《人工智能通识》（科技素养）

第6讲 智能之眼—视觉感知

主讲：丛润民

章节知识点概览

知识点1：　智慧之眼——开启机器的超能力

知识点2：　给画质开美颜——图像增强

知识点3：　机器鹰眼——细粒度分类

知识点4：　AI侦探——目标检测

知识点5：　剪影艺术家——图像分割

知识点6：　揭秘阿凡达中的三维视觉

知识点1：智慧之眼——开启机器的超能力

01 计算机视觉简介

02 计算机视觉发展历史

计算机视觉定义

- 计算机视觉是一门让计算机“看见”并理解视觉信息的技术科学。其核心目标是让计算机像人类一样，能够**感知、分析和理解图像或视频**。简言之，计算机视觉就是**让机器“看懂”世界**，从图像或视频中提取有用的信息，并做出相应的判断或决策。

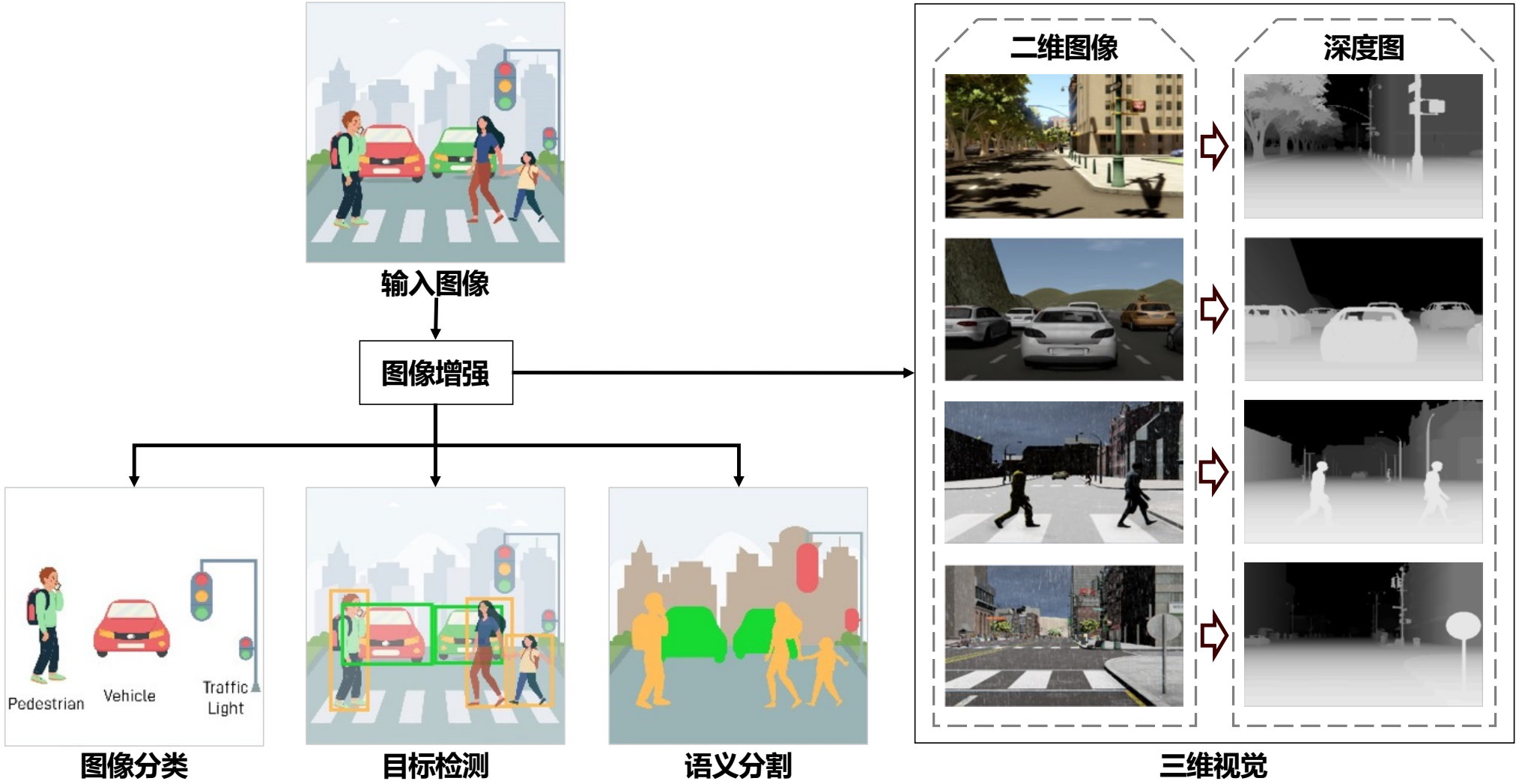
视频来源：好看视频-百葭苍苍

- 图像增强
- 图像分类
- 目标检测
- 语义分割
- 三维视觉
- 图像生成

- ✓ 自动驾驶
- ✓ 智慧医疗
- ✓ 无人安防
- ✓ 目标检测
- ✓ 工业检测
- ✓ 虚拟现实

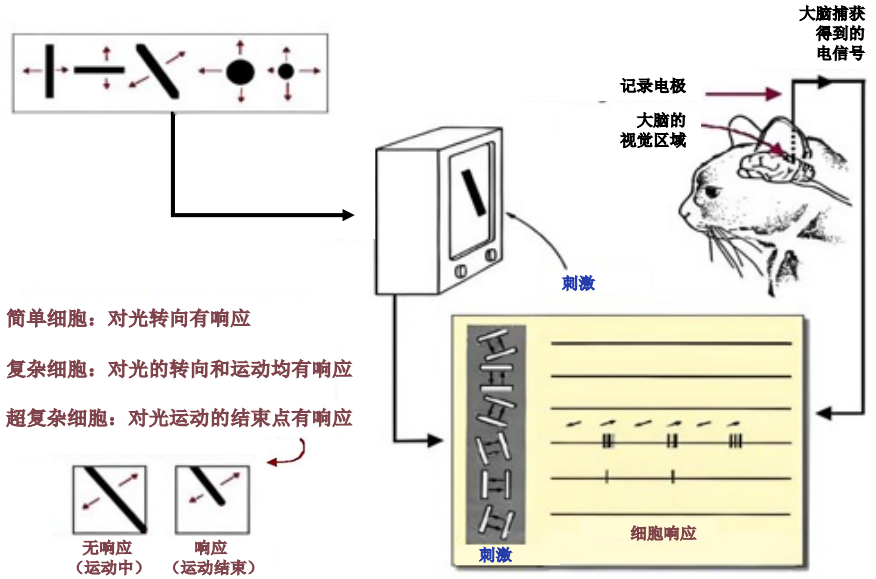
计算机视觉的应用——自动驾驶为例

■ 计算机视觉的技术链条密切相连



20世纪50年代——二维图像的分析与识别

- 1959年，大卫·休伯尔和托斯登·威塞尔通过猫的视觉实验，揭示了大脑皮层细胞对不同视觉特征的感知机制，奠定了生物视觉和计算机视觉的基础。
- 1959年，拉塞尔·基尔希等人发明了首台数字图像扫描仪，推动了数字图像处理技术的发展。

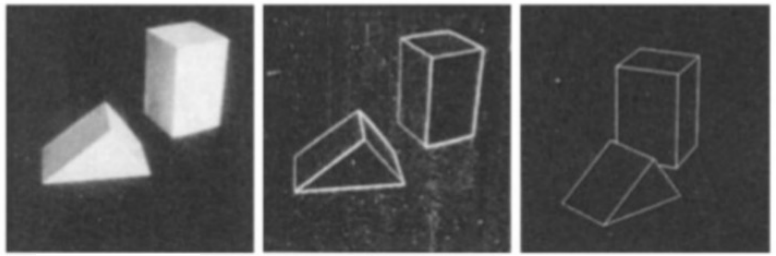


猫视觉实验

20世纪60年代——聚焦三维视觉理解

- 1965年，拉里·罗伯茨在《三维固体的机器感知》中提出了从二维图像提取三维信息的方法，开启了计算机视觉研究的新纪元。
- 1969年，威拉德·博伊尔和乔治·史密斯研发了电荷耦合器件（CCD），极大提高了数字图像采集的效率。

