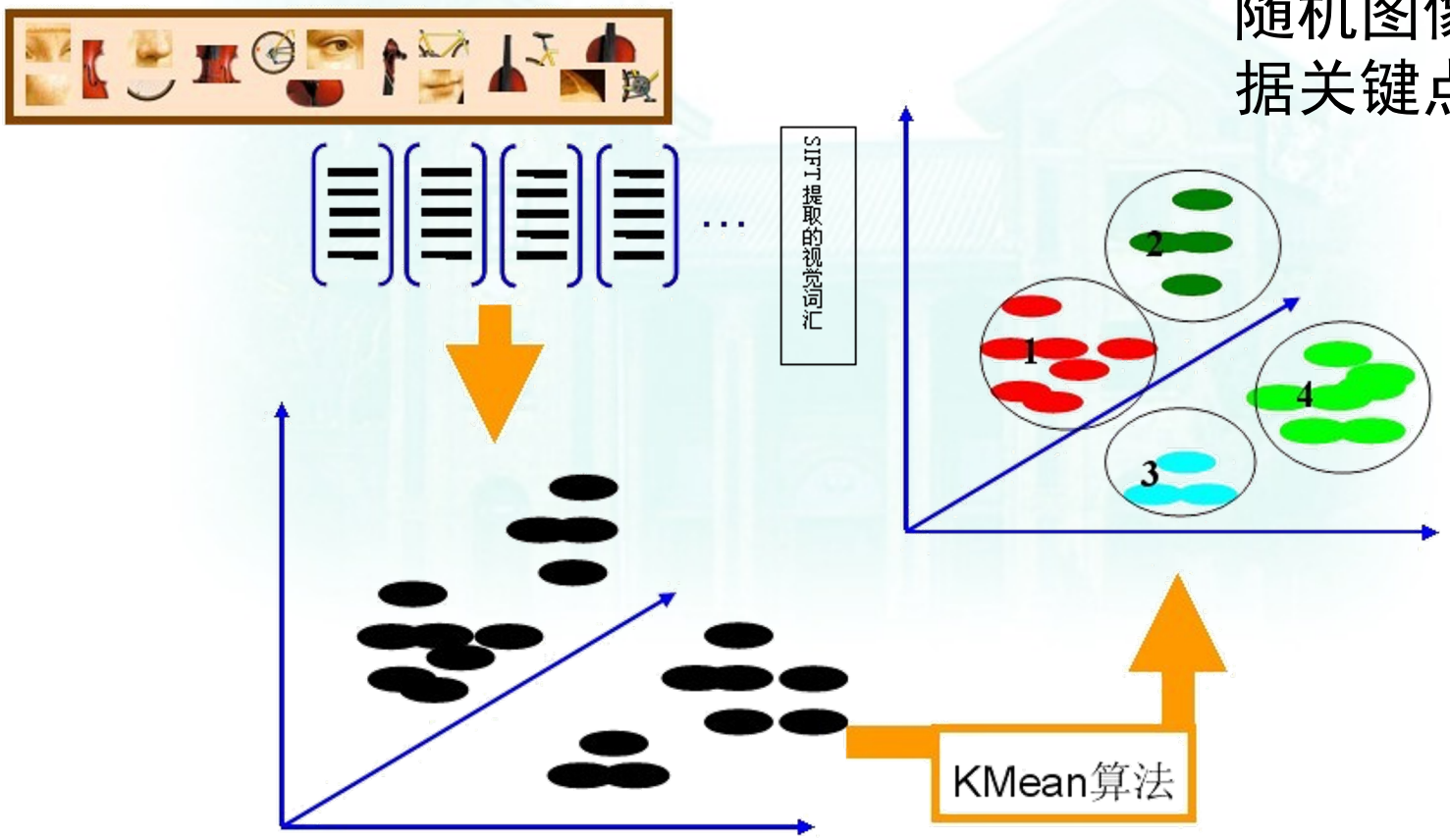


Bag-of-Visual-Words （BoW）

- 步骤：
- 从数据库采样视觉词汇（提取相应特征组）
- 利用K-Means算法构造单词表

如何采样？

随机图像块或者根据关键点（SIFT）



Bag-of-Visual-Words （BoW）

- 步骤：
- 利用单词表的中词汇表示图像(统计单词表中每个单词在图像中出现的次数)

人脸： [3, 30, 3, 20]　　自行车：[20, 3, 3, 2]　　吉他： [8, 12, 32, 7]

分类器

- 最近邻 （无参数，计算量大）
- 线性回归 （假设过于理想）
- Logistic回归 （假设仍然过于理想）
- 朴素贝叶斯 （假设样本间是相互独立）
- 支持向量机 （映射到高维空间，计算复杂）
- 决策树 （可解释，训练耗时）
- 随机森林、 AdaBoost（众志成城，不稳定）
- 深度神经网络（**目前几乎一统江湖！**）

深度神经网络往往集成了特征提取、特征编码、分类（端到端）

第三部分：基于深度学习的图像理解与处理

深度学习三（四）大天王

Yann LeCun

Geoffrey E. Hinton

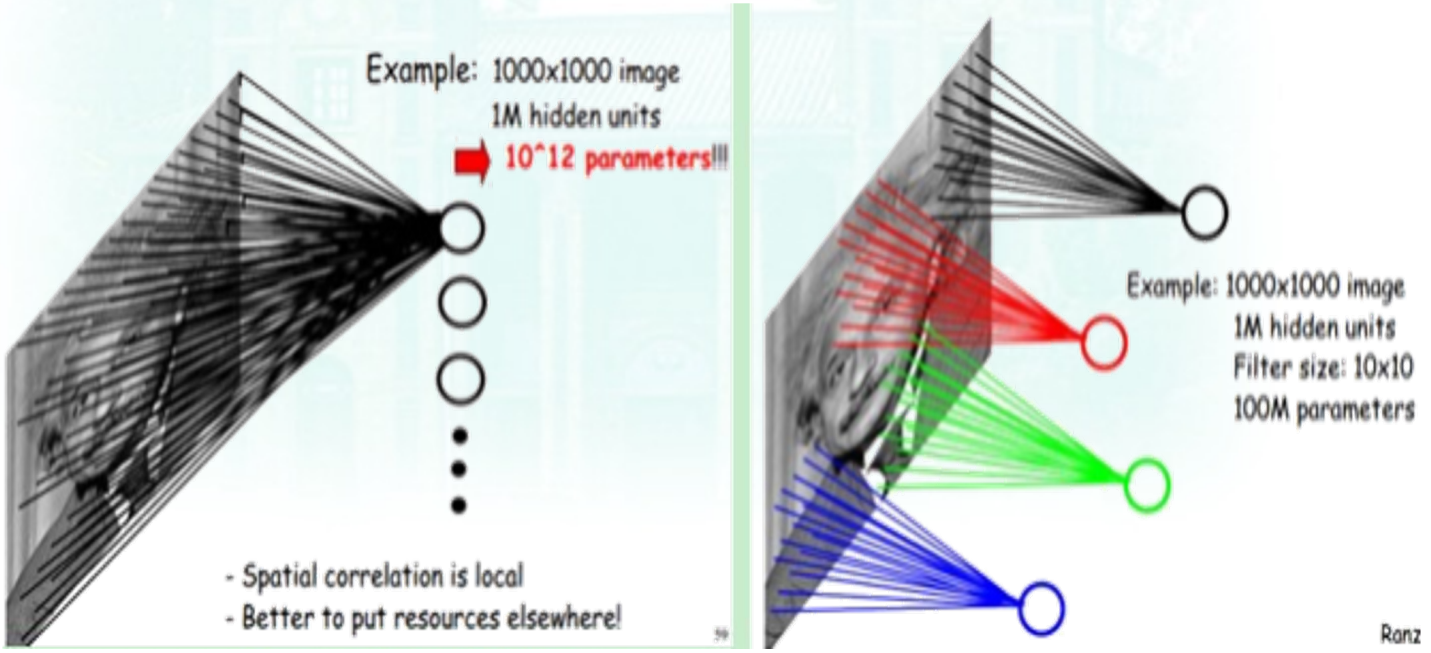
Yoshua Bengio

Andrew Ng

局部连接（卷积神经网络）

- 通过局部连接和权值共享减少神经网络参数个数（远距离像素的相关性很小、不同局部可能出现相似模式）
- 对1000x1000像素的图像，设1百万个隐层神经元

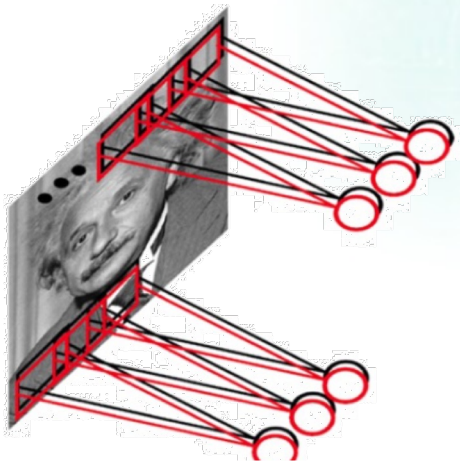
——全连接有1000x1000x1000000=10^12个参数
——局部连接只有100M=10^8个参数（用M个10x10的滤波器）

