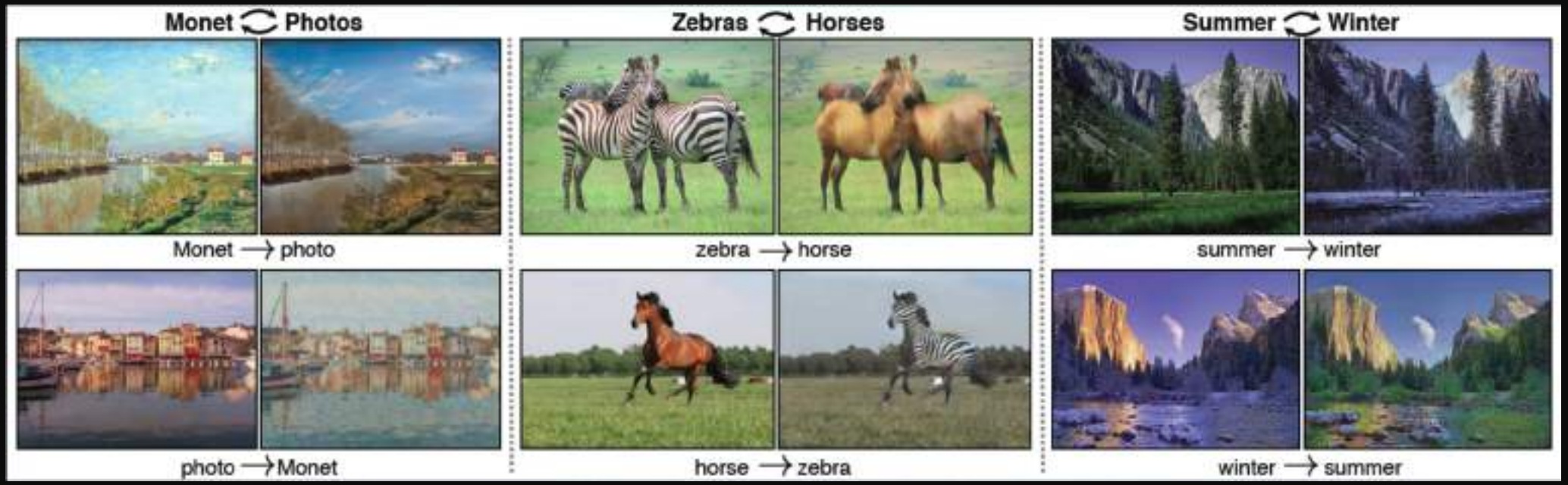
深度学习为计算机视觉带来的进步

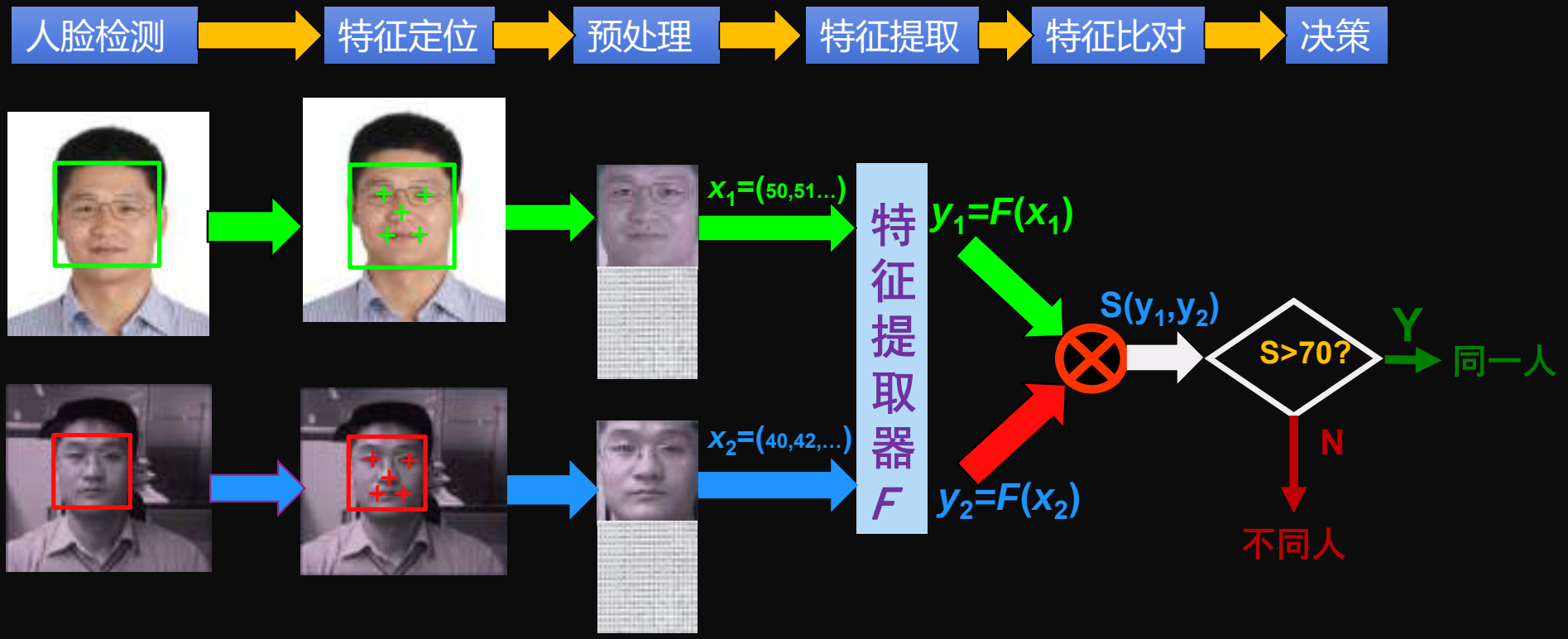
SeetaTech 中科视拓

口图像合成及风格转换



以人脸识别为例...

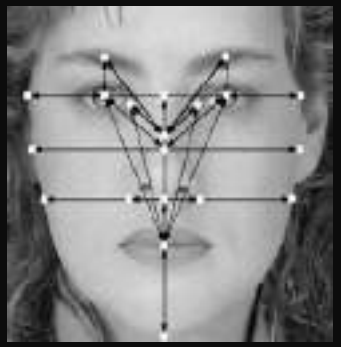
全自动人脸识别系统流程

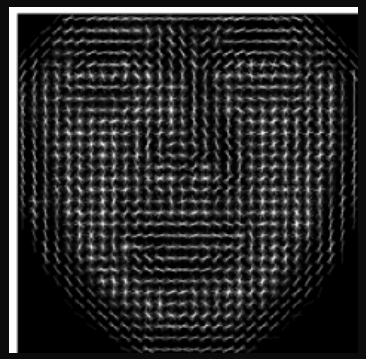


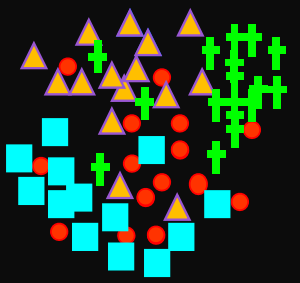
特征提取器F

SeetaTech 中科视拓

- 第一代：完全人工设计特征
  - 形状，颜色，纹理，频谱
- 第二代：(子空间)变换特征
  - PCA，LDA，LPP，SR... $y = Wx$
- 第三代：人工设计局部特征 + 变换特征
  - Gabor滤波器，LBP + PCA，LDA等 $y = W(f(x))$