Larry Roerts
1963, 1st thesis of Computer Vision



输入图像　2×2梯度操作　新视角的计算后3D模型

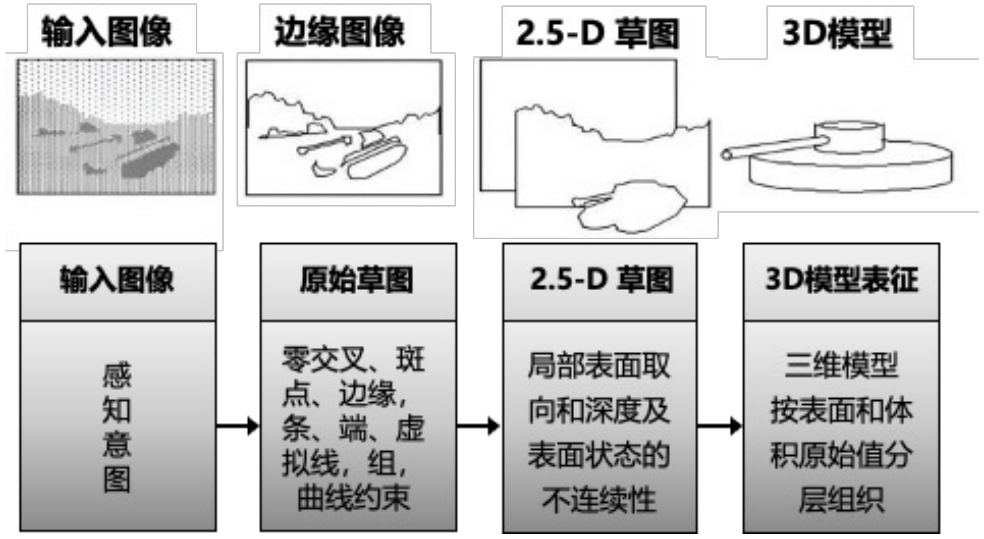
拉里·罗伯茨与《三维固体的机器感知》

20世纪70年代——理论体系的系统化构建时期

- 1977年，大卫·马尔提出“计算视觉”理论，创新性地描述了视觉信息处理的阶段性模型。

20世纪80年代——独立学科的确立与理论向应用的飞跃

- 1982年，大卫·马尔的《Vision》一书推动了计算机视觉理论的成熟，标志着该领域成为独立学科。
- 1982年，日本COGEX公司推出了首套工业光学字符识别系统，标志着计算机视觉开始进入工业应用。



“计算视觉”理论

David Marr被认为是计算机视觉之父，他在视觉信息处理方面的理论对后来的研究产生了深远的影响。为了纪念他，IEEE国际计算机视觉大会设立了马尔奖，每两年评选一次，授予在计算机视觉领域做出杰出贡献的论文。这个奖项代表了计算机视觉研究方面的最高荣誉之一。

发展历史

20世纪90年代——特征对象识别开始成为研究重点

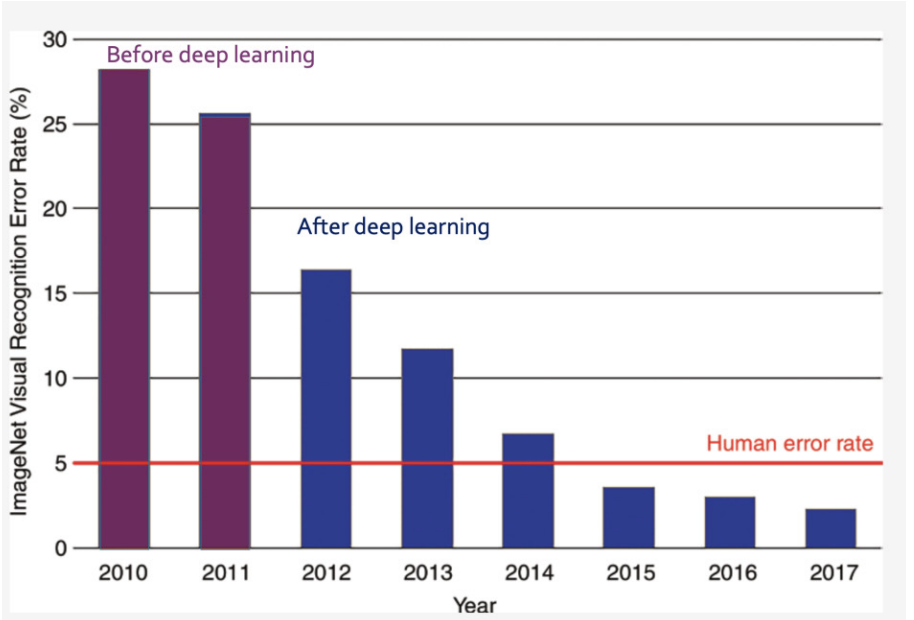
- 1992年，支持向量机提出，在分类、回归和异常检测等任务中表现出色，成为当时机器学习的主流方法之一。
- 1998年，杨立昆改进了卷积神经网络，提出了LeNet-5模型，奠定了现代卷积神经网络的基础。
- 1999年，NVIDIA推出GPU，开启了并行计算和图像处理的新纪元。

21世纪初，从机器学习迈向深度学习

- 2001年，提出了第一个实时人脸检测框架，首次实现了在普通硬件上实时检测人脸的目标。
- 2005年，提出了基于方向梯度直方图（HOG）的行人检测方法，成为重要的视觉工具。
- 2006年，Pascal VOC数据集推出，极大推动了对象分类技术的发展。
- 2006年，辛顿等人提出深度信念网络，标志着深度学习时代的开始，为后续的卷积神经网络等技术展奠定基础。
- 2009年，李飞飞教授及其团队发布ImageNet数据集，极大丰富了计算机视觉领域的研究数据资源。

2010年至今——深度学习引领的全面爆发

- **2010-2017年，ImageNet数据集与挑战赛的推动：**从2010年到2017年，ImageNet挑战赛（ILSVRC）成为推动深度学习技术快速发展的重要平台，极大提升了目标检测等视觉任务的性能。
- **2012年，AlexNet的突破：**2012年，亚历克斯·克里切夫斯基、伊尔亚·苏茨克维和杰弗里·辛顿提出的AlexNet深度卷积神经网络赢得ImageNet竞赛，标志着深度学习正式成为计算机视觉的核心技术，推动了卷积神经网络的广泛应用。



深度学习给ImageNet挑战赛带来的性能突破

“AI教父”辛顿：从木工到诺贝尔物理学奖得主

杰弗里·辛顿（Geoffrey Hinton），1947年12月6日出生于英国温布尔登，2018年图灵奖得主、2024年诺贝尔物理学奖得主，英国皇家学会院士，加拿大皇家学会院士，美国国家科学院外籍院士，多伦多大学名誉教授。辛顿在人工神经网络和深度学习领域做出了重大贡献，包括著名的反向传播算法、ImageNet竞赛首个冠军网络AlexNet均出自其手，被誉为“AI教父”，是深度学习领域的先驱和奠基人之一。

华裔之光李飞飞的逆袭之路：从洗碗工到改变世界的AI女神

李飞飞，1976年出生于中国北京，美国国家工程院院士、美国国家医学院院士、美国艺术与科学院院士，美国斯坦福大学首位红杉讲席教授。她主导了著名的ImageNet项目，该项目通过构建大型图像数据库，极大地推动了深度学习在计算机视觉中的应用。 2024年12月2日，李飞飞创办的AI创业公司World Labs宣布向“空间智能”迈出第一步：从单张图像即可生成三维世界。

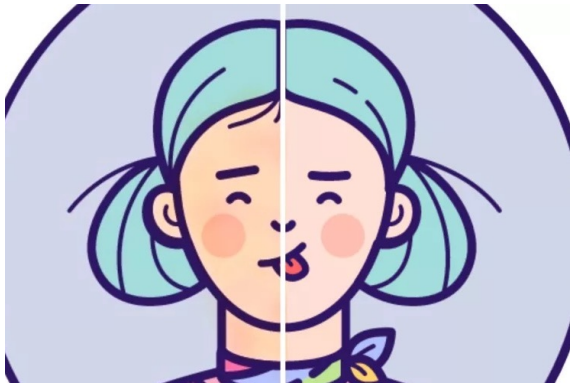
基地边缘


AI女神李飞飞
一个底层移民的逆袭

发展历史

- 2014年，伊恩·古德费洛提出**生成对抗网络（GAN）**，通过生成器与判别器竞争优化，革新了图像生成技术，成为计算机视觉领域的重大突破之一。
- 2016年，林宗毅等人提出**特征金字塔网络**，实现多尺度特征融合。
- 2017年，**PyTorch和TensorFlow的普及：**PyTorch和TensorFlow两个深度学习框架的崛起，迅速成为研究人员的主流工具，为包括图像分类在内的多项任务提供了强大的支持。
- 2017年，何恺明等人提出的**Mask R-CNN**在实例分割任务上实现了重大突破，其创新的RoIAlign技术显著提升了分割精度，成为该领域的里程碑式工作。
- 2020年，**Transformer模型进军计算机视觉：**Vision Transformer（ViT）的提出标志着Transformer模型正式应用于视觉任务，并迅速成为目标检测、语义分割和图像分类等任务的基础架构。
- 2020年，**扩散模型**作为生成模型的新范式，它不仅能够生成高质量、高分辨率的图像，还支持灵活的图像编辑功能，为图像修复、风格迁移、超分辨率重建等任务提供了新的解决方案。
- 2023年，**Segment Anything Model（SAM）**模型凭借其零样本学习能力，在视觉分割领域取得突破，为下游应用提供了广泛的可能性，预示着计算机视觉向智能化和自动化迈进。