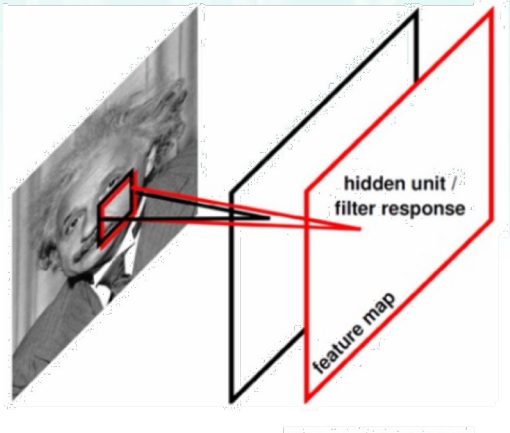


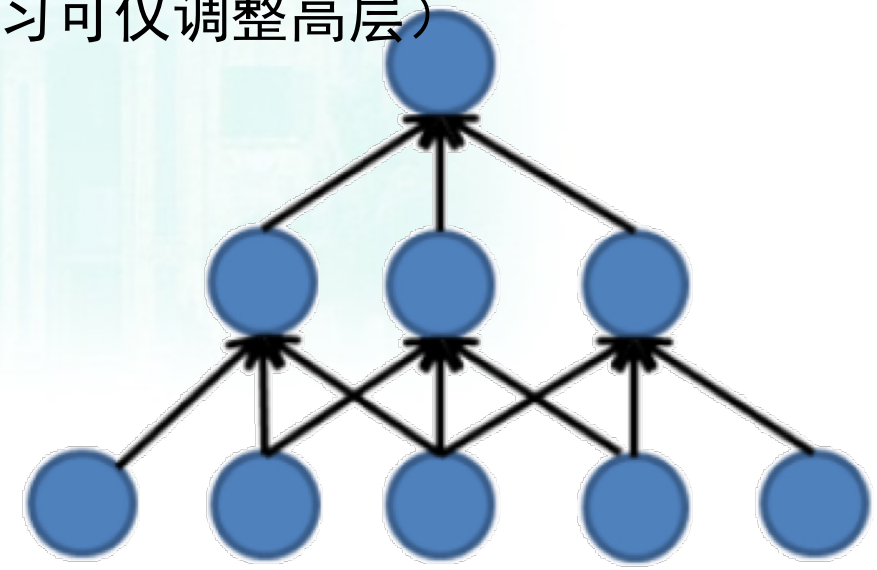
全连接vs. 局部（稀疏）连接

卷积神经网络：从局部到全局感知

- 每一层的滤波器都是捕捉空间局部模式（图像不同位置存在相似边缘）
- 同层可以采用多个滤波器，捕捉多种特征
- 越高层的滤波器具有更大的感受野（receptive field）
- 多层堆积实现全局感知
- 跨任务低层权重可共享（迁移学习可仅调整高层）