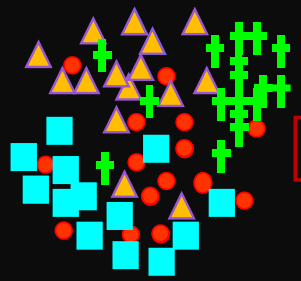






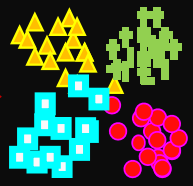
输入图像空间

特征变换



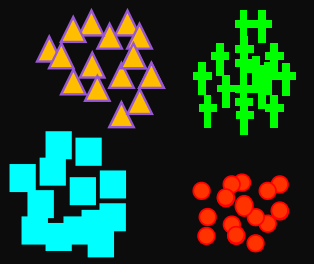
输入图像空间

局部特征提取



高维局部特征空间

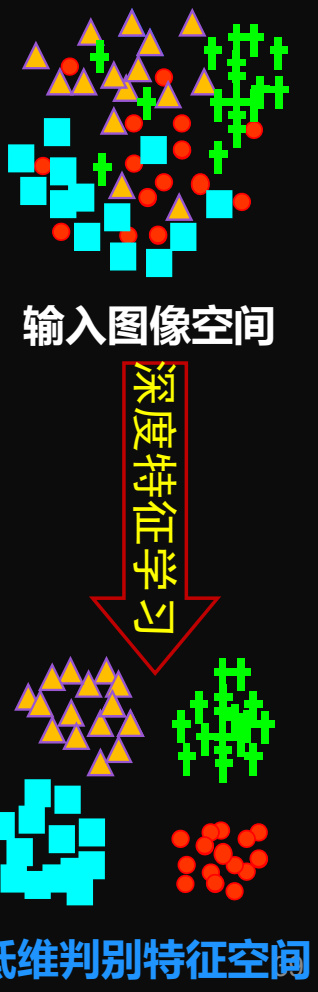
特征变换



低维判别特征空间

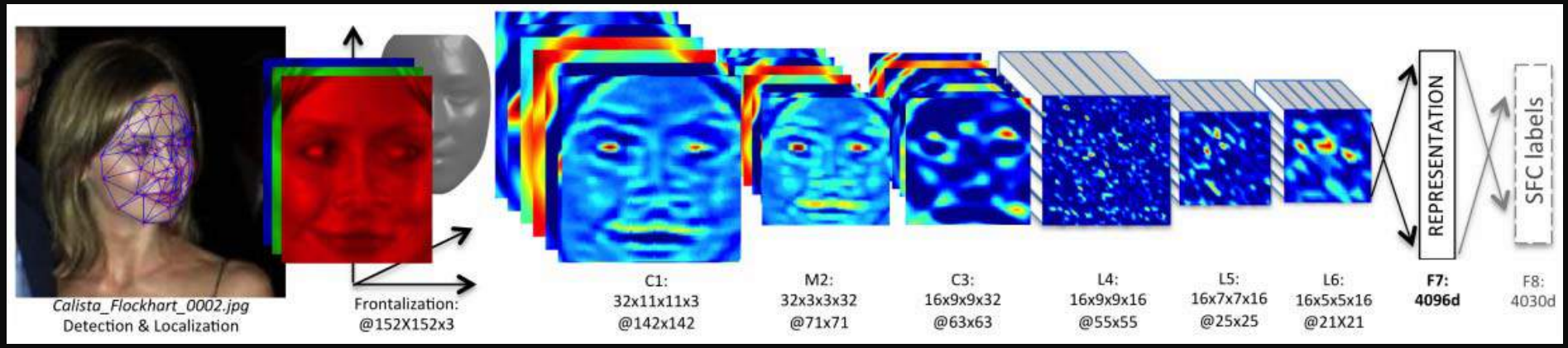
特征提取器F

- 第一代：完全人工设计特征 ——知识驱动
  - 形状，颜色，纹理，频谱
- 第二代：(子空间)变换特征 ——数据驱动(学 W矩阵)
  - PCA，LDA，LPP，SR... $y = Wx$
- 第三代：人工设计局部特征 + 变换特征 ——知识+数据驱动
  - Gabor滤波器，LBP + PCA，LDA等 $y = W(f(x))$
- 第四代：深度特征学习 ——完全数据驱动
  - 非线性
  - 局部特征参数可学习



DeepFace(Facebook)

- 8层网络，人脸3D正面化预处理
- 训练数据：4K人，4.4M图像

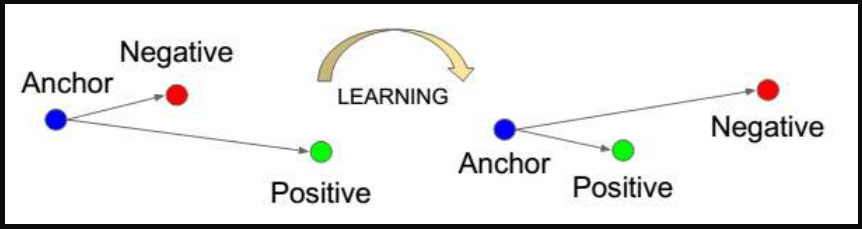


Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification. CVPR, 2014.

FaceNet(Google)

□GoogleNet(22层)+海量数据(8M人,2亿张图像)+Triplet Loss

- [F. Schroff, D. Kalenichenko, and J. Philbin, CVPR15]



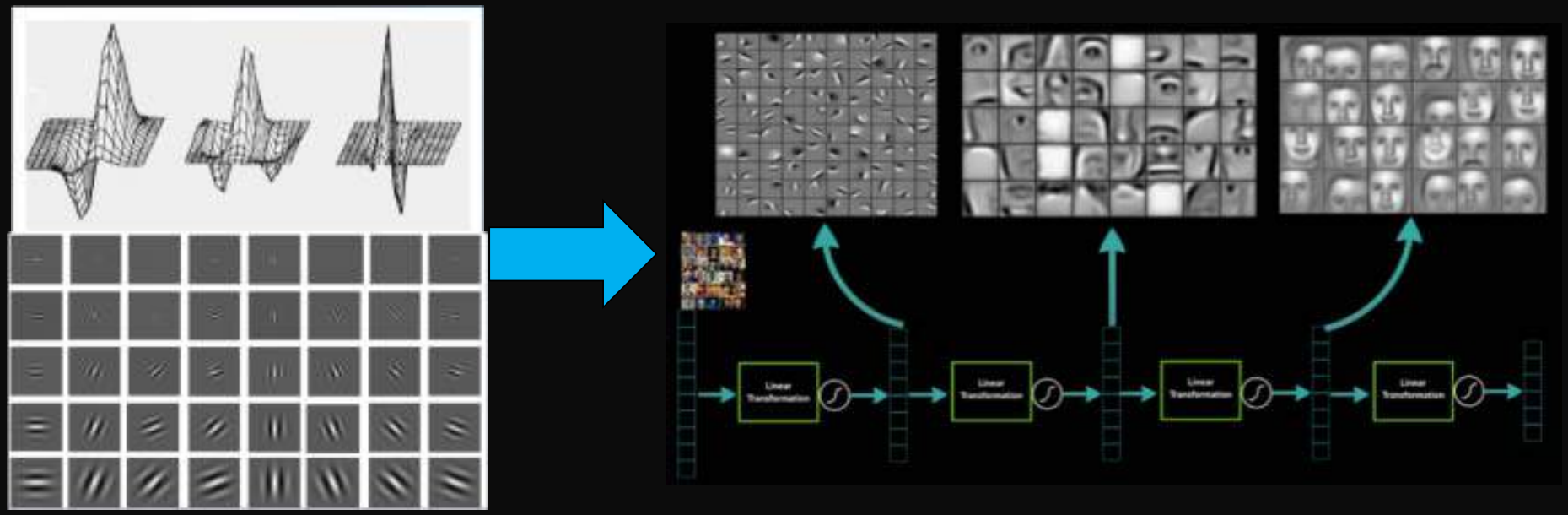
$$\sum_{i}^{N}\left[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha\right]_+$$