知识点2：给画质开美颜——图像增强



01 图像增强概述

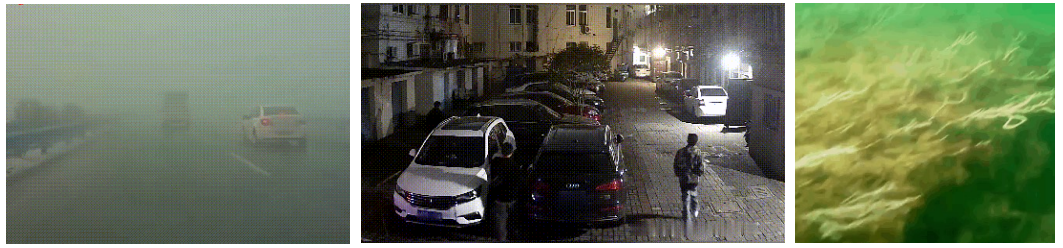
02 典型图像增强任务

图像获取时面临的质量问题

01

外界环境影响

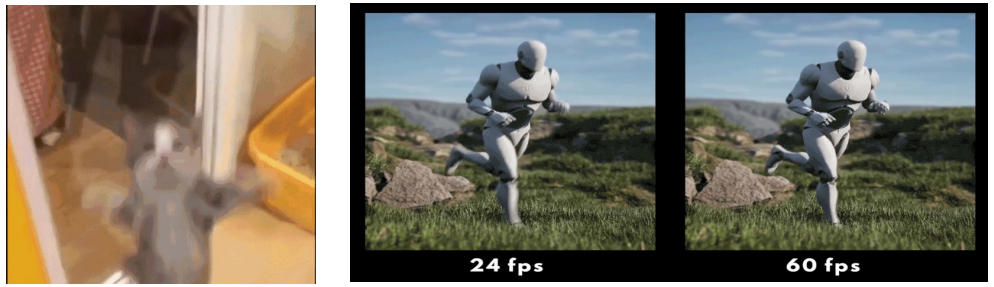
光照不足、雨天/雾天环境、水下介质作用等，图像往往会出现退化现象，如暗区、阴影和噪声等。



02

摄像设备限制

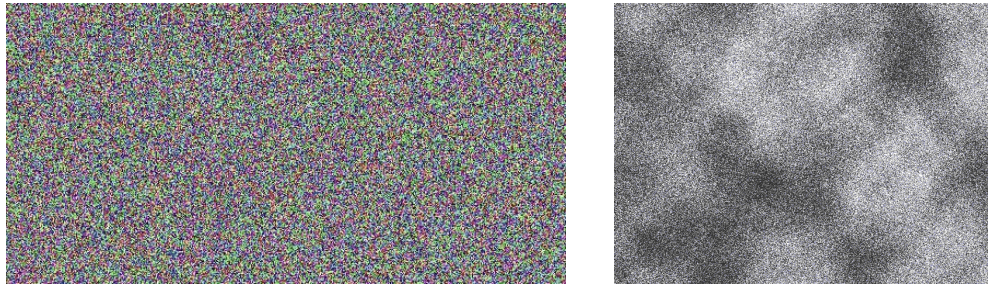
摄像设备的性能限制或拍摄抖动等，会导致图像细节丢失、模糊失真、色彩失真等问题。



03

传输过程干扰

图像在传输过程中可能会受到各种干扰，如电磁干扰、信号衰减等，导致图像质量下降。



图像增强概述

■ 图像增强目的

提高图像的清晰度，或是突出图像中的某些有用信息，同时压缩或去除其他无用的信息，从而使图像更加符合人眼视觉习惯或便于机器解析。

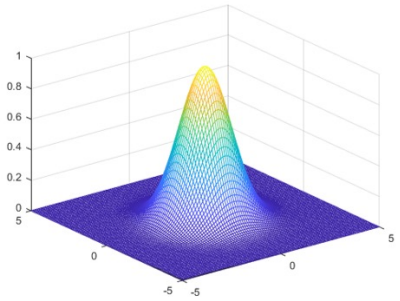
■ 应用场景

图像增强在多个领域中展现出重要价值，如医疗图像的增强、监控系统的图像处理、遥感图像的分析等，并作为目标识别、跟踪、特征匹配、图像融合等高级图像处理算法的预处理步骤，发挥着至关重要的作用。