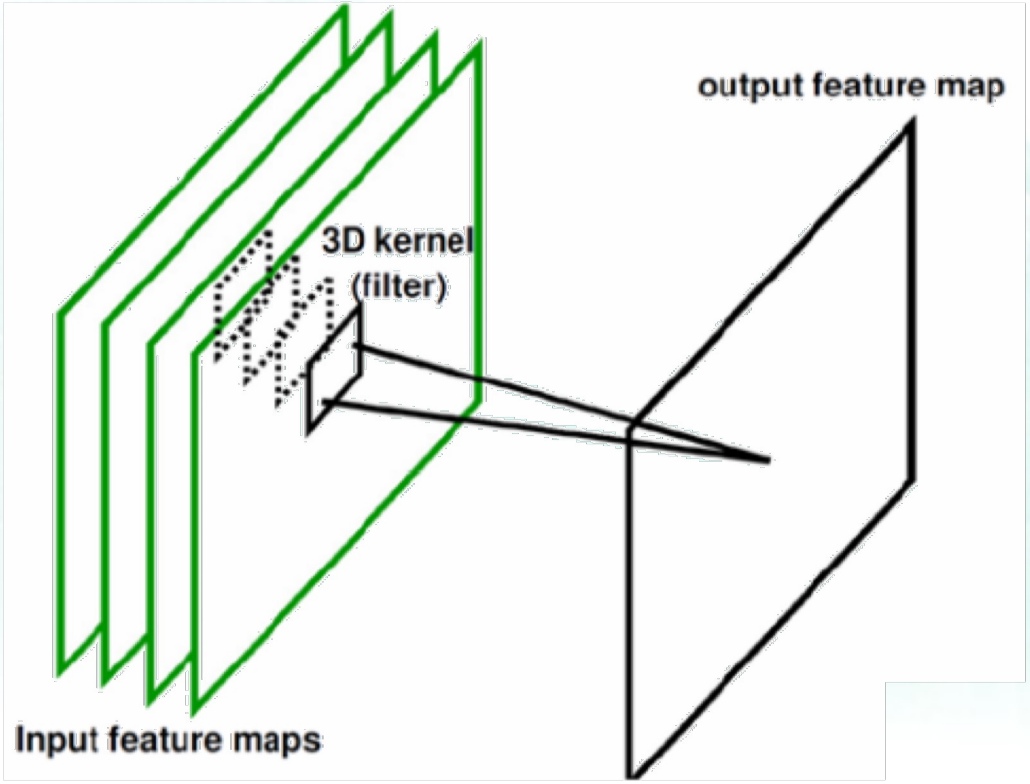






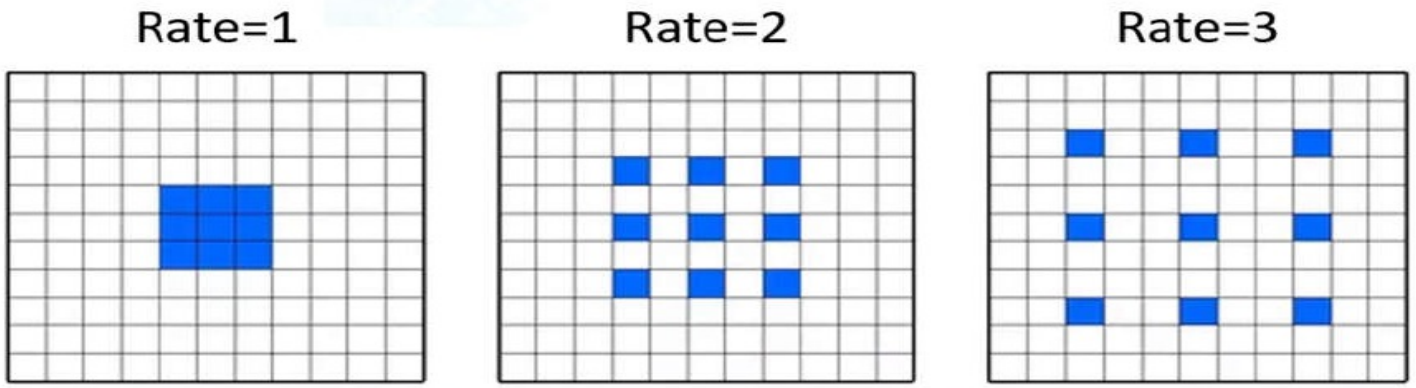
卷积神经网络： 3D卷积

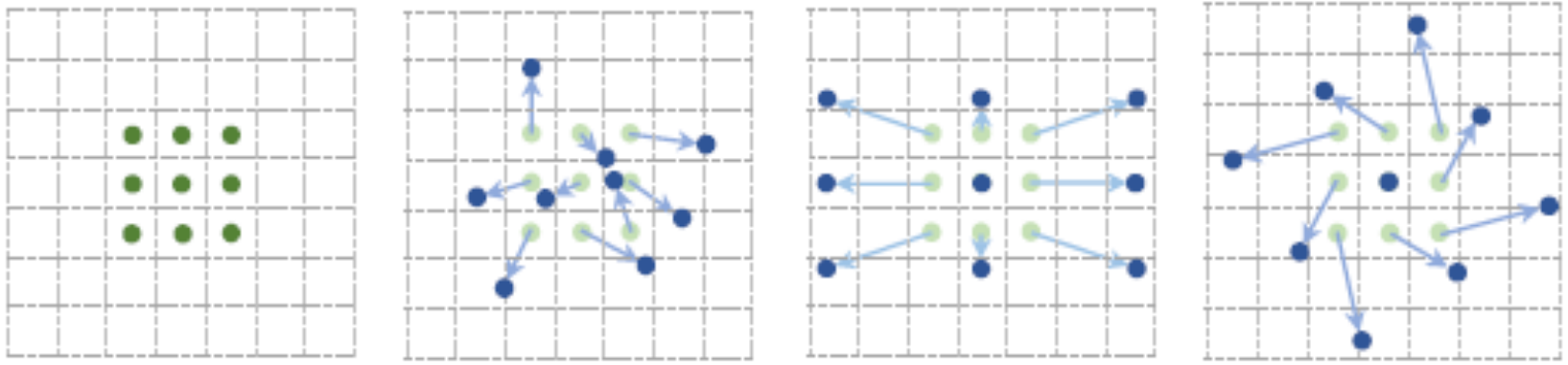
- 对多层特征图（多通道图像、视频）



膨胀卷积

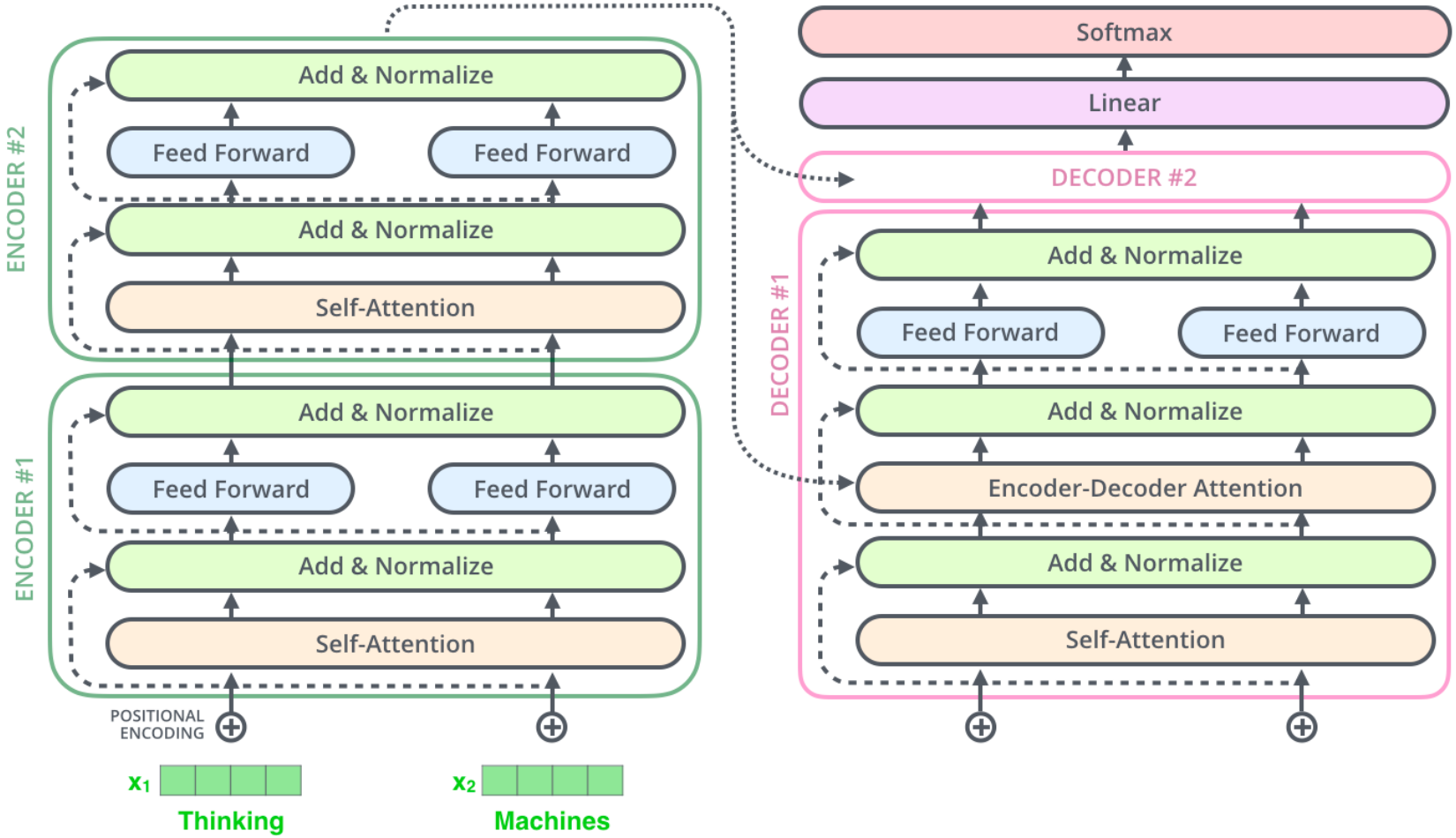
- Dilated (Atrous) convolution:不减小特征图尺寸情况下提取更大尺度的图像特征
- 拓展：Deformable CNN