从特征设计到特征学习

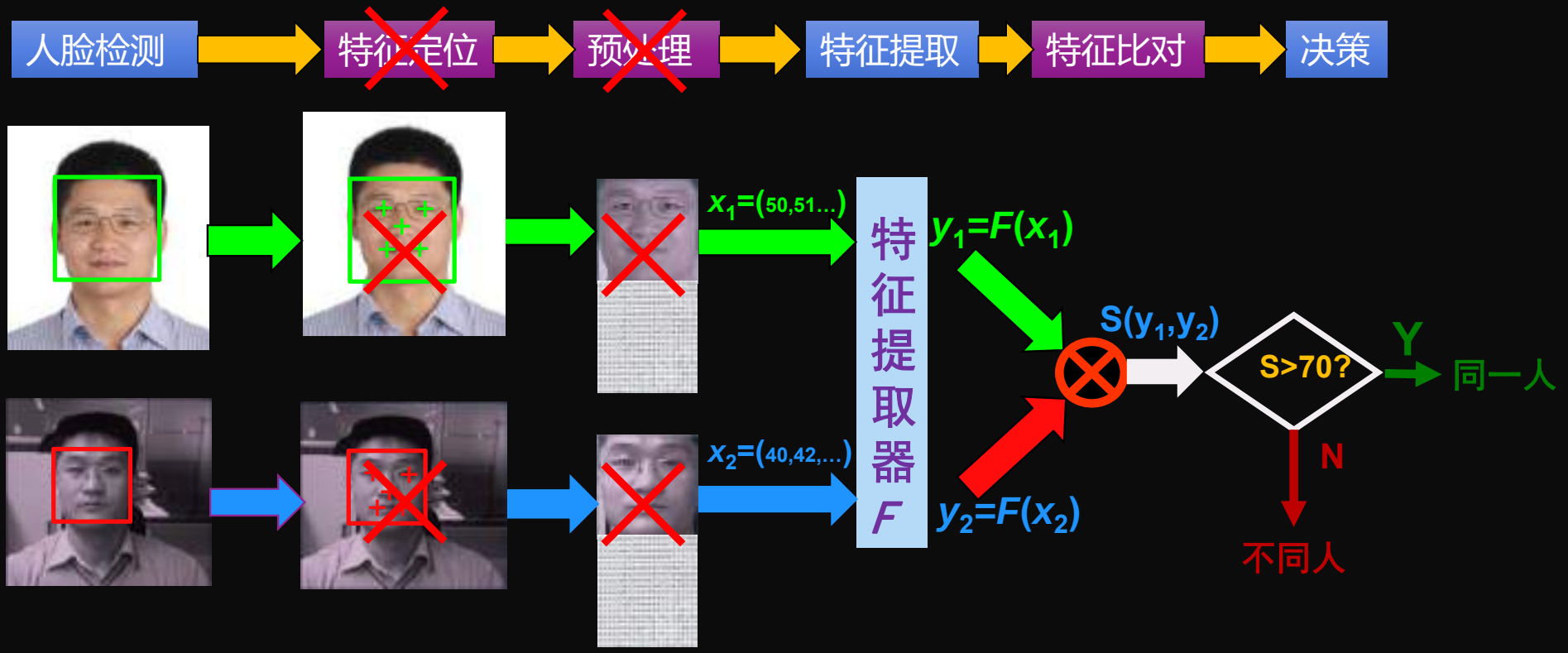
SeetaTech 中科视拓

□本质上是层级抽象的滤波型局部特征

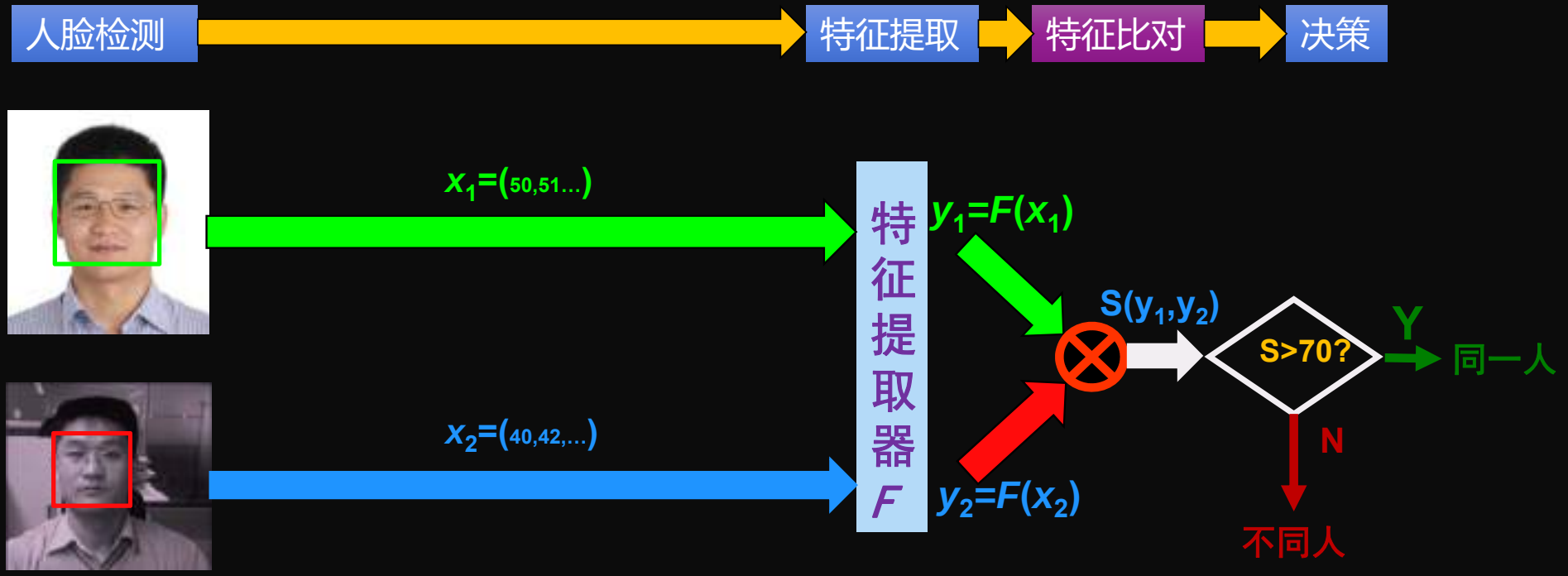
- ■与之前局部特征的不同
  - □Gabor：权值固定，人为设定（加窗傅里叶型函数），没有目标函数
  - □CNN：数据驱动的权值学习（最有利于目标函数达成的）



全自动人脸识别系统流程



全自动人脸识别系统流程



SeetaTech 中科视拓

人脸识别错误率2-4个数量级的下降！
（iPhone X：百万分之一的错误率）

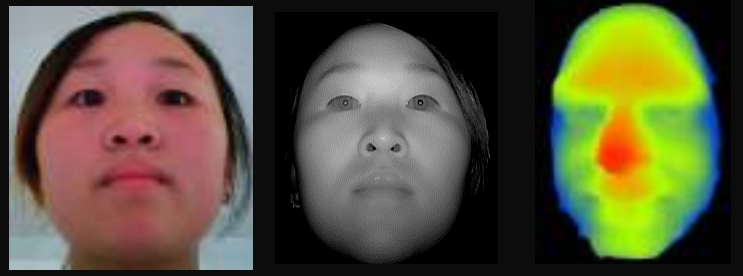
得益于百万人、亿级规模人脸图像的特征学习！

人脸识别技术进展——从iPhone X说起

SeetaTech 中科视拓

口大概的精度情况

- ■误识率：百万分之一
- ■拒识率：~5%（即通过率/识别率约95%）



口是最容易的人脸识别应用场景！

- ■RGBD传感器：彩色照片(RGB)，近红外图，深度图（距离/立体信息）
- ■注册阶段：多张多角度，近照，还可以越用越多（熟人识别）
- ■识别阶段：近景，0.3米~1米
- ■识别模式：1:1比对（不是1:N）

人脸识别技术进展——性能依赖于场景

SeetaTech 中科视拓

□场景1——1:1人证合一验证系统

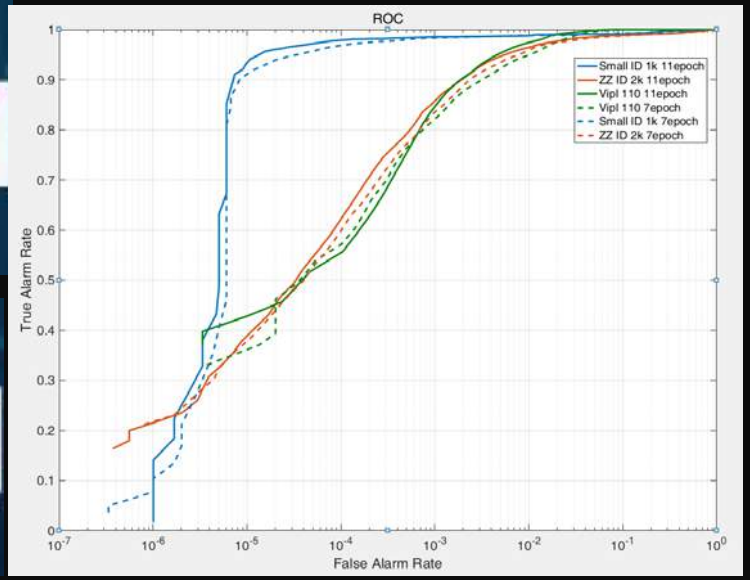
- ■1-A场景：二代证卡内照片(102*126) vs. 被识别人**配合**现场拍照
  - □精度：FAR=0.01%，验证通过率>94%——**超过人类！**
- ■1-B场景：二代证大图(358*441) vs. 被识别人**配合**现场拍照
  - □精度：FAR=0.01%，验证通过率>98%——**超过人类！**
- ■1-C场景：二代证网纹图(178*220) vs.被识别人**配合**现场拍照
  - □精度：FAR=0.01%，验证通过率>96%——**超过人类！**
- ■1-D场景：企业员工刷卡 vs. 被识别员工**配合**现场拍照
  - □精度：FAR=0.01%，验证通过率>99%

人脸识别技术进展——性能依赖于场景

SeetaTech 中科视拓

口场景1-A/C——1:1人证合一验证系统(中科视拓SeetaFace系统)

- ■TAR>94% @ FAR=0.01%



人脸识别技术进展——性能依赖于场景

SeetaTech 中科视拓

□场景2——1:N静态照片检索系统

- ■**2-A场景**：N=1000万或亿级，目标库中人脸和查询图均为**证件照**
  - □首选识别率**90%以上**
- ■**2-B场景**： N=1000万，目标库中图像为**证件照**，查询图为监控**视频帧**
  - □视频帧为质量较优的准正面截图
  - □首选识别率**~80%**【个人猜测】
- ■**2-C场景**： N=1000万，目标库图像来自监控**视频帧**，查询图片为**证件照**
  - □首选识别率**<70%**【个人猜测】
- ■**2-D场景**：N=100万，目标库人脸图和查询图均来自**生活照（新闻照）**
  - □首选识别率**~90%**（例如： MegaFace, 90%）

人脸识别技术进展——性能依赖于场景

SeetaTech 中科视拓

□场景3——1:N+1动态人脸识别系统

□N=10000，注册照片质量可控

- ■3-A场景：被识别人配合
  - □例：单位无卡考勤/门禁系统
  - □**误识率<1%**时，首选识别率**>98%**
- ■3-B场景：无感用户
  - □例：VIP识别系统（被识别人既不配合也不刻意回避）
  - □**误识率<1%**时，首选识别率**70%-90%**
- ■3-C场景：被识别人不配合、甚至刻意回避
  - □例：黑名单布控系统
  - □**虚警率<1%**时，首选识别率**<80%**

人脸识别技术进展

□场景3-A——刷脸考勤或脸控闸机(中科视拓SeetaFace系统)

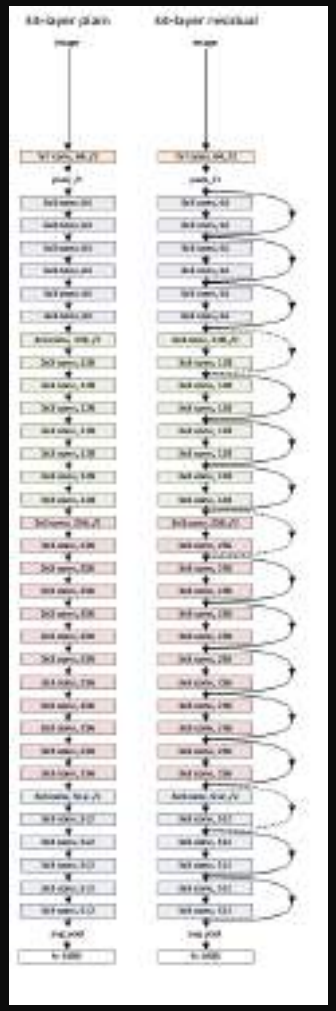
■N=1万人，注册照片质量可控，被识别人基本配合

趋势和问题

SeetaTech 中科视拓

- **开放环境下，如何确保识别的鲁棒性?**
- **活体检测（截防假体攻击）：安全性极高的情景，RGBD相机!**
- **极端情况——即使是少数熟人的识别也成问题**
  - 光照：背光，强光，侧光...
  - 成像：低分辨率，运动模糊，视角...
  - 表情和遮挡：夸张表情，墨镜，口罩，刘海...
- **黑名单监控场景下的精度挑战**
  - **人员规模：1万→10万→100万→...**
  - **首选识别率：80%→90%→95%→**
  - **虚警率要求：1%→0.1%→0.01%→0.001%→**

SeetaTech
中科视拓



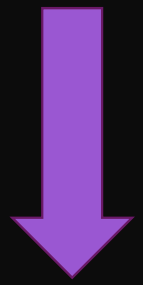
A

B

C

SeetaTech 中科视拓

人类专家知识驱动的AI方法论



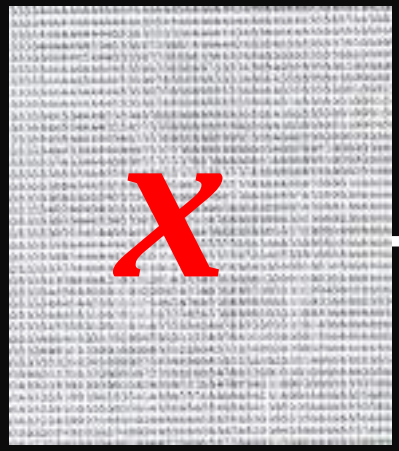
有监督大数据驱动的AI方法论

AI=A+B+C

VAI=A+B+C的范式改变了什么?