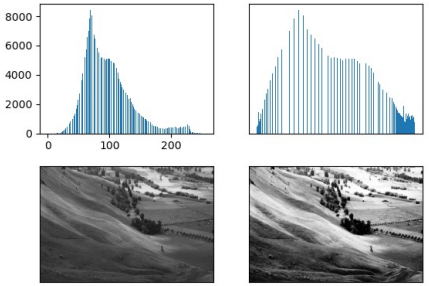
■ 早期阶段的信号处理技术

➢ 时域处理方法

滤波技术、非凸低秩优化策略、直方图均衡化、灰度变换、锐化滤波等，直接对图像像素进行操作，改善图像对比度与亮度。



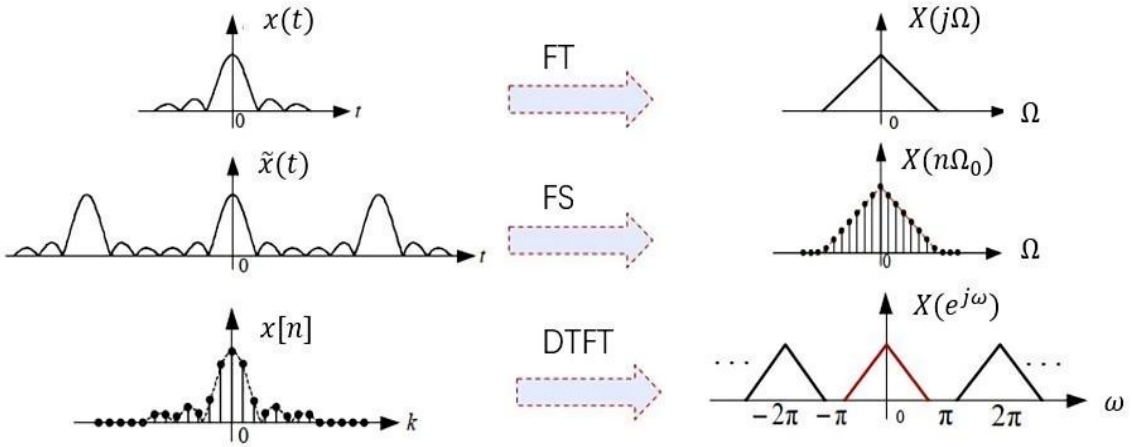
滤波技术

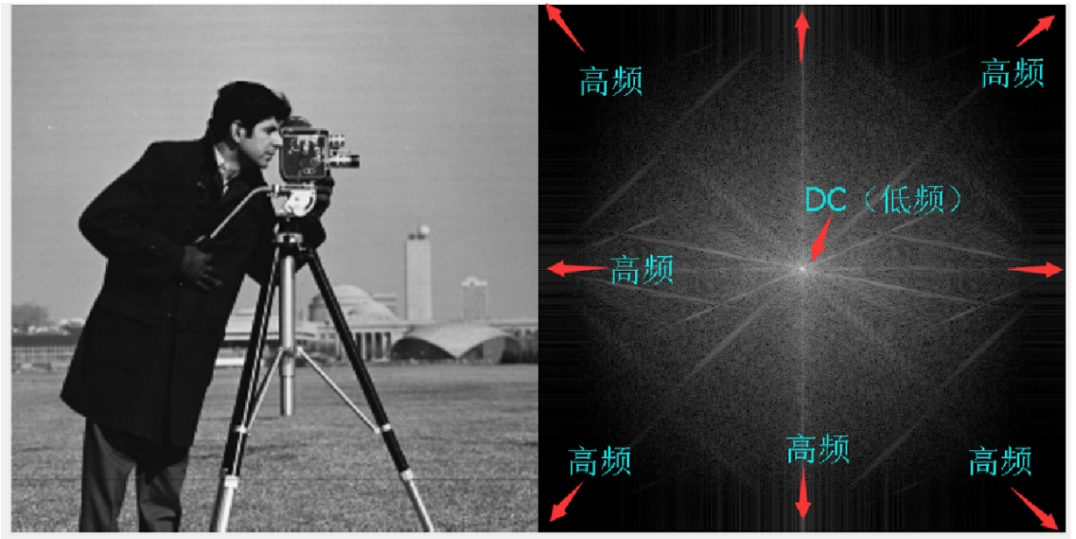


直方图均衡化

➢ 频域处理方法

通过傅里叶变换、小波变换、离散余弦变换等将图像转换至频率域进行处理，实现去噪、边缘增强等效果。





■ 深度学习的引入

➢ 单域网络架构原理

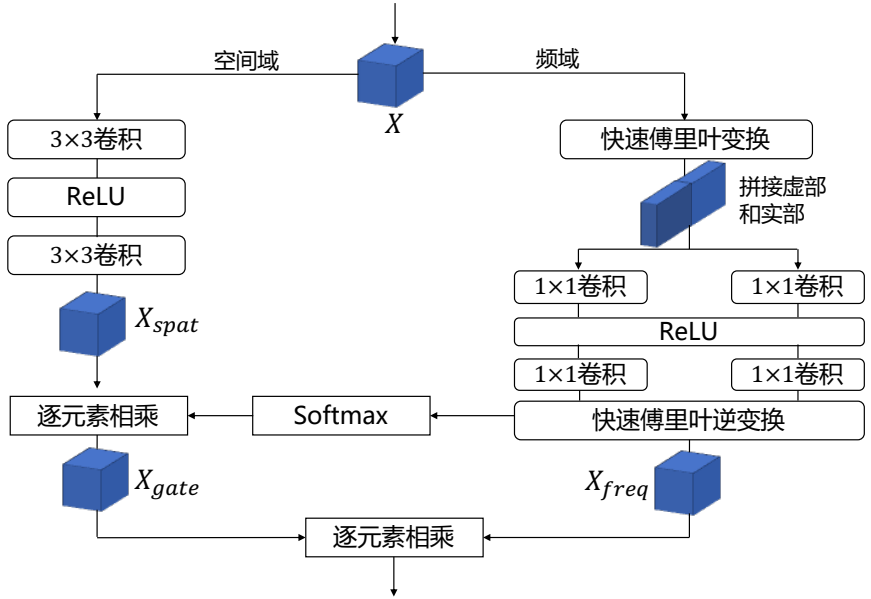
单域网络架构主要针对单个图像域进行训练，如空域或频域，通过卷积神经网络等模型，提取图像特征并重建高质量图像。

➢ 单域网络架构局限性

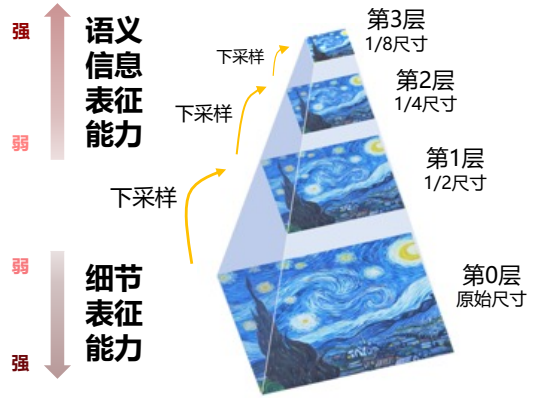
单域网络架构在处理具有多种复杂降质因素的图像时，难以同时恢复图像的细节和纹理信息，导致图像质量不尽如人意。