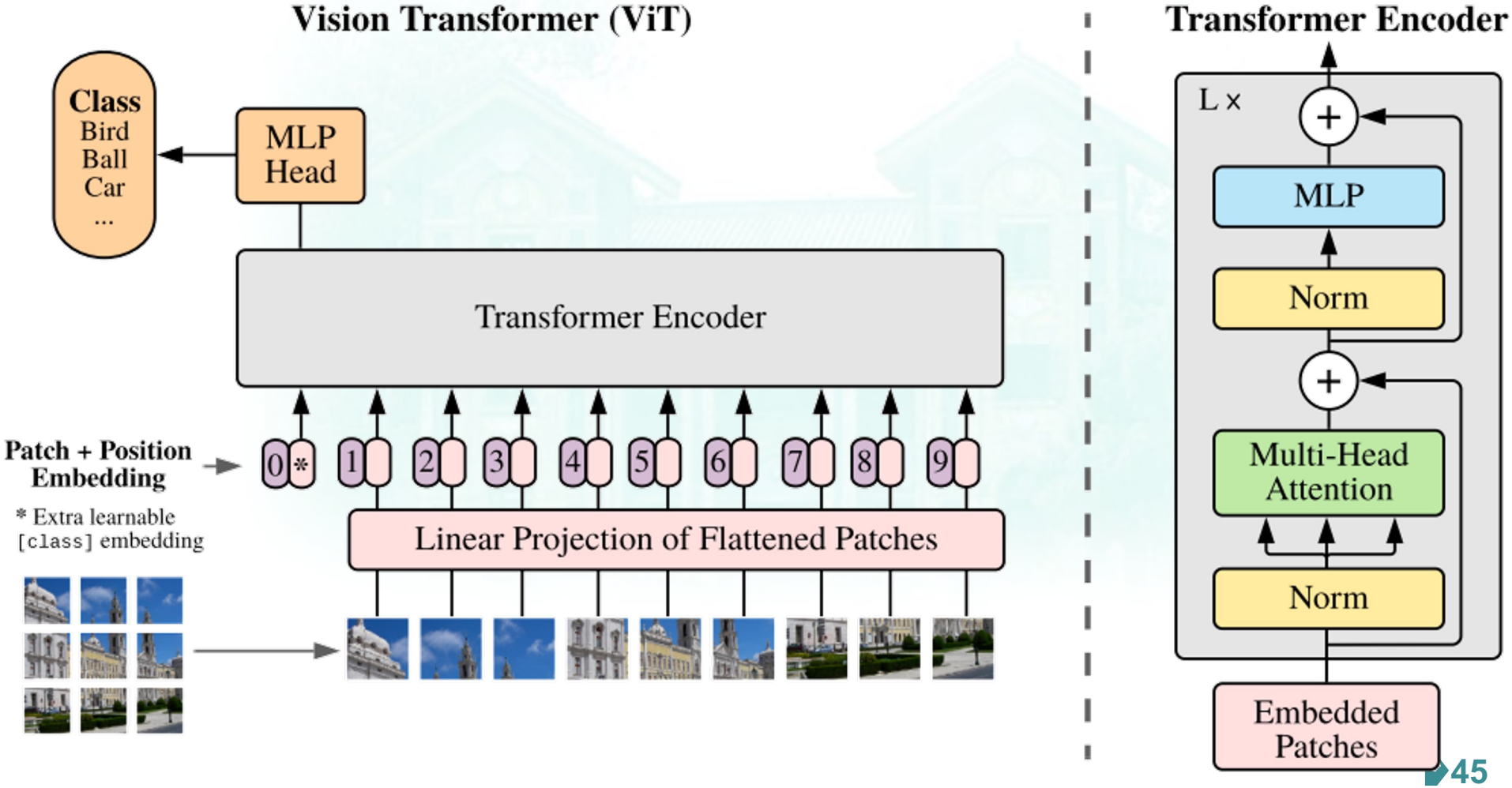
Transformer

- Encoder-Decoder架构
- 位置编码，自注意力，多头注意力



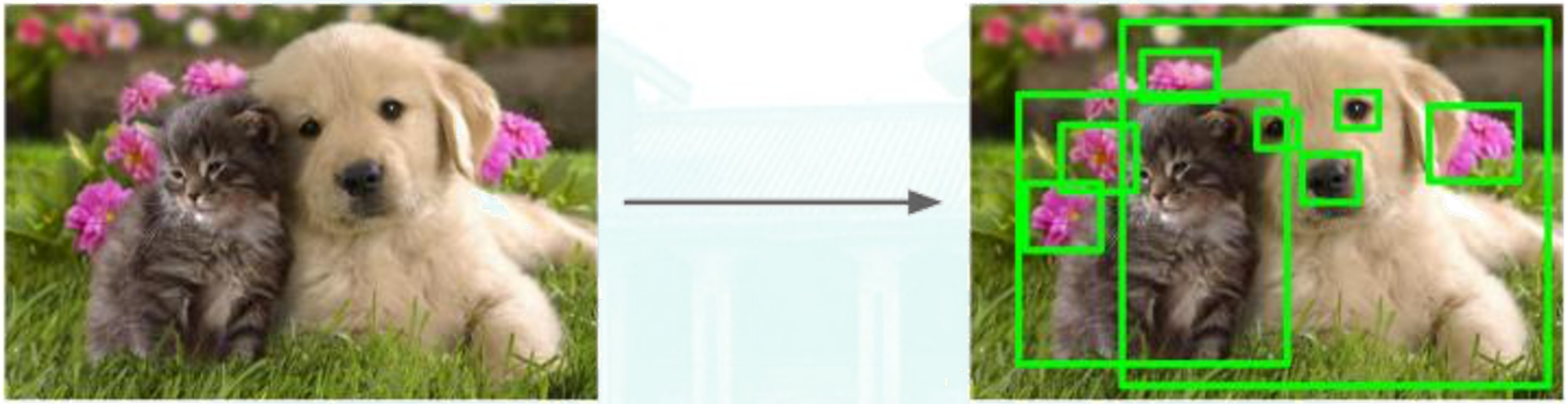
视觉Transformer

- 图片拆分成16x16个patches
- 线性变换降维嵌入位置信息，输入Transformer



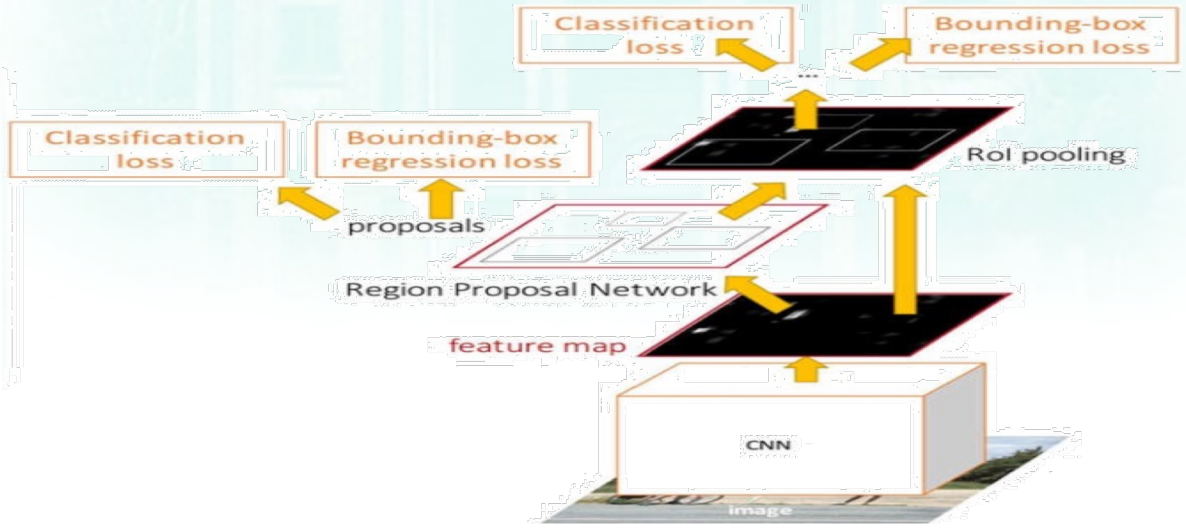
物体检测

- 锚点的选择严重影响到效率（多尺度滑窗？）
- 出现一个相关研究：似物检测(Objectness、Proposal)