SeetaTech 中科视拓

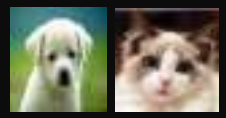
- 推动了一大类**非线性映射函数学习问题**的解决



x → $F(x)$ → y

y 可以是：

- **类别标签（分类/识别）**：{dog, cat, horse…}



- **向量/图像（回归/预测）**

深度学习带来的思想变迁

SeetaTech 中科视拓

□解决非线性问题的方法

- Kernel方法：黑盒子；有限种类核函数；核函数不可学
- 流形学习：非显式变换；分段线性；Novel样本不适用

□人工编码知识到从数据中学习知识

- Deep Learning = Feature Learning
- 通用特征(Gabor, HOG, SIFT…)不适应特定问题，面向特定问题的特征设计困难

□从分而治之到全盘考虑

- Divide and Conquer → End-to-end
- 子问题最优未必全局最优，各个步骤最优未必全局最优

□数据智能：从重算法到重数据

- 经验从数据中来，知识从数据中来

SeetaTech
中科视拓

从客户的狗屎检测需求说起...

前深度学习时代，我们这么做...

前深度学习时代，我们这么做...

SeetaTech 中科视拓

- 步骤1：花几个月时间收集并标注几百张便便图
- 步骤2：花几个月观察便便图，并绞尽脑汁选择或设计一些特征
  - 形状，颜色，纹理；SIFT，HOG，Gabor，LBP，Haar...
- 步骤3：用某种分类器训练和测试，结果不好回到步骤2

+ 专家知识驱动的特征设计 + 专家选择的的分类器

前深度学习时代，我们这么做...

SeetaTech 中科视拓

□步骤1：花几个月时间收集并标注几百

□步骤2：花几个月观察便便

特征

■形状，颜色，纹理

□步骤3：

骤2

需要多久？1年甚至更久！
人脸检测用了20年！
行人检测用了10年！

知识驱动的特征设计

+

专家选择的的分类器

深度学习时代，我们这么做...

深度学习时代，我们这么做...

SeetaTech 中科视拓

- 步骤1：花4个星期时间收集并标注(框出狗屎位置)数万张便便图
- 步骤2：花1个星期，挑几个深度模型，选几组模型超参数
- 步骤3：交给机器绞尽脑汁优化学模型中的数千万/数亿权重参数

\+ 专家选择深度模型 + 机器优化深度模型

深度学习时代，我们这么做...

SeetaTech 中科视拓

- 步骤1：花4个星期时间收集并标注数[illegible]
- 步骤2：花1个星期，挑几个[illegible]超参数
- 步骤3：交给机器[illegible]的数千万/数亿权重参数

需要多久？2个月

+ 专家选择深度模型 + 机器优化深度模型

SeetaTech 中科视拓

后深度学习时代，我们怎么做？

后深度学习时代，我们怎么做?

后深度学习时代，我们期待这么做

SeetaTech 中科视拓

- 步骤1：花几分钟时间收集并标注几张便便图
- 步骤2：交给机器绞尽脑汁完成任务

+

机器选择和优化模型

后深度学习时代，我们期待这么做

- 步骤1：花几分钟时间收集并标
- 步骤2：交给机器

需要多久？

器选择和

优化模型

后深度学习时代，我们期待这么做

SeetaTech 中科视拓

- 步骤1：花几分钟时间收集并标注
- 步骤2：交给机器经[illegible]

需要多久？
1个星期？
几个小时？

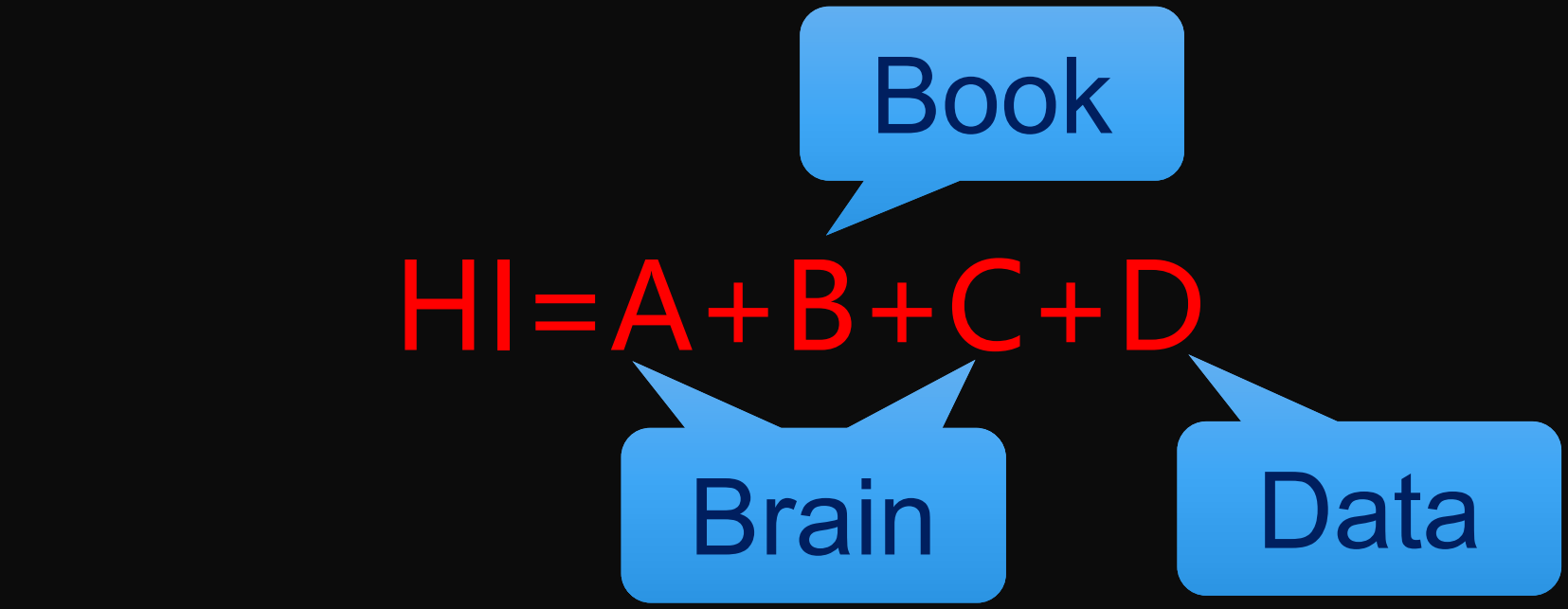
[illegible]器选择和
优化模型

想想人类智能HI...

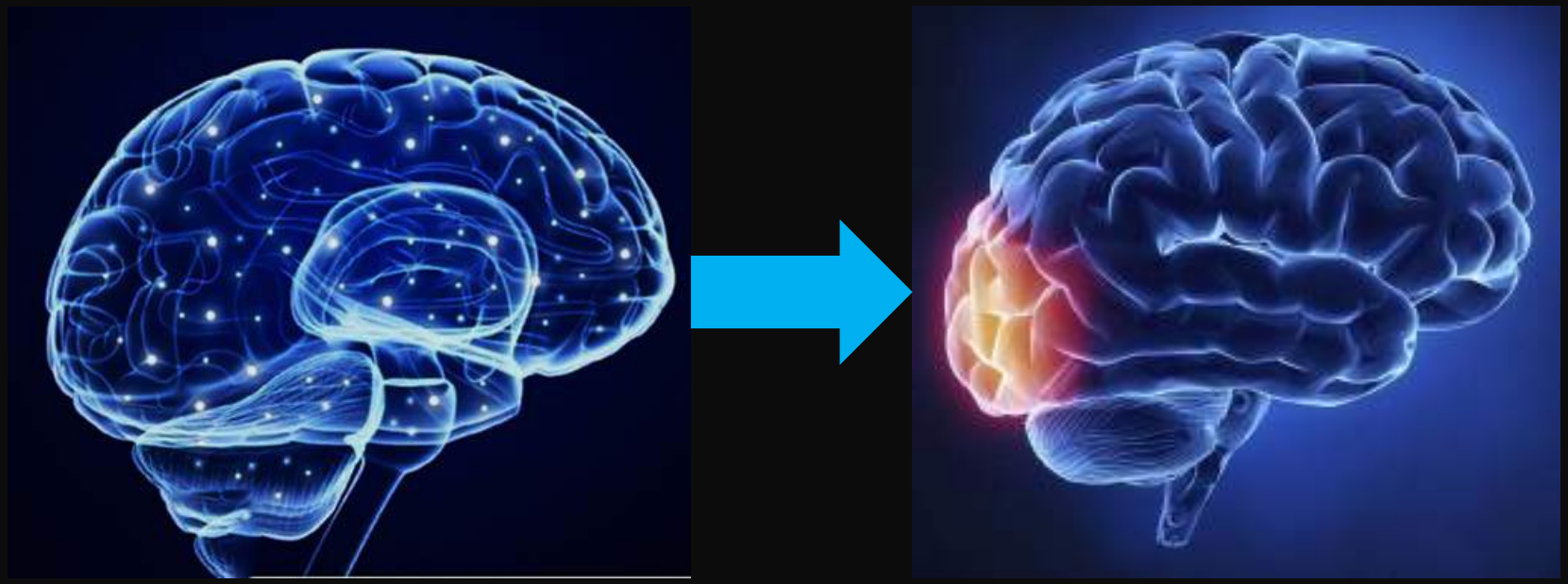
SeetaTech
中科视拓

需要多久？
几分钟？几秒钟？

想想人类智能HI...