➢ 时域与频域结合方法

综合运用时域与频域处理技术，探究多域协同和多尺度融合技术提高图像增强效果，满足复杂场景需求。



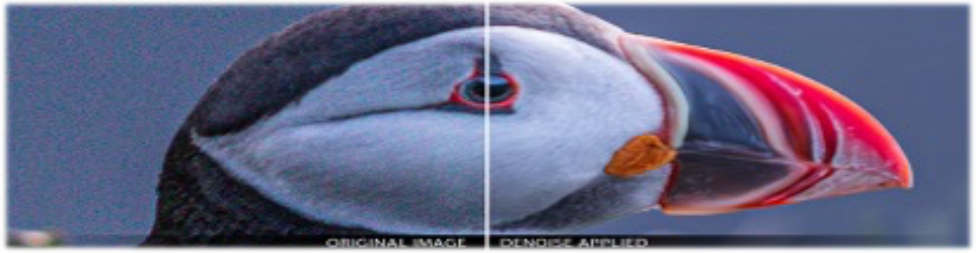
多域协同模块示例



多尺度思想

图像增强概述

（a）图像超分辨率重建



（b）图像去噪

（c）低光照增强

图像增强
任务概览

（d）图像复原

（e）图像去雾

（f）图像去模糊

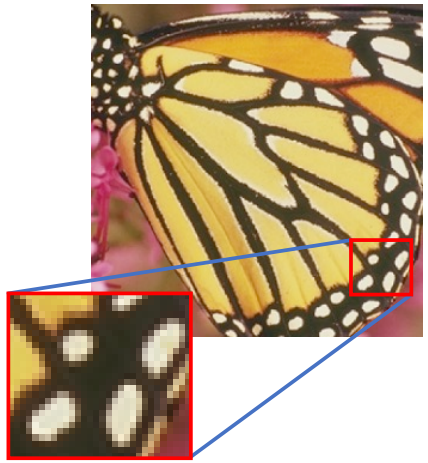
（g）水下图像增强

典型图像增强任务——图像超分辨率

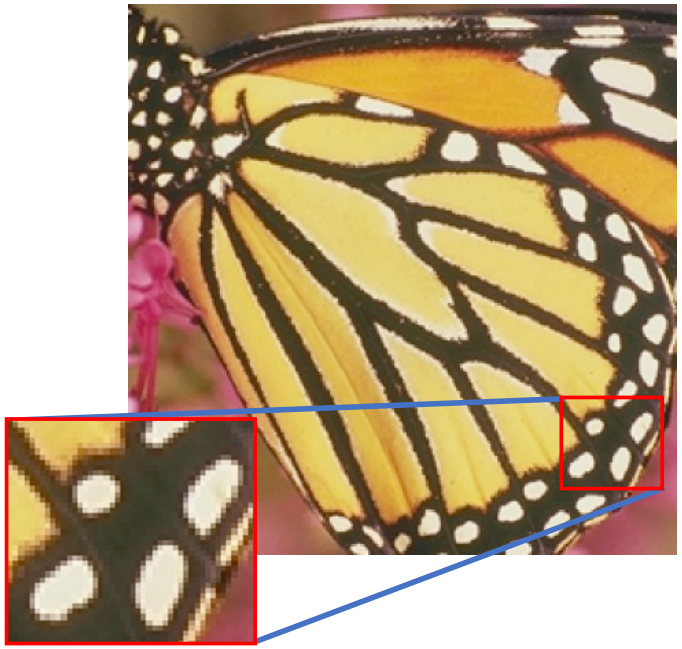
■ 图像超分辨率

典型图像增强任务——图像超分辨率

■ 图像超分辨率



超分辨率重建



任务描述

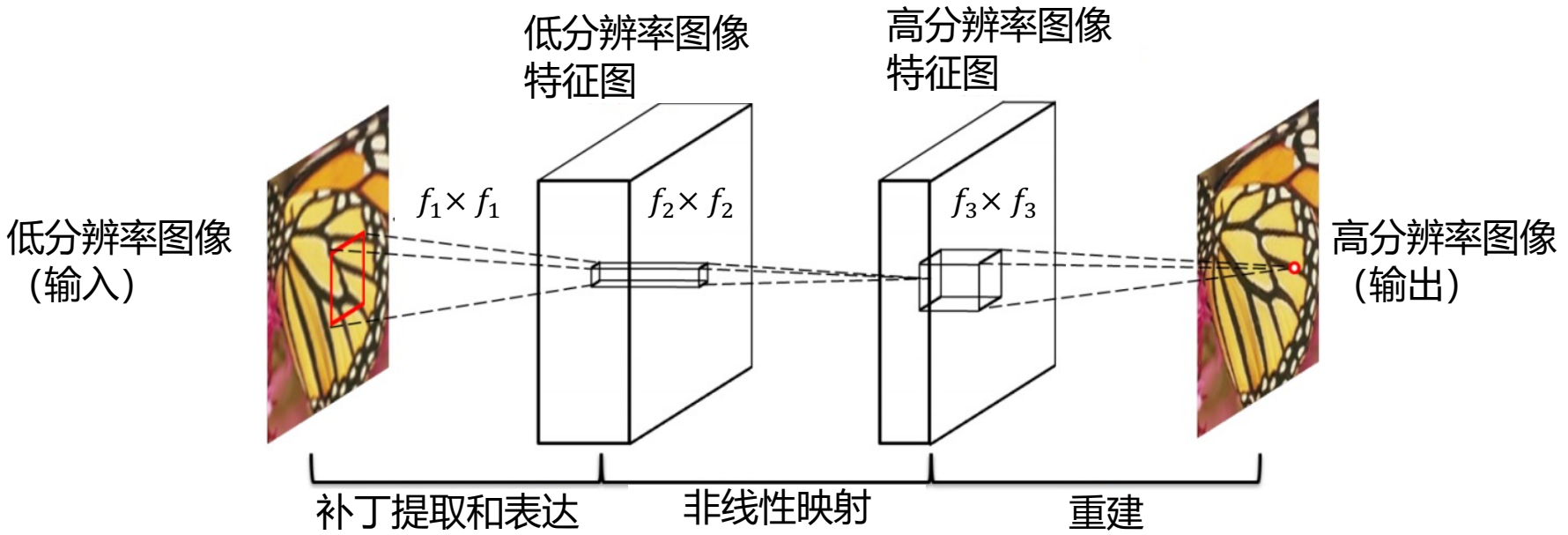
旨在从低分辨率图像中重建出高分辨率图像，同时尽可能地保留和恢复原始图像中的细节和纹理信息

难点分析

图像细节丢失、图像噪声、算法复杂度高

典型图像增强任务——图像超分辨率

■ 经典方法 —— SRCNN（Super-Resolution Convolutional Neural Network）

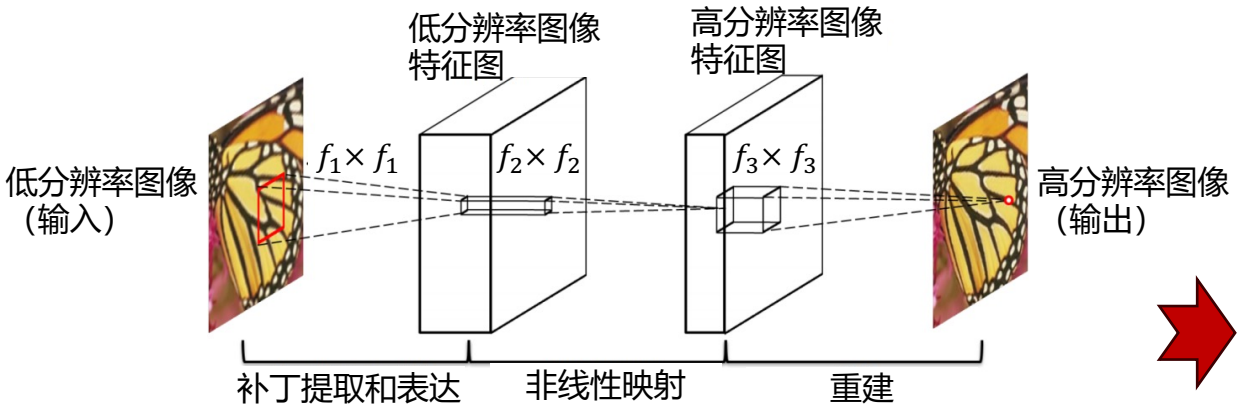


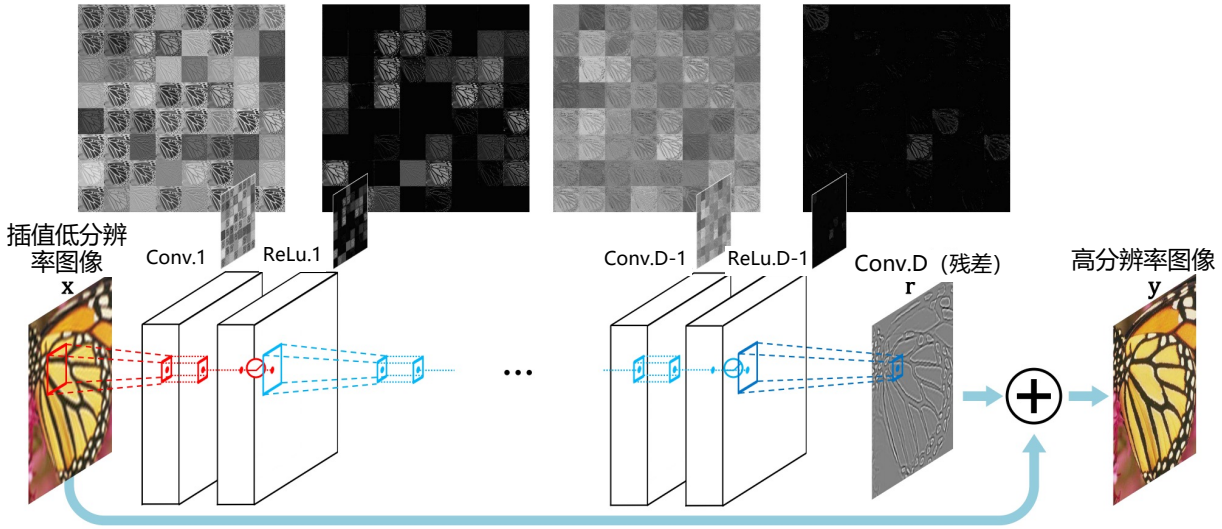
- **开山之作**：2014年提出，图像超分辨率的首个深度学习模型，截止到2025年3月，该论文的谷歌学术引用高达6700+次。
- 它通过卷积网络建立低分辨率图像到高分辨率图像的**直接映射**。流程包括：上采样、特征提取和重建高分辨率图像。
- SRCNN以其简洁高效的设计为图像超分辨率研究奠定了基础，但存在网络深度不足、训练慢以及仅适用单一尺度的局限。