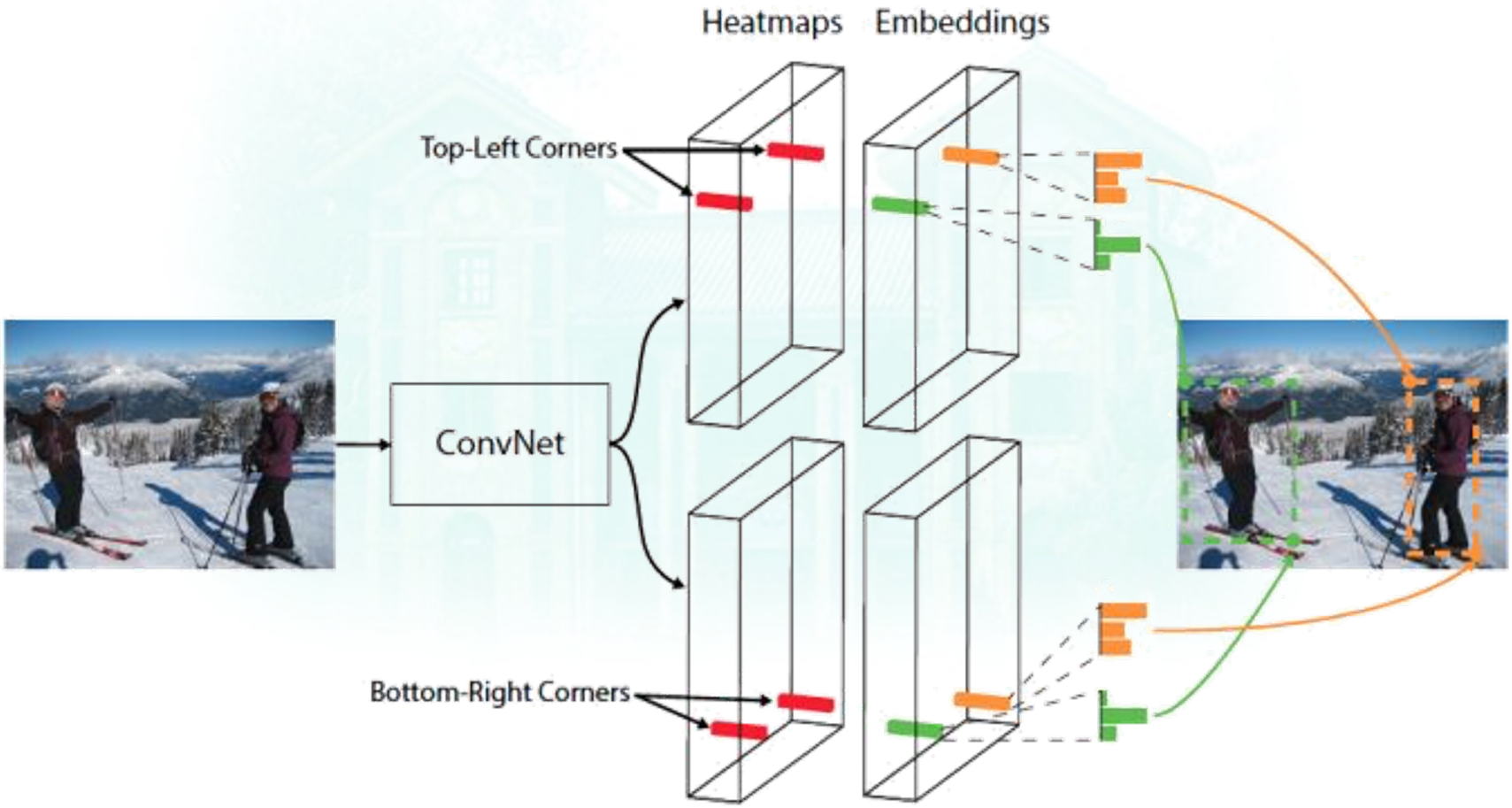
用于目标检测的神经网络（基于锚点的方法）

- 代表：R-CNN、Fast R-CNN、Faster R-CNN、YOLO、SSD
- 有效机制：
- 采用多层特征图进行预测（本质为局部特征+全局特征）
- 采用不同尺度和长宽比的Anchor
- 多任务学习：分类 + 位置回归
- 设计专门的似物检测网络(如Region Proposal Network)或者密集采样（如SSD）。