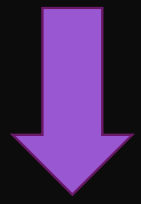

Book
HI=A+B+C+D
Brain
Data

3岁前的脑是大数据学习(历代祖先)的结果，是进化脑！3岁后个体脑的后天发育是利用小数据和知识对进化脑进行适应性修改的过程！

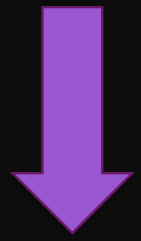


SeetaTech 中科视拓

人类专家知识驱动的AI方法论



有监督大数据驱动的AI方法论



知识与数据联合驱动的AI方法论

SeetaTech 中科视拓

但是...

计算机视觉远不是一个解决了问题

SeetaTech 中科视拓

□什么问题已经解决？数据富饶问题，靠ABC基本解决

- ■通用目标检测问题：百类→千类→万类→十万类
- ■通用目标分类/识别：千类→万类→十万类→百万类
- ■Fine-grained识别（人脸/花鸟虫鱼）：千类→万类→十万类→百万类…
- ■语义分割：几十类→百类→千类
- ■深度与3D：通用条件下的深度计算基本解决？（更多靠传感器的进步）
- ■视觉QA及推理：概念，更多知识特别是常识的嵌入
- ■从视觉到语言：概念，更多知识特别是常识的嵌入
- ■多模态的协同：视、听、语言、触觉、嗅觉

计算机视觉远不是一个解决了问题

SeetaTech 中科视拓

□什么问题已经解决？数据富饶问题

- ■通用目标检测问题：百类→千类
- ■通用目标分类/识别：千类
- ■Fine-grained识别：千类→十万类→百万类...
- ■语义分割
- ■深度
- ■本解决？（更多靠传感器的进步）
- 知识特别是常识的嵌入
- 更多知识特别是常识的嵌入
- ■视、听、语言、触觉、嗅觉

Scalability问题！

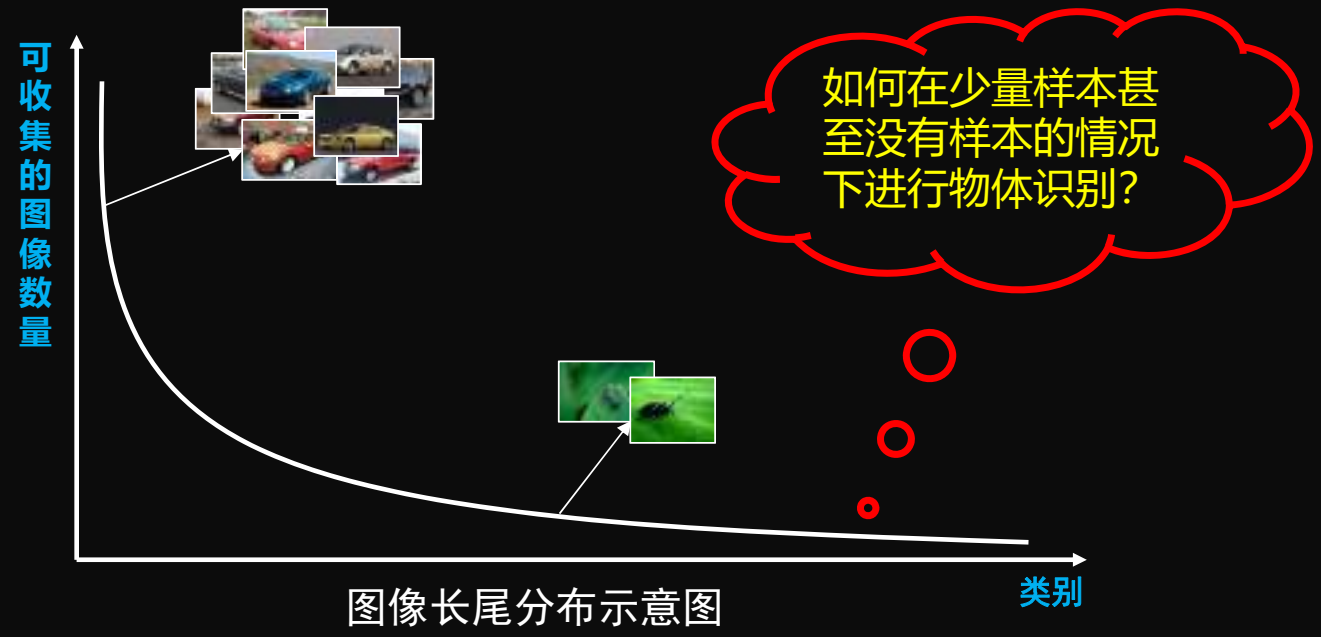
（如果不靠数据暴力来解决）

计算机视觉远不是一个解决了问题

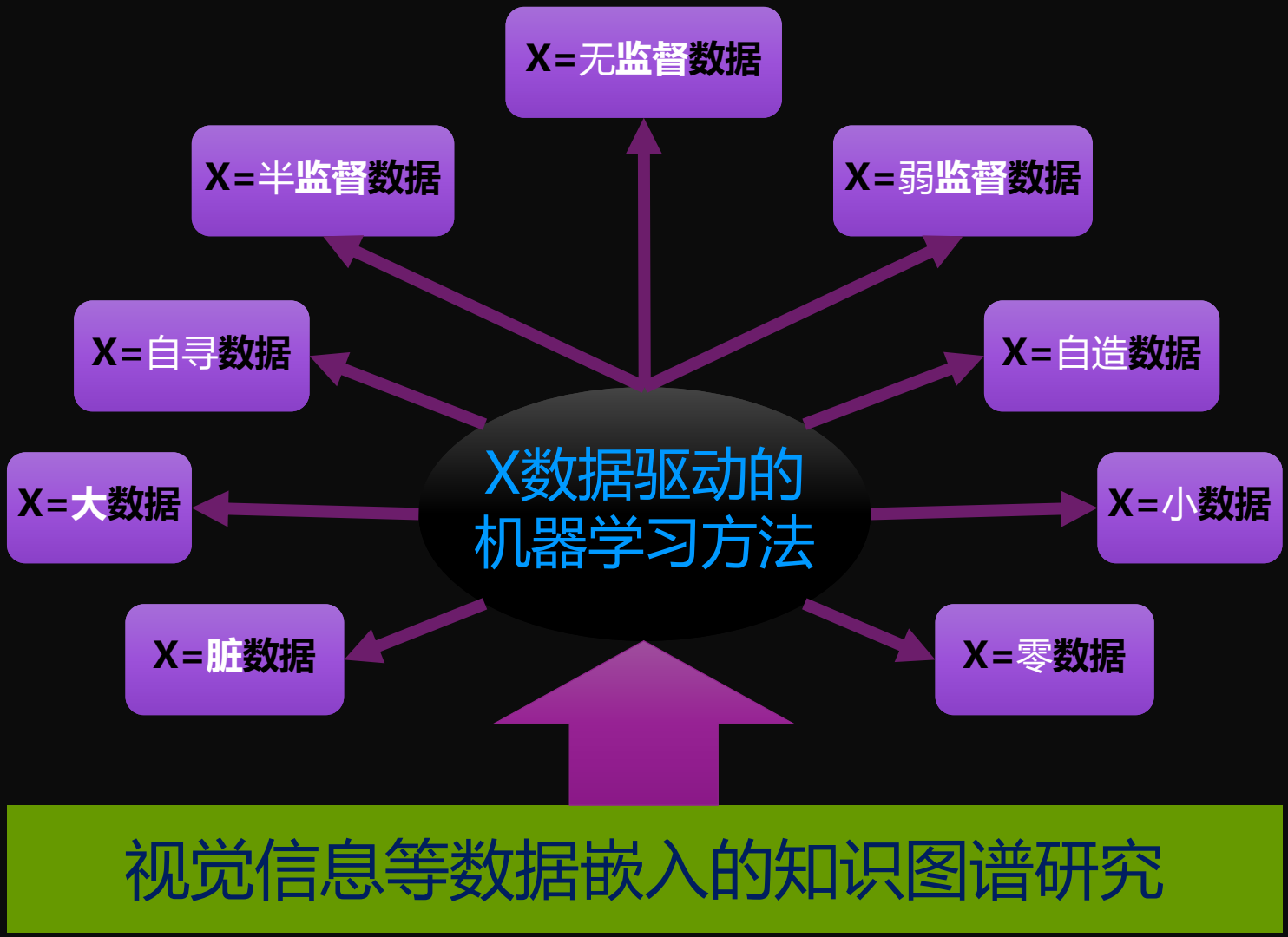
SeetaTech 中科视拓

口严重的Scalability（可扩展性）问题

- ■图像类别(至少3万类)不计其数，针对每类收集图像不切实际
- ■图像的数量呈现长尾分布，某些类别图像很难收集



图像长尾分布示意图


X=无监督数据
X=半监督数据
X=弱监督数据
X=自寻数据
X=自造数据
X数据驱动的
机器学习方法
X=大数据
X=小数据
X=脏数据
X=零数据
视觉信息等数据嵌入的知识图谱研究