C. Dong, et al, Learning a deep convolutional network for image super-resolution, ECCV 2014.

典型图像增强任务——图像超分辨率

■ 经典方法 —— VDSR（Very Deep Super-Resolution）





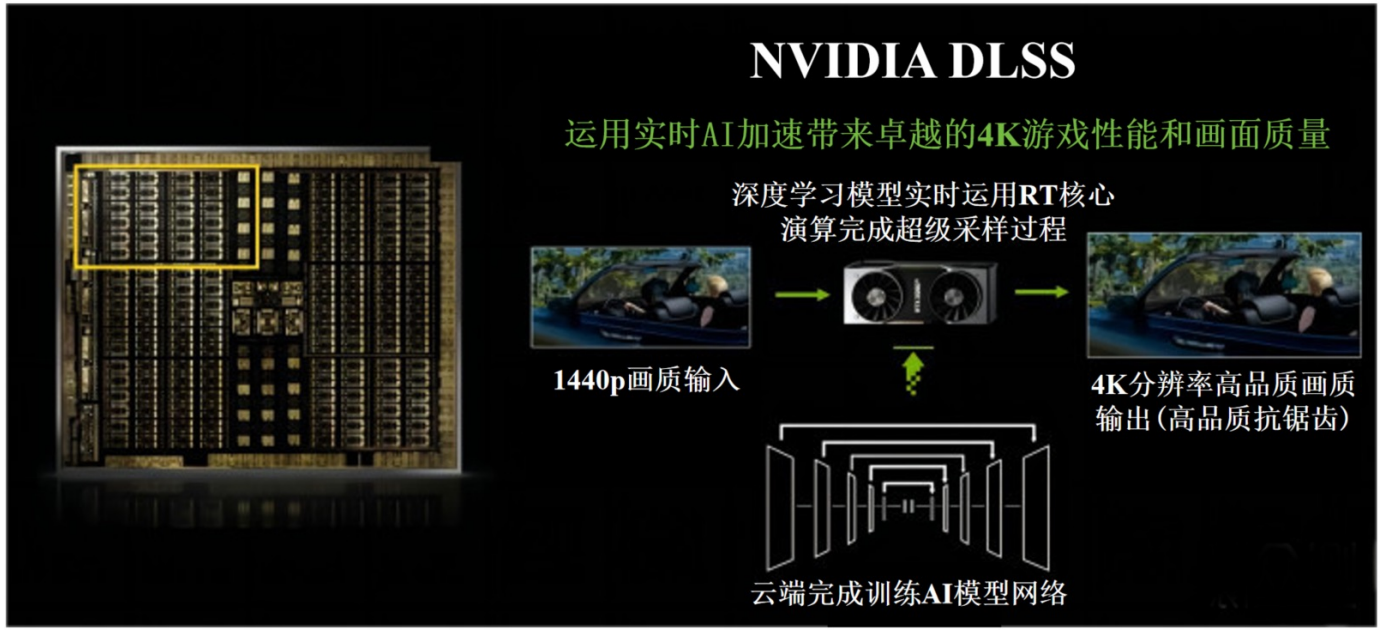
- 通过非常深的卷积神经网络来学习图像细节的残差信息，并将这些信息与插值放大的图像相结合，从而有效地恢复图像细节并提升图像分辨率。

- 残差学习加速训练收敛并缓解梯度消失问题
- 单一模型能够处理多种缩放因子
- 高学习率与动态梯度剪裁的训练策略，VDSR有效提升了图像细节恢复能力和训练效率

Kim J, Lee J K, Lee K M, Accurate image super-resolution using very deep convolutional networks, CVPR 2016.

典型图像增强任务——图像超分辨率

■ 图像超分辨率的应用



消费电子产品领域

修复老照片、扩大图片尺寸及智能修复图像破损

数字媒体和娱乐领域

实现了从低分辨率输入到接近原生高分辨率输出的高效转换，从而在降低显卡负担的同时，保证了游戏画面的流畅性和视觉质量

■ 未来发展趋势

01

实时超分辨率重建

模型更加高效和智能，能够在更短的时间内生成更高质量的高分辨率图像。

02

多模态数据融合

通过结合深度信息、红外图像等其他数据源，进一步提升重建效果 。

03

跨领域应用扩展

在虚拟现实、增强现实和医疗影像等领域应用，为更多行业带来创新和变革。

典型图像增强任务——低光图像增强

右图最前方车辆旁边几个人？

低光图像增强

任务描述

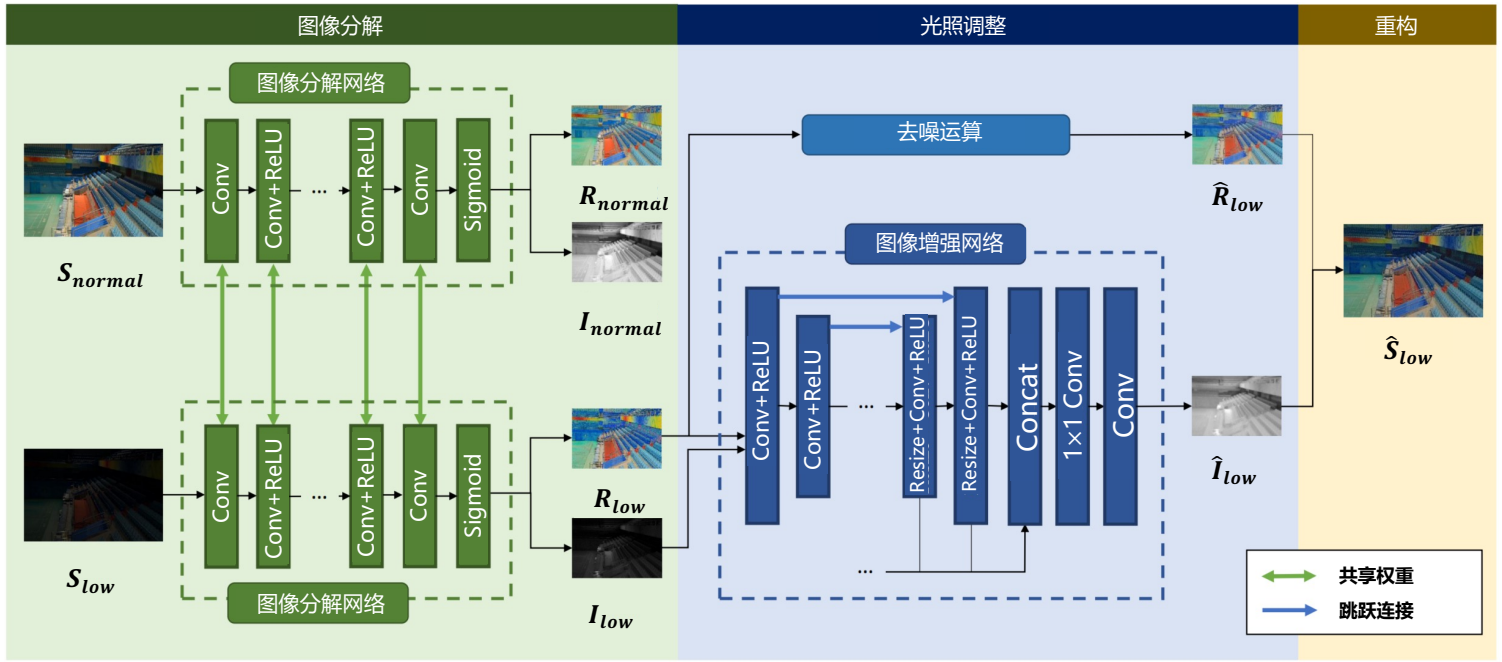
在低光环境下，图像通常模糊且噪声多，这会影响后续处理和分析。低光照图像增强技术通过改善亮度和对比度，使图像更加清晰。

技术应用

低光照图像增强技术在夜间摄影、安全监控、自动驾驶等领域得到了广泛的应用，拓宽了图像技术的应用范围，增强了信息的可用性和准确性。

典型图像增强任务——低光图像增强

■ 经典方法 —— RetinexNet



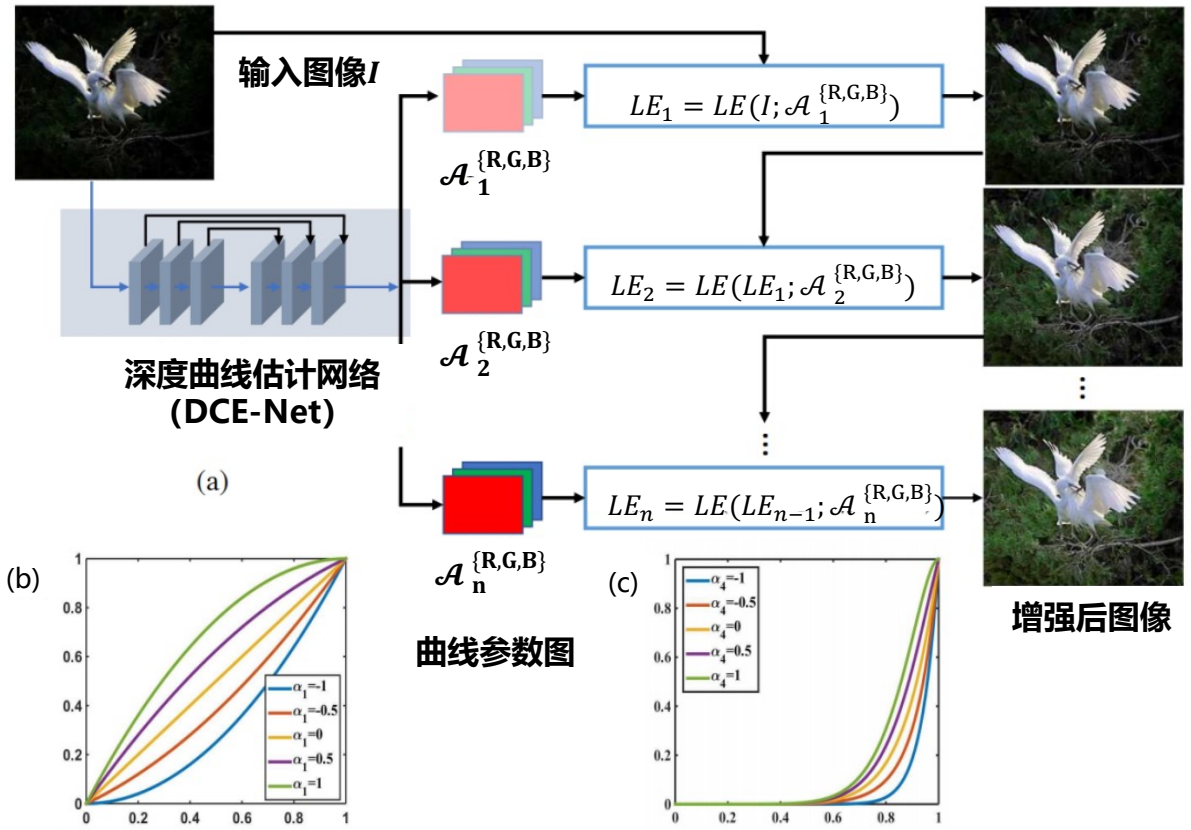
Retinex理论是一种用于低光照图像增强的经典方法，宝丽来公司创始人、即显摄影发明者埃德温·赫伯特·兰德提出。它通过分离图像的反射率和光照分量来调整光照，保持物体颜色稳定性，广泛应用于图像增强、非均匀光照校正等领域。