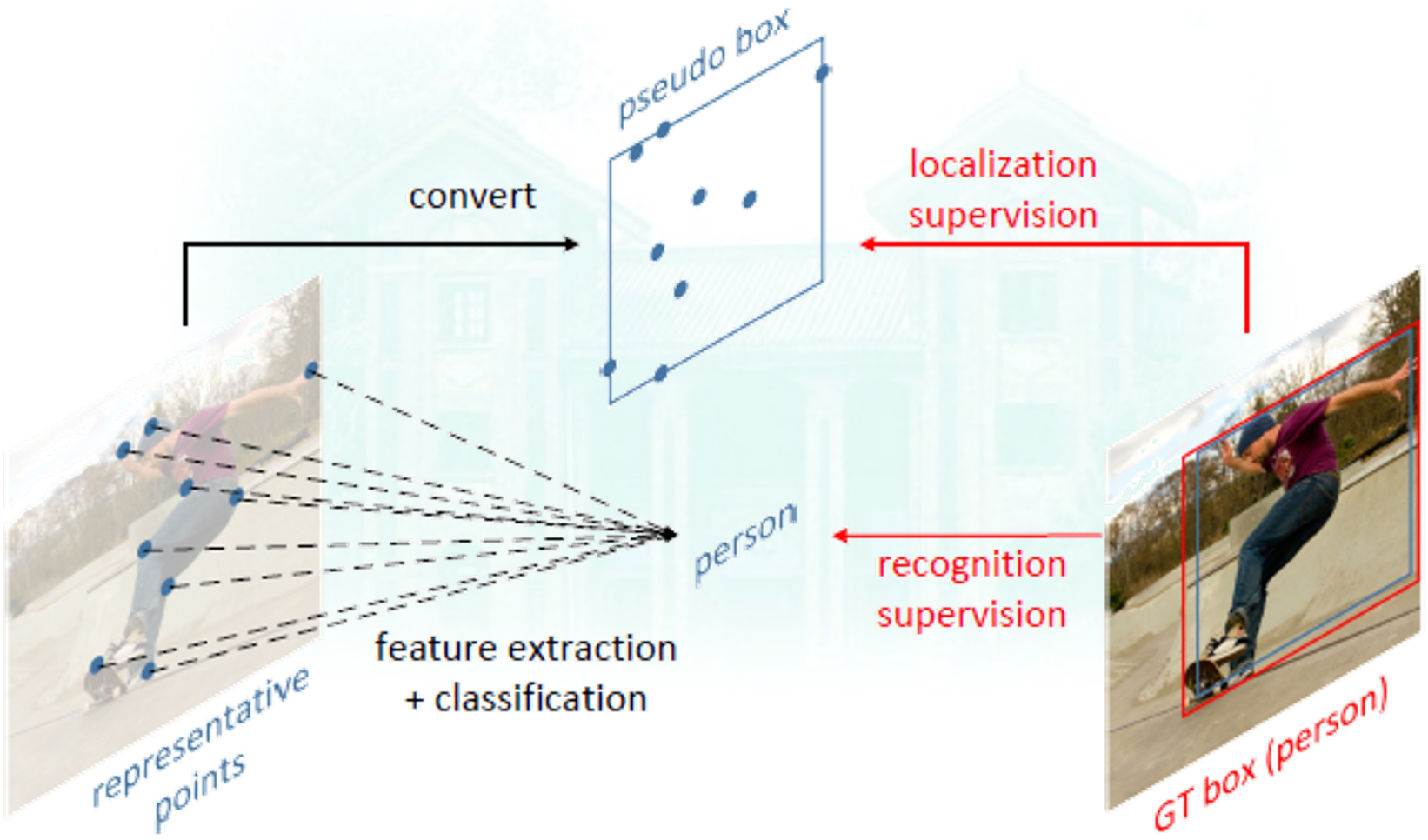
目标检测之CornerNet

- 一步法：舍弃传统的 anchor boxes思路，提出预测目标边界框的左上角和右下角一对顶点，使用单一卷积模型生成热点图和连接矢量



目标检测之RepPoints

- 一步法：舍弃传统的 anchor boxes思路，基于点集检测
- 使用可形变的卷积处理点集



目标检测之DETR

- 检测文本对应的视觉目标，多模态
- 使用transforer框架

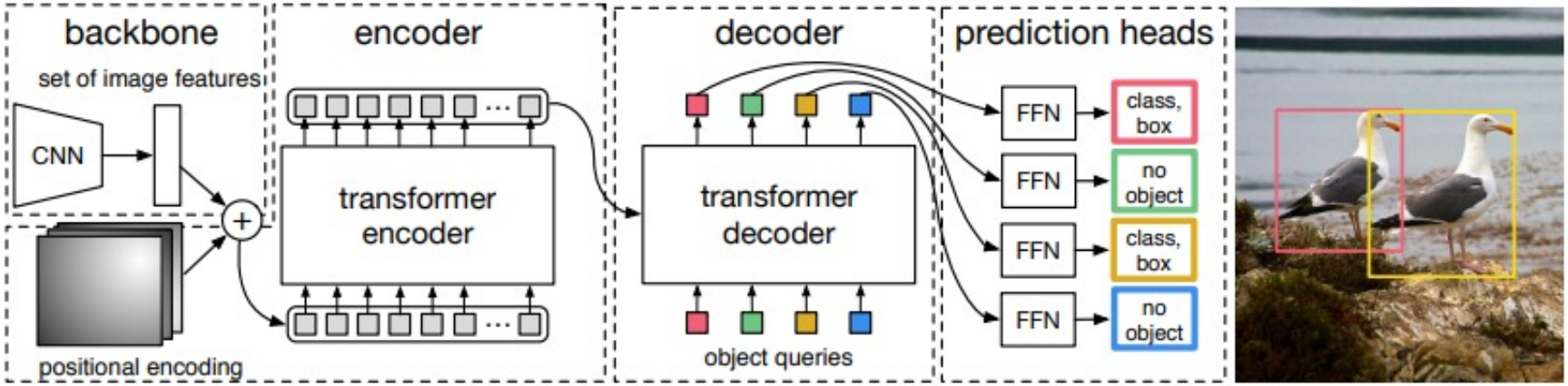
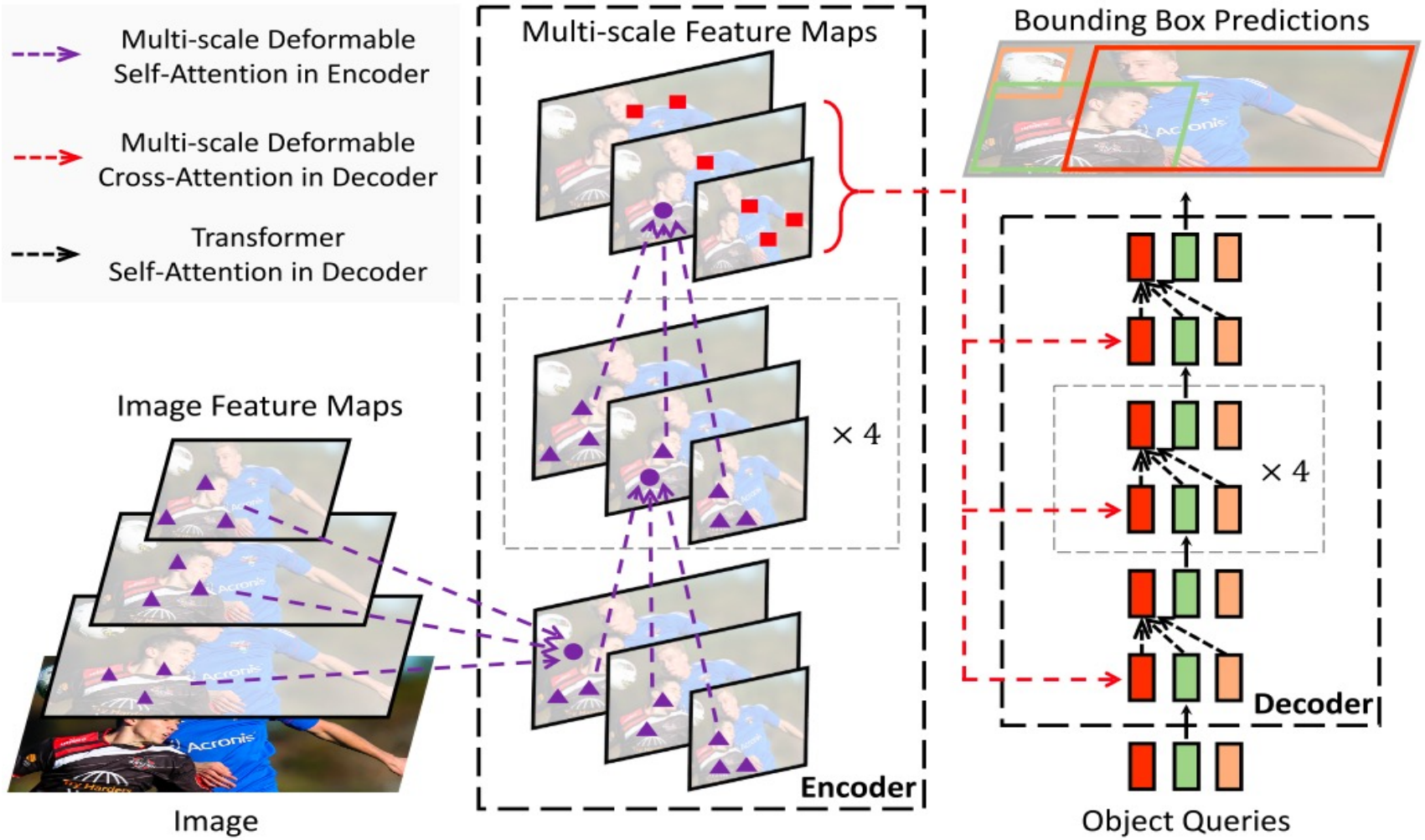


Fig. 2: DETR uses a conventional CNN backbone to learn a 2D representation of an input image. The model flattens it and supplements it with a positional encoding before passing it into a transformer encoder. A transformer decoder then takes as input a small fixed number of learned positional embeddings, which we call *object queries*, and additionally attends to the encoder output. We pass each output embedding of the decoder to a shared feed forward network (FFN) that predicts either a detection (class and bounding box) or a "`no object`" class.

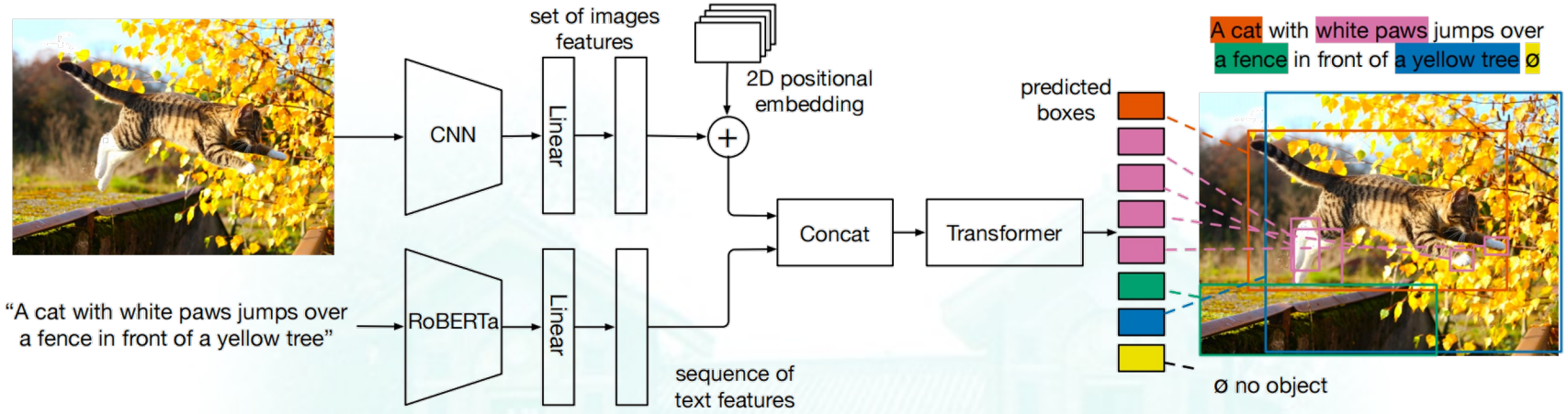
目标检测之Deformable DETR

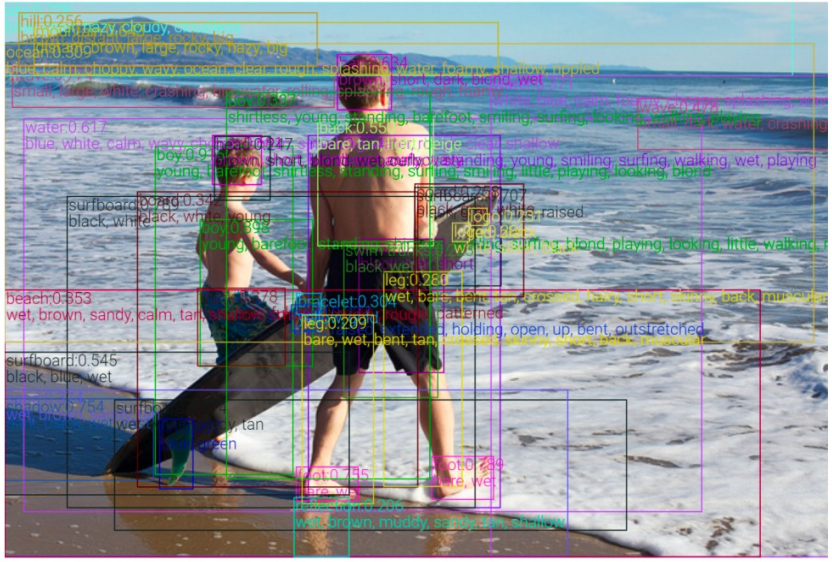
- 将DETR中的attention替换成Deformable Attention
- 收敛速度加快10倍