例如：一种值得期待的解决方案

SeetaTech 中科视拓

- 步骤1：接受任务——“请给我生产一个安全帽检测引擎”
- 步骤2：完成任务——AI生产平台完成全部过程，提供完整解决

例如：一种值得期待的解决方案

SeetaTech 中科视拓

- **步骤1：接受任务——“请给我生产一个安全帽检测引擎”**
- **步骤2：完成任务——AI生产平台完成全部过程，提供完整解决**
  - 分析任务类型——检测任务；检测目标：安全帽
  - 数据收集
    - 搜索互联网得到高可靠“安全帽”图像，大量似是而非的“安全帽”图像
  - 知识收集
    - 安全帽属于帽子，颜色各异，多为圆形，经常戴在建筑工人的头部，也可能放在桌子上...
  - 自动选择算法并基于上述数据和知识进行学习（或许包括自动数据生成阶段）
  - 自动测试并返回用户所需的检测引擎

一些小样本学习方法

SeetaTech 中科视拓

- 有标注数据的自动获取和生成式“自造”
- 零样本学习：充分利用语义知识
- 小样本学习：除了结合语义知识之外
  - 充分利用大量无监督样本（半监督学习）
  - 基于辅助集的迁移学习（共享知识的迁移）
  - 基于辅助集的表示学习（表示方法的迁移）
  - 基于辅助集的元学习（学习方法的迁移）

研究方向1——从数据出发

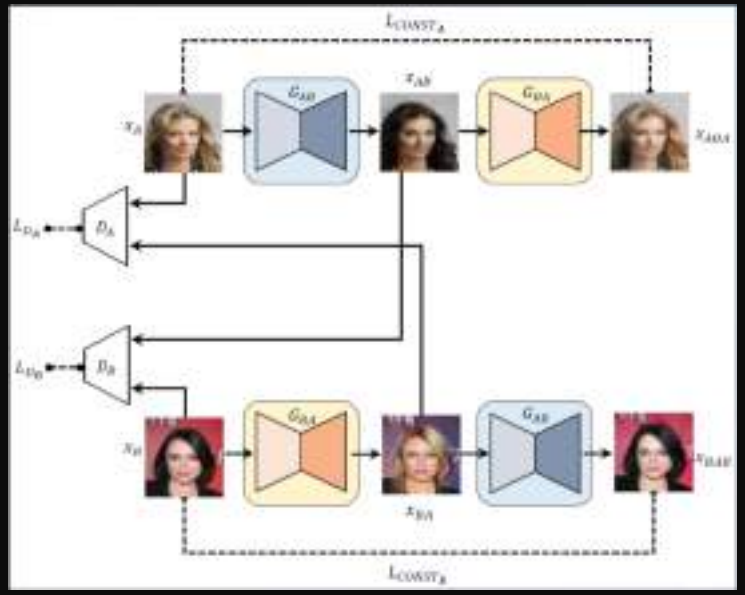
口从数据出发：从X-Data → Big Data

■图形学（物理建模）的方法举一反三：图形学重构现实世界

研究方向1——从数据出发

口从数据出发：从X-Data → Big Data

- ■图形学（物理建模）的方法举一反三：图形学重构现实世界
- ■GAN-like方法举一反三（learning-based generative models）



研究方向1——从数据出发

□从数据出发：从X-Data → Big Supervised Data

- ■图形学（物理建模）的方法举一反三：图形学重构现实世界
- ■GAN-like方法举一反三（learning-based generative models）
- ■自寻数据及无监督数据的自动标注：人类知识和其他模态的协同增效
  - □例1：通过**跟踪**获得类别不变的大量样本（用于目标检测，甚至分割任务）

研究方向1——从数据出发

□从数据出发：从X-Data → Big Supervised Data

- ■图形学（物理建模）的方法举一反三：图形学重构现实世界
- ■GAN-like方法举一反三（learning-based generative models）
- ■自寻数据及无监督数据的自动标注：人类知识和其他模态的协同增效
  - □例1：通过**跟踪**获得类别不变的大量样本（用于目标检测，甚至分割任务）
  - □例2：对唇读（视觉）而言，可以通过语音识别大量获得唇读数据

研究方向1——从数据出发

SeetaTech 中科视拓

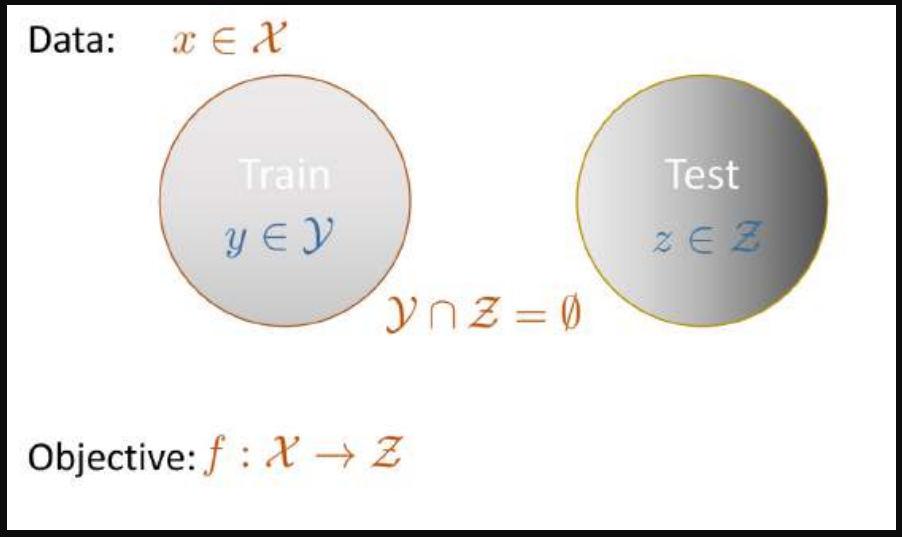
□从数据出发：从X-Data → Big Supervised Data

- ■图形学（物理建模）的方法举一反三：图形学重构现实世界
- ■GAN-like方法举一反三（learning-based generative models）
- ■自寻数据及无监督数据的自动标注：人类知识和其他模态的协同增效
  - □例1：通过**跟踪**获得类别不变的大量样本（用于目标检测，甚至分割任务）
  - □例2：对唇读（视觉）而言，可以通过语音识别大量获得唇读数据
  - □例3：对机器人而言，可以通过**操作**物体（比如拿起来看）来获得大量标注数据
  - □例4：**其他模态自动标注**，比如小孩子通过用胳膊够东西，获得深度信息

研究方向2——零样本学习

□零样本学习（Zero-shot Learning）——问题

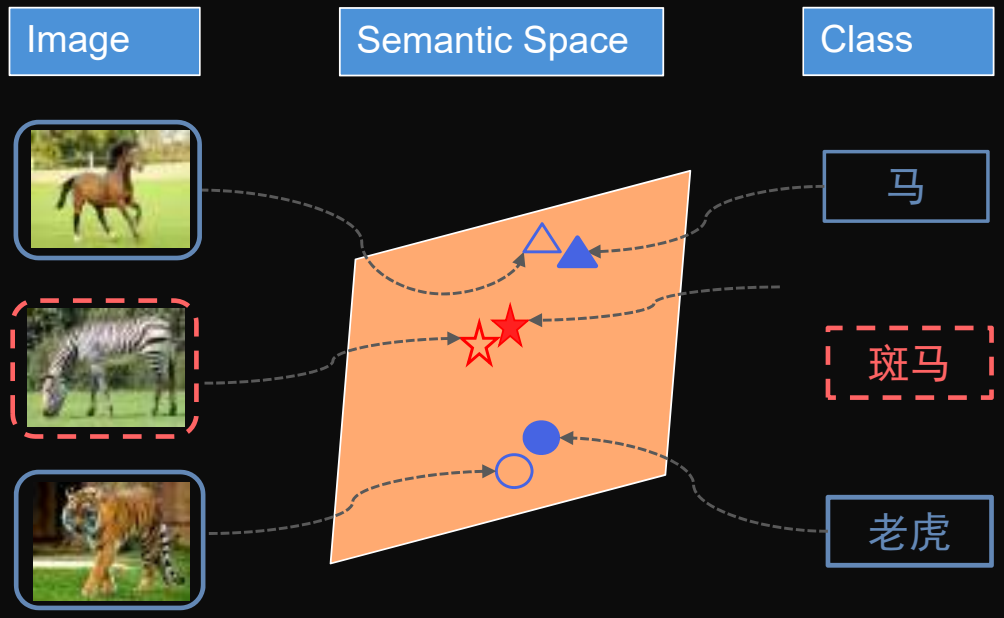
- ■如何在没有训练样本的情况下进行识别
- ■训练集与测试集的类别不相交
- ■零样本类别：**没有样本，但通过描述定义**