- **RetinexNet是基于Retinex理论的深度学习改进版**
- 通过两个子网络分别处理反射率和光照分量，增强低光照图像

Wei C., et al, Deep Retinex Decomposition for Low-light Enhancement, BMVC 2018.

典型图像增强任务——低光图像增强

■ 经典方法 —— Zero-DCE



- Zero-DCE是一种**无需参考图像**的低光图像增强方法。该方法通过**像素级曲线估计**调整图像亮度和对比度，不依赖成对数据，适用于复杂光照条件。
- Zero-DCE使用轻量级网络结构和非参考损失函数，模型通过7层卷积层和跳跃连接保留图像细节。其核心是通过多次迭代的曲线估计来精确调节亮度，避免过度增强。
- 为了实现无参考训练，设计了空间一致性、曝光控制、颜色恒常性和光照平滑等损失函数，确保增强后的图像自然、清晰。

Guo C, et al., Zero-reference Deep Curve Estimation for Low-light Image Enhancement, CVPR 2020.

典型图像增强任务——低光图像增强

■ 低光图像增强的应用

阿里达摩院研发的车载摄像头ISP处理器（AliISP），通过3D降噪和图像增强算法，提升了夜间车载摄像头的图像识别能力，应用于自动驾驶物流车，增强了自动驾驶的安全性和可靠性。

■ 面临的挑战

- 鲁棒性与泛化能力：提高技术在复杂光照条件下的稳定性和适应能力需要研究更智能的算法，以应对不同光照条件下图像质量的差异。
- 实时处理与资源效率：随着移动设备和嵌入式系统的普及，低光照图像增强技术需要更高效、轻量化的算法，提升实时处理能力，减少资源消耗。
- 多模态数据融合：未来研究应探索结合不同传感器（如红外、雷达等）的数据，获取更全面的场景信息，从而实现更智能、精确的图像处理。
- 自适应增强技术：研究基于图像内容和光照条件自动调整增强参数的技术，以获得最佳视觉效果，进一步提升用户体验。

知识点3：机器鹰眼——细粒度分类

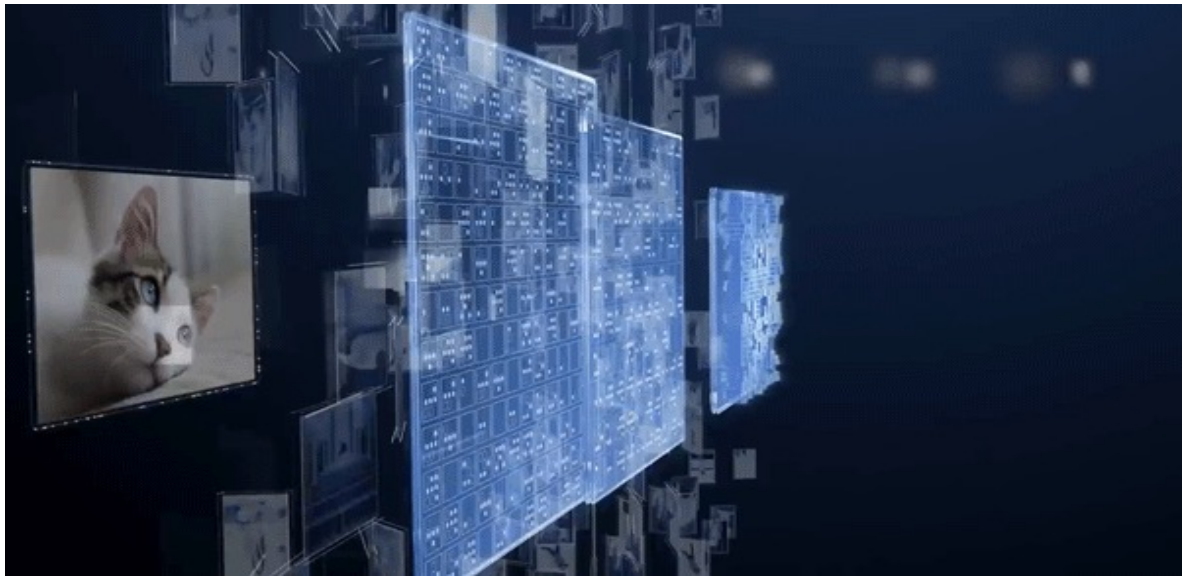
01 图像分类概述

02 细粒度图像分类

图像分类概述

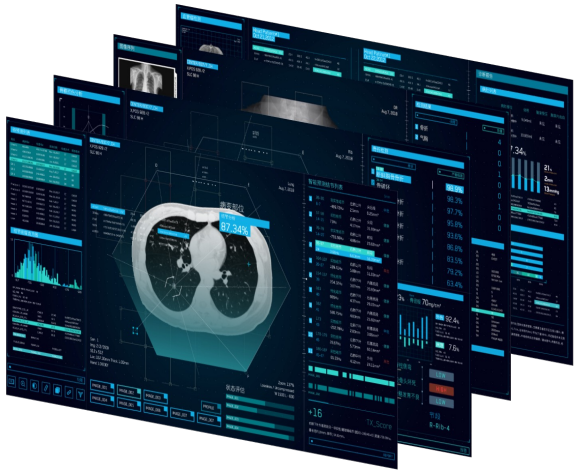
■ 图像分类目的

图像分类是指将输入的图像分配到一个预定义的类别集合中，类别可以是“猫”、“狗”或“车”等特定物体。



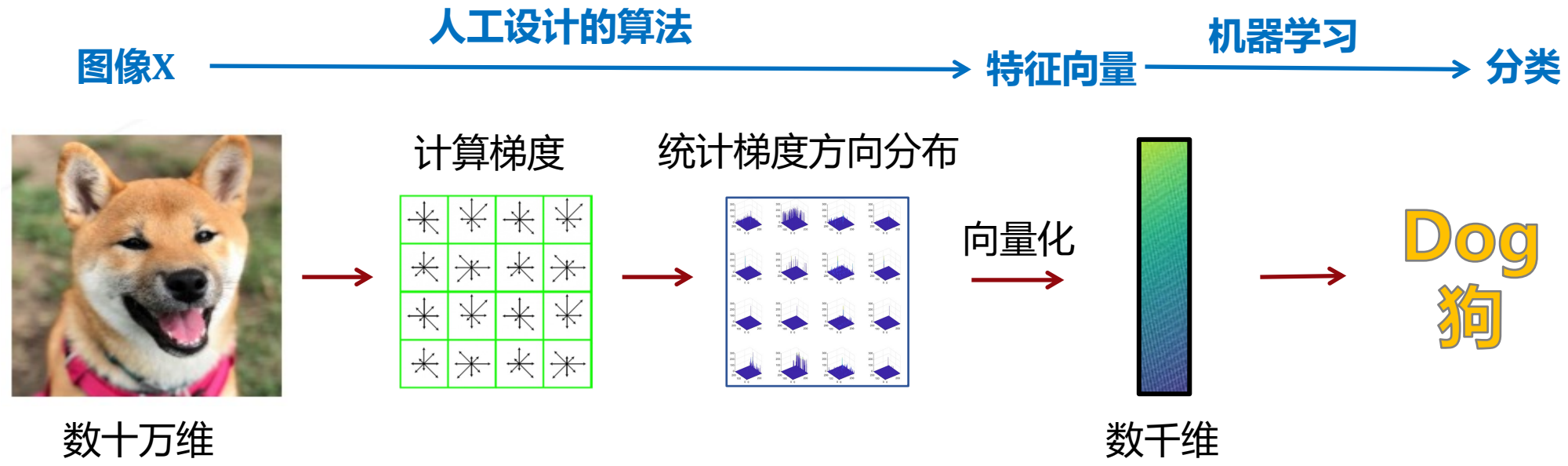
■ 广泛的应用场景

在多个领域都有广泛应用，例如自动驾驶中的行人识别、医学图像中的疾病检测、农业中的病虫害识别等。它是许多视觉任务的基础，例如面部识别、物体检测、场景理解等。随着深度学习的发展，图像分类已成为人工智能的重要应用之一。