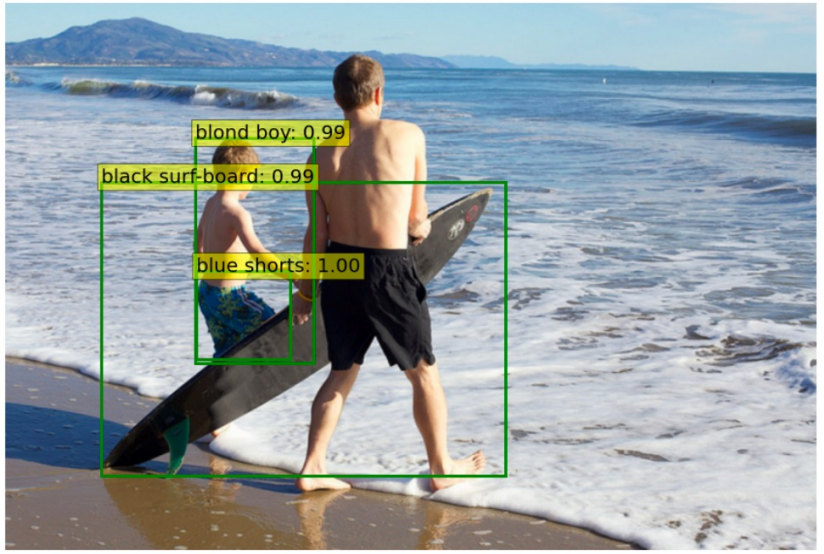


目标检测之Deformable DETR

- 多模态输入：文本+图像





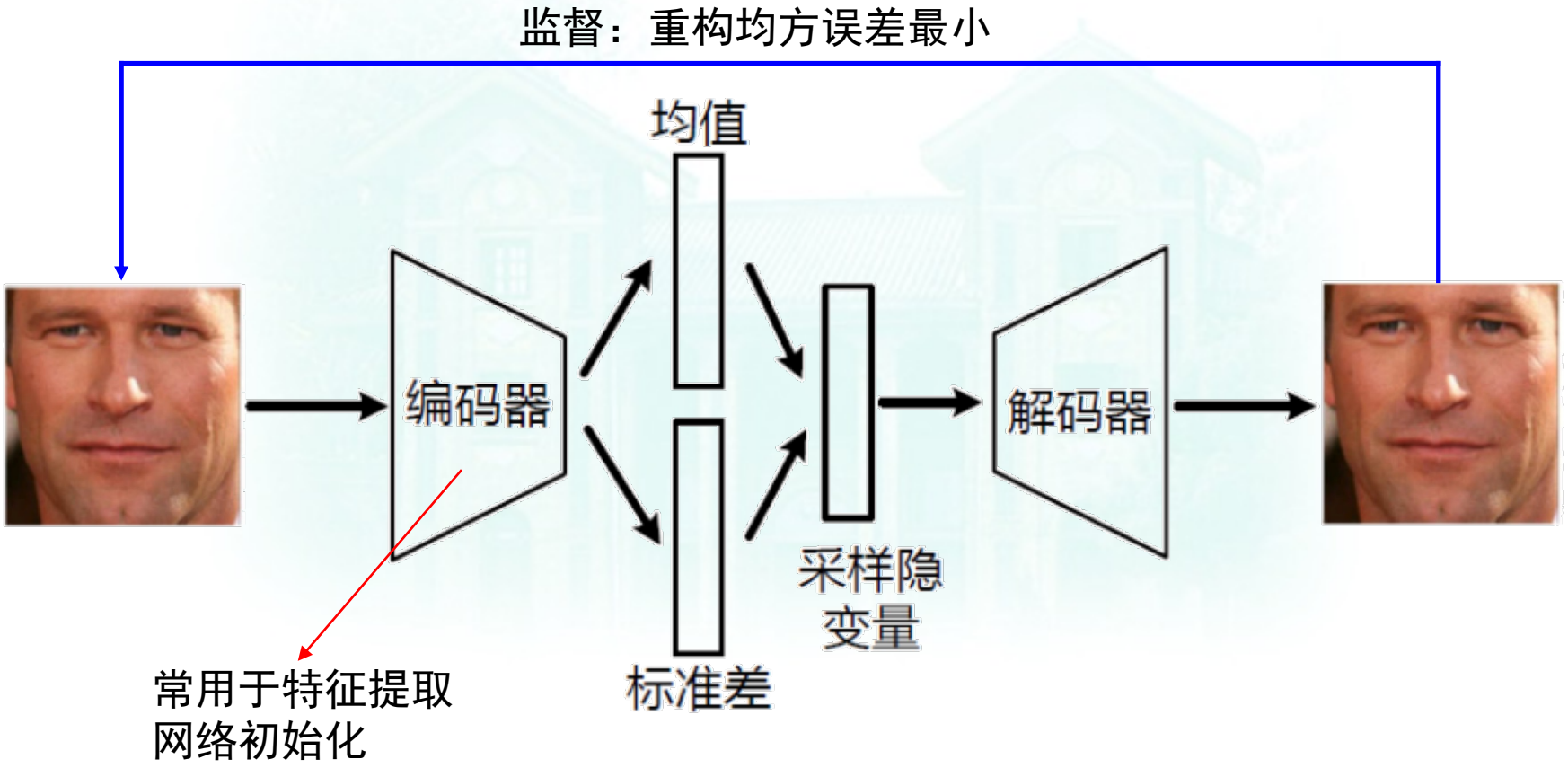


(a) Current object detection pipeline outputs, predicting all possible objects in the image. This extensive annotation is essential to multi-modal understanding systems that treat detection as a black box.

(b) MDETR predicts boxes relevant to the caption and labels them with the corresponding spans from the text. Here we use the caption: "blond boy wearing blue shorts. a black surf-board."

生成模型：Variational Auto-Encoder （VAE）

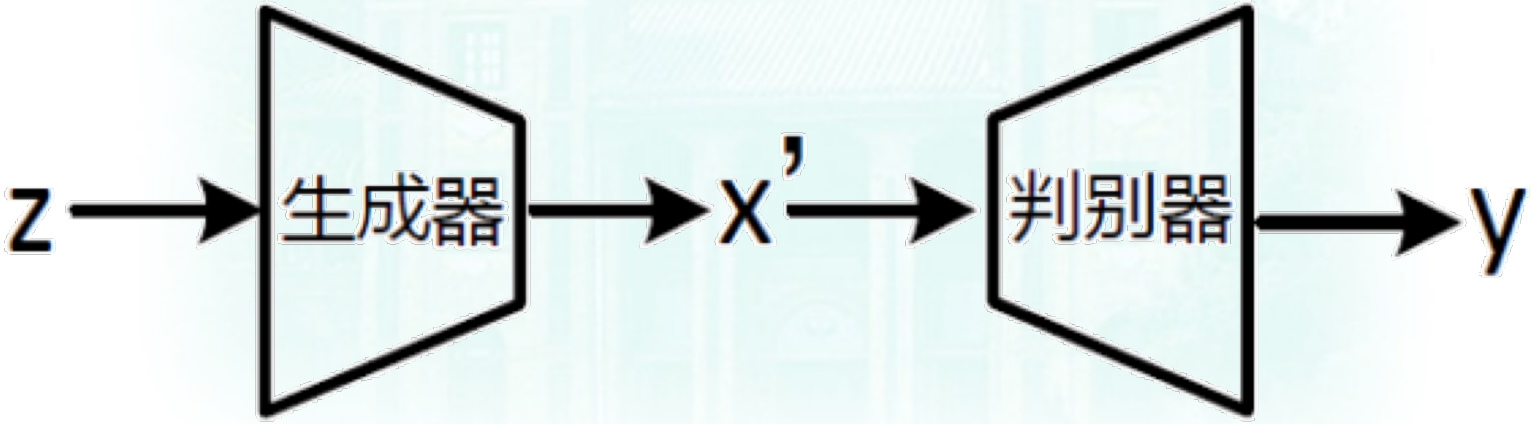
- 要求隐变量的概率分布符合高斯分布，泛化能力强
- 生成的图像比较模糊



Kingma D P, Welling M. Auto-encoding variational bayes, 2013.