研究方向2——零样本学习

口零样本学习（Zero-shot Learning）——解决思路

■通过**语义空间**完成图像空间样本和类别空间之间的连接

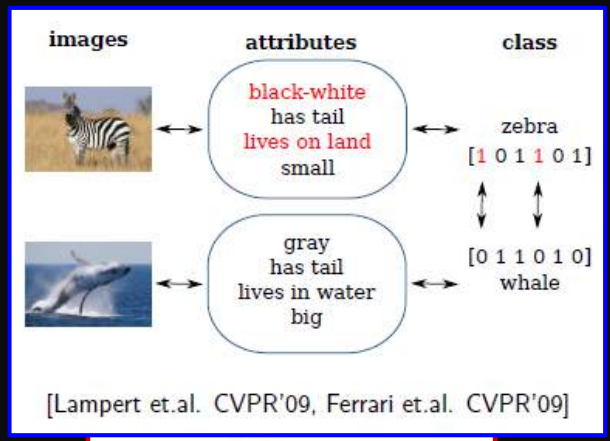


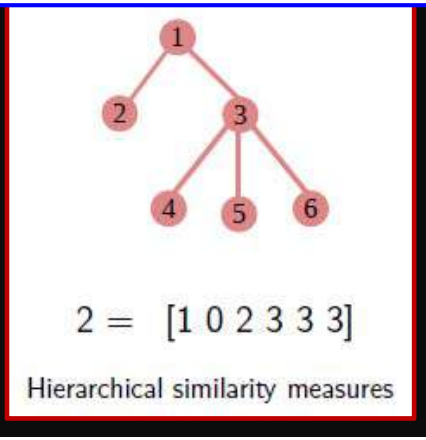
研究方向2——零样本学习

SeetaTech 中科视拓

□零样本学习（Zero-shot Learning）——语义空间选择

- ■文本空间
- ■属性空间
- ■相似性空间



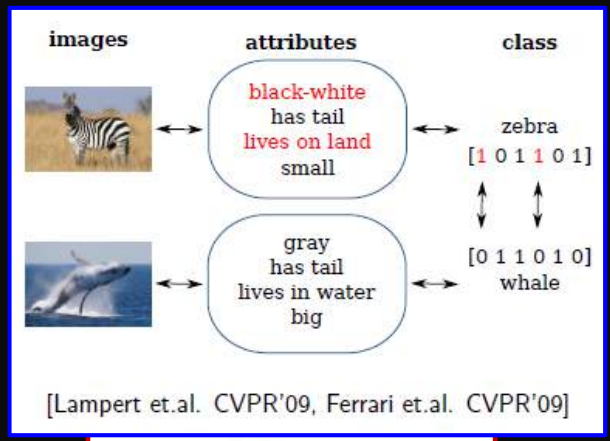


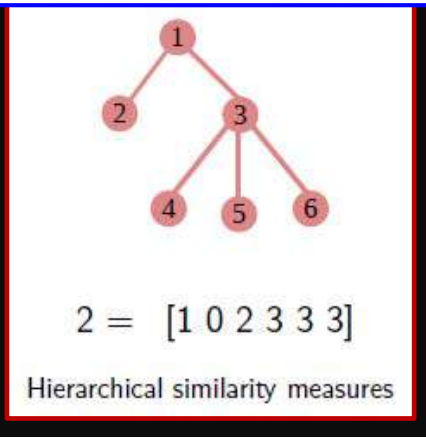
研究方向2——零样本学习

SeetaTech 中科视拓

□零样本学习（Zero-shot Learning）——语义空间选择

- ■文本空间
- ■属性空间
- ■相似性空间

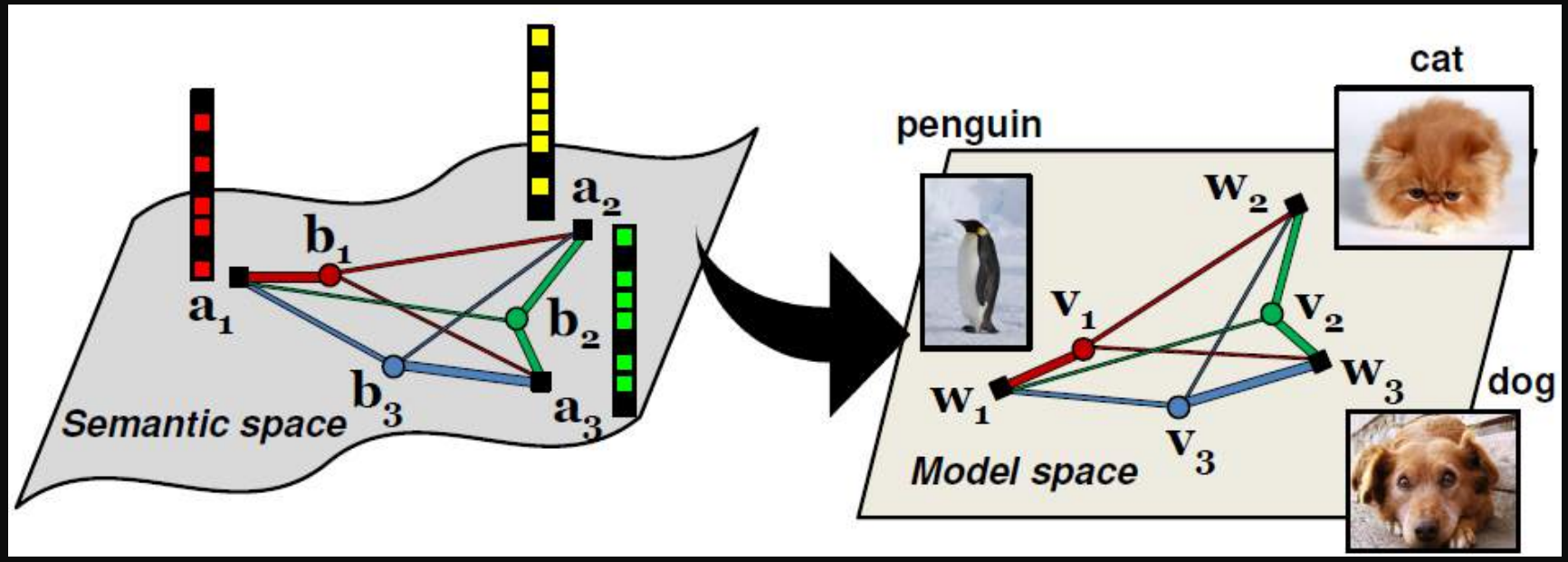




研究方向2——零样本学习

□零样本学习（Zero-shot Learning）——空间变换

■共享图像空间和语义空间的结构信息，然后在图像空间生成未知类分类器



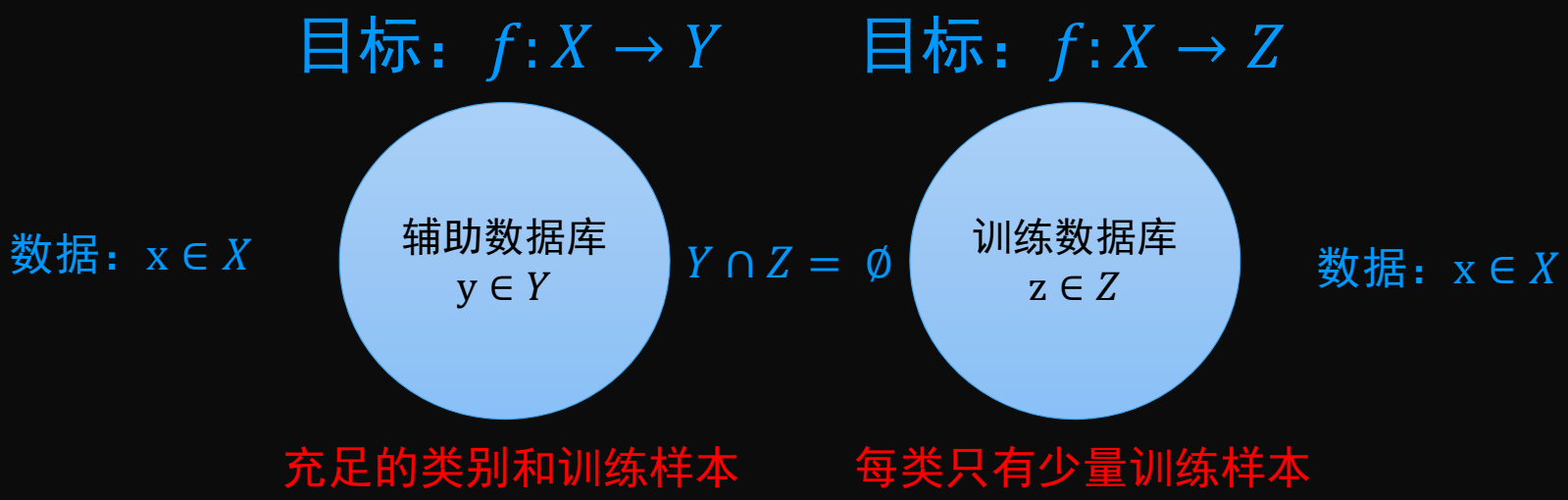
Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, Fei Sha. Synthesized Classifiers for Zero-Shot Learning. CVPR, 2016.

研究方向3——小样本学习

SeetaTech 中科视拓

问题定义

- 待识别的每个类别（Novel Classes）只有少量的样本可用于训练分类器
- 通常假定存在一个辅助集合（或基类数据集）
  - 有大量数据可以进行学习



研究方向3——小样本学习

SeetaTech 中科视拓

问题定义

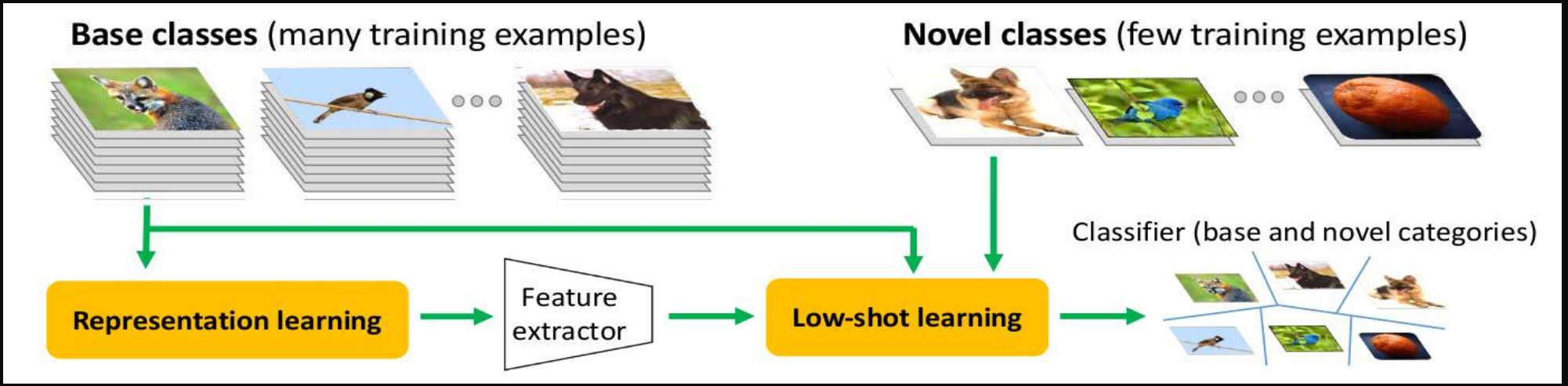
- 待识别的每个类别（Novel Classes）只有少量的样本可用于训练分类器
- 通常假定存在一个辅助集合（或基类数据集）
  - 有大量数据可以进行学习

?

研究方向3——小样本学习

口方法1：基于表示学习的方法

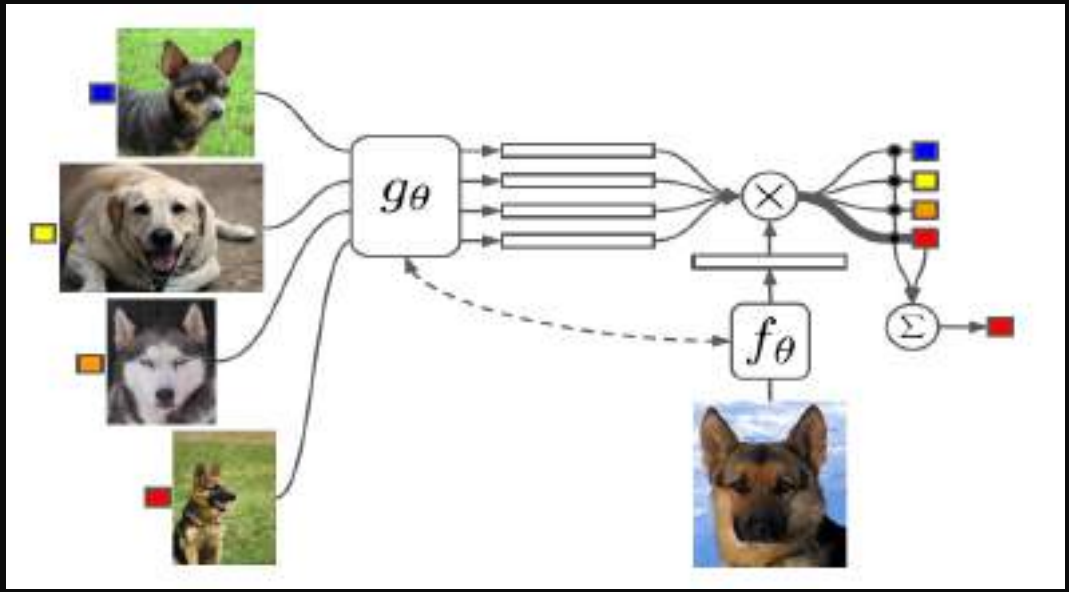
- ■步骤1：利用辅助集合（或基类数据集）学习表示，即：学习如何提取特征可以更好的区分不同类别的物体
- ■步骤2：基于few-shot学习分类器



研究方向3——小样本学习

口方法2：基于迁移学习的方法

- ■方法1：用few-shot进行新类别的finetuning
- ■方法2：通过辅助数据集，利用meta-learning思想进行训练

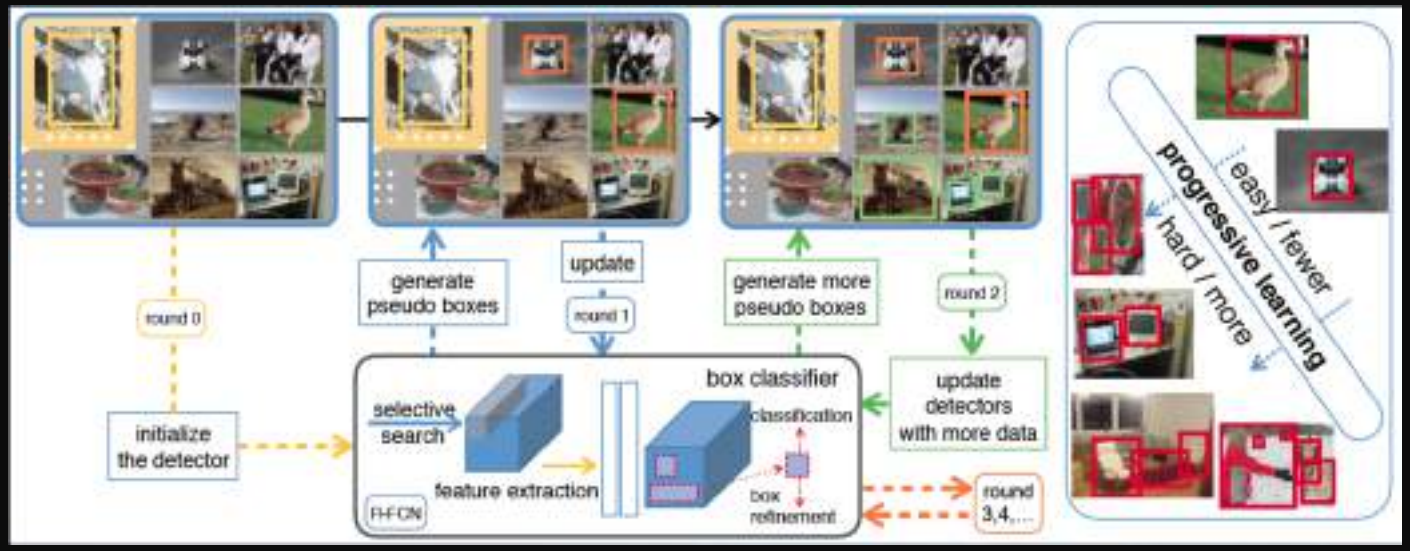


O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. in NIPS, 2016.

研究方向3——小样本学习

口方法3：利用大量无监督数据集的方法

- ■步骤1：用few-shot学习一个初始检测器D_0
- ■步骤2：对无监督样本进行检测，得到Pseudo标注样本
- ■步骤3：训练一个新的检测器，返回第一步形成迭代



研究方向3——小样本学习

□方法4：元学习（Meta-Learning）

■学习策略（N-class-K-shot）

- □1. 从辅助集中采样N类每类K个样本
- □2. 利用上述support set训练模型
- □3. 重新采样N类，每类K个样本
- □4. 更新模型
- □5. 重复3~4过程

■框架优势

- □在学习过程中，每阶段只利用少量数据更新模型，使得模型本身具有更强的适应能力
- □模拟小样本学习过程，在只有少量数据的情况下，模型可以很快适应新任务

总结与讨论

SeetaTech 中科视拓

□理论方法层面：仍然需要机器学习的本质进步！

- ■迁移学习，transfer learning
- ■自主学习（特别是对自主系统）
  - □主动发现对学习最有利的数据；自纠错学习
- ■多任务学习
  - □多模态的协同增效学习：
    - ■ 数据标签的自动获取；交叉验证（你对我也对，我好你也好）
  - □对抗学习：此消彼长（你对我错，你好我不好）
- ■进化学习：模型进化（由易及难；Never-ending learning）

总结与讨论

SeetaTech 中科视拓

传感器层面

- 必然超越人眼Retina的视觉信息获取能力！
- 四高一深（监控、iPhone X、自动驾驶、工业视觉将持续牵引）
  - 高清，高速，高动态，高光谱，深度(RGBD相机)
- 弱信号检测
  - 弱光成像，远距离成像，PPG
- 主动视觉——机器人产业必须的硬件基础设施
  - 模拟人眼的主动视点聚焦，Attention能力

总结与讨论

SeetaTech 中科视拓

口从计算设备出发：3-5年后端侧计算能力1000倍?

- ■云计算：脱机训练和inference都在云上
- ■端云协同阶段1：脱机训练在云，简单inference在端（edge）
- ■端云协同阶段2：脱机训练在云，全部inference在端（edge）
- ■端云协同阶段3：脱机训练和在线学习在云，全部inference在端（edge）
- ■端云协同阶段4：脱机训练在云，在线学习和全部inference在端（edge）
- ■完全端侧计算：脱机和在线学习一体化在端侧；全部inference在端侧

总结与展望

SeetaTech 中科视拓

- AI需要智慧之眼，视觉智能大有可为！
  - 会看的AI更智能！
- 学术界亟需突破知识和数据联合驱动的方法论！
  - 解决Scalability的问题！
- 当前阶段，对工业界，数据为王(Data is King)！
  - 但也要着眼未来，布局新的可能性（培育王后）

欢迎批评指正！