Generative Adversarial Networks (GAN)

- 二元极小极大博弈问题（如：造假钞vs.验假钞）
- 随机变量通常服从高斯分布或者均匀分布
- Conditional GAN 、CycleGAN （最大优势是什么？）
- 生成的图像比较清晰，但是很容易丢失个性化的结构信息
- 容易产生模式坍塌（生成图像局限于有限ID）

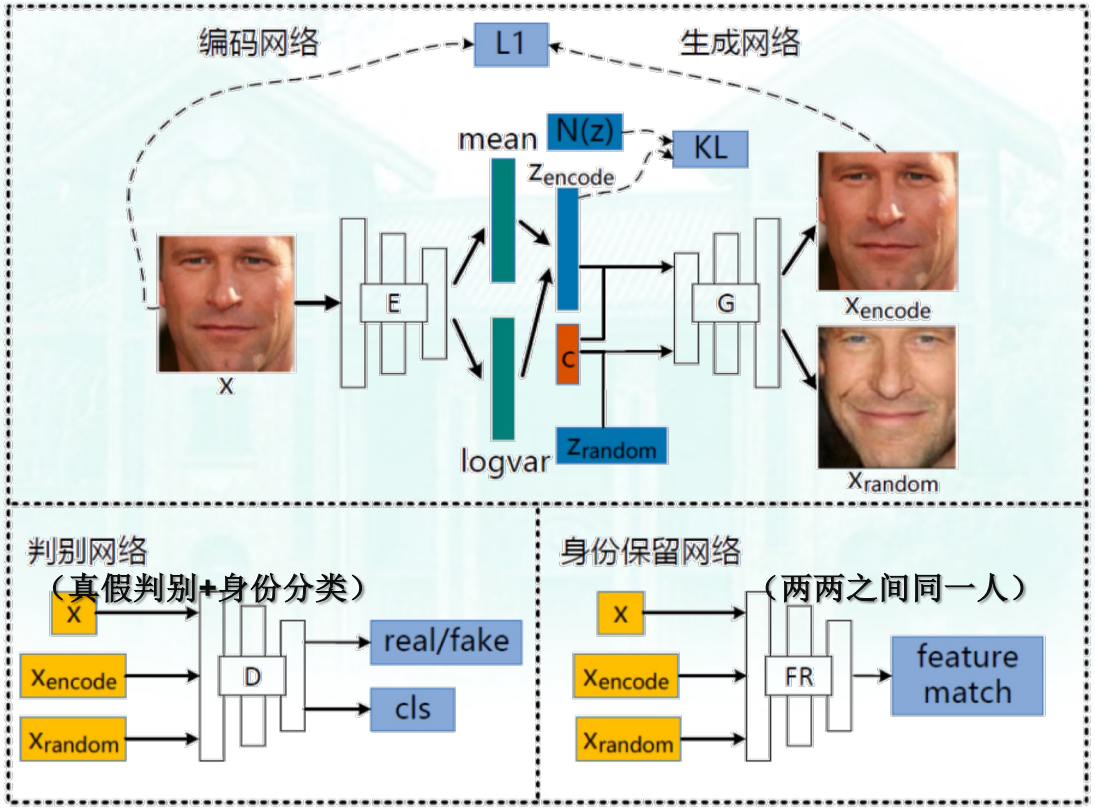


$$\min_G \max_D V(D,G) = E_{x\sim p_d(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1-D(G(z)))]$$

Goodfellow et al. Generative adversarial nets. NIPS, 2014.

GAN+VAE

- 采用感知损失（特征匹配）度量重构，稳定训练过程；
- 引入身份保留损失和人脸分类损失确保生成人脸的身份不变；
- 采用多尺度判别器保证图像质量



“重构”+“对抗”衍生CycleGan、StarGan……

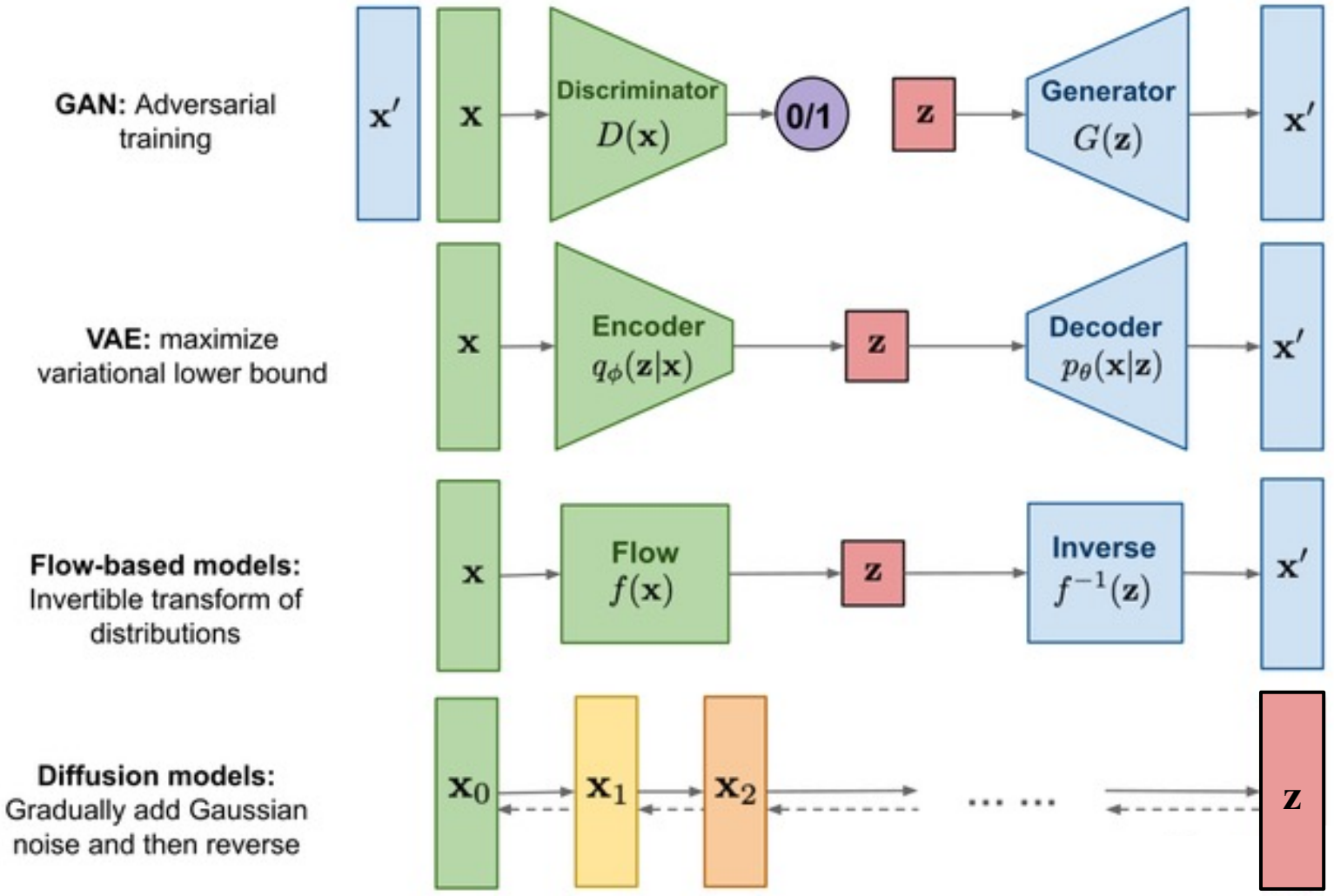
VAE+GAN的人脸生成结果

(a) 真实图像　(b) CVAE　(c) CGAN　(d) ours

分别注意表观细节和人脸轮廓

Diffusion模型

- 逐步加噪声，逐步去噪声；加噪声与去噪为互逆过程



Diffusion模型

- 逐步加噪声，逐步去噪声；加噪声与去噪为互逆过程