■ 传统图像分类方法

- 传统方法通常包括数据预处理、特征提取（如SIFT、HOG等）和分类器的训练（如支持向量机、K-近邻等）。
- 虽然这些方法在某些情况下能取得较好的效果，但其依赖于手工设计的特征提取过程，难以应对复杂任务。



■ 深度学习的崛起

- 随着计算能力和数据规模的提升，深度学习模型（如AlexNet、VGGNet、ResNet等）可以自动从图像中学习层次化的特征表示，极大提高了图像分类的准确性和效率。
- 深度学习通过卷积神经网络（CNN）等架构，减少了对手工特征的依赖，并且能够在大规模数据集上取得优异表现。

图像X → 可学习的特征提取 $X^{(1)} = f_{\theta_1}^1(X)$ → 可学习的特征提取 $X^{(2)} = f_{\theta_2}^2(X^{(1)})$ → …… → 可学习的特征提取 $X^{(L)} = f_{\theta_L}^L(X^{(L-1)})$ → 分类

学习如何产生适合分类的特征

Dog
狗

多个简单特征变换复合构成一个复杂的端到端分类器

$$P(y|X) = F_\Theta(X) = \left(f_{\theta_1}^1 \circ f_{\theta_2}^2 \circ \cdots \circ f_{\theta_L}^L\right)(X)$$

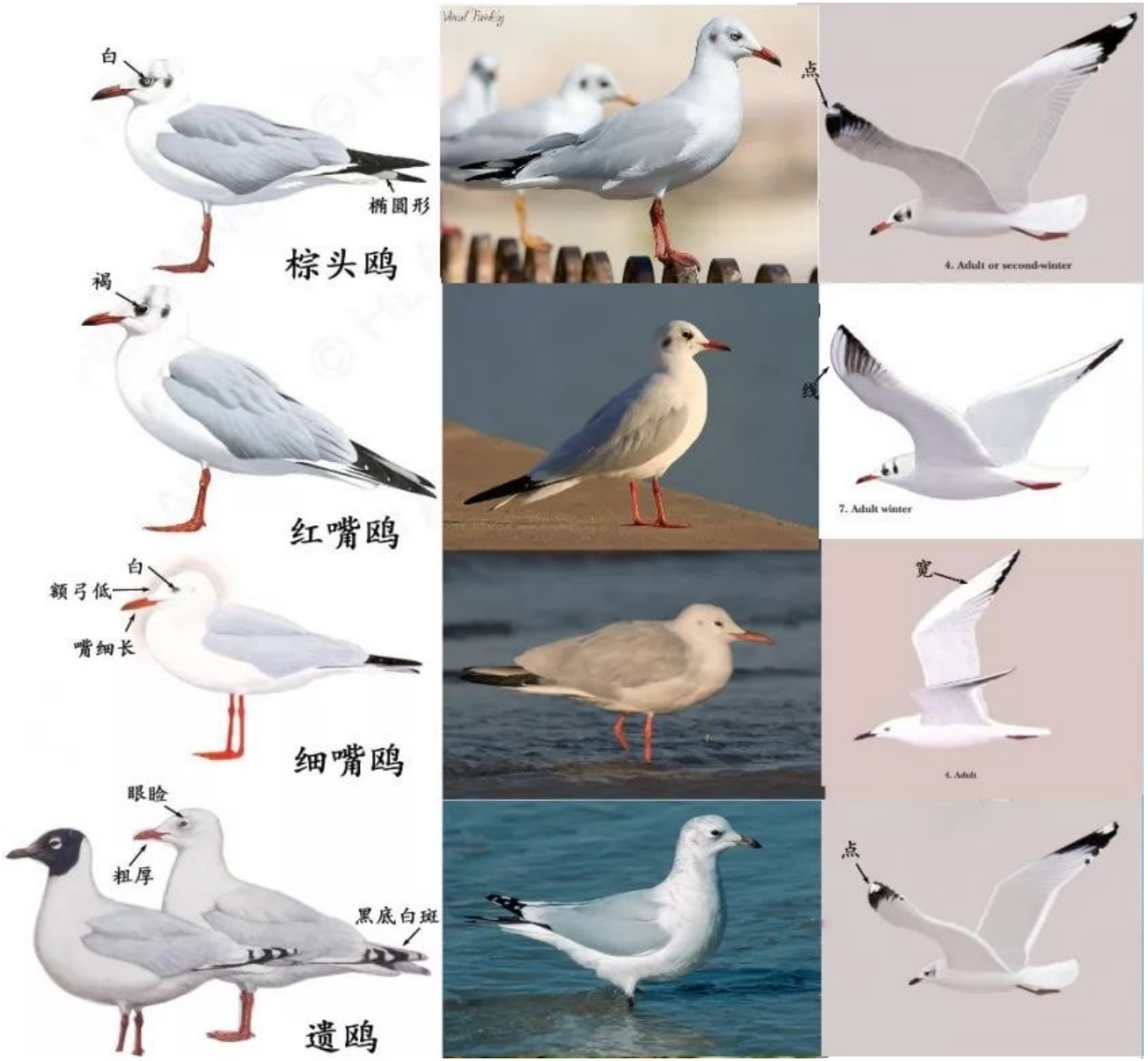
■ 图像分类的实际应用

- **人脸识别：**杭州萧山国际机场通过阿里云ET航空大脑的人脸识别技术，大幅提升了安检和食堂支付效率，旅客通过安检时仅需3秒，人脸识别还能有效应对整容、照片更新不及时等复杂情况，方便特殊人群使用。

- **医疗影像诊断：**针对青光眼筛查，腾讯觅影团队研发了基于深度学习的青光眼分类模型，准确率超过95%，显著提高了筛查效率和早期青光眼的诊断准确性，帮助缓解医疗资源不均问题。

鹰眼——超越人类的视力

学学测测



棕头鸥：虹膜白色，头白色，嘴红，尖端黑色，眼后有一黑褐色斑，头顶有两道不明显斑纹，上体浅灰色，翼尖黑色，下体白色，停歇时候可以观察到初级飞羽下部有一椭圆形白斑。

学学

红嘴鸥：虹膜暗褐色，头白色，嘴红，尖端黑色，眼后有一黑褐色斑，头顶有两道不明显斑纹上体浅灰色，翼尖黑色，下体白色，有些个体泛粉红色，停歇时候可以观察到初级飞羽翼下全白，飞行时表现为初级飞羽外侧两根羽毛为白色，仅先端黑色。

细嘴鸥：虹膜白色，其额弓明显较低，嘴细长为暗红色，眼后黑色斑模且较小，上体浅灰色，翼尖黑色，下体白色或粉红色，停歇时候可以观察到初级飞羽翼下全白，飞行时候初级飞羽外侧四根羽毛为白色，仅尖端黑色，观感为翅上白色羽毛较宽。

遗鸥：虹膜褐色，具有白色眼睑，头污白色，枕部及颈部一般具有灰褐色斑，嘴红色较粗厚，因脖子较短，整体看起来较矮胖，停歇时候可以观察到初级飞羽上部白色斑块较明显、下部有一椭圆形白斑。

开始测试

请将图片与其正确的名字连起来




棕头鸥　　红嘴鸥　　细嘴鸥　　遗鸥

细粒度图像分类

■ 定义

细粒度图像分类是指在大类别（如“狗”或“鸟”）中，进一步细分到子类别（如不同品种的狗或鸟），从而**识别具有细微差异的物体**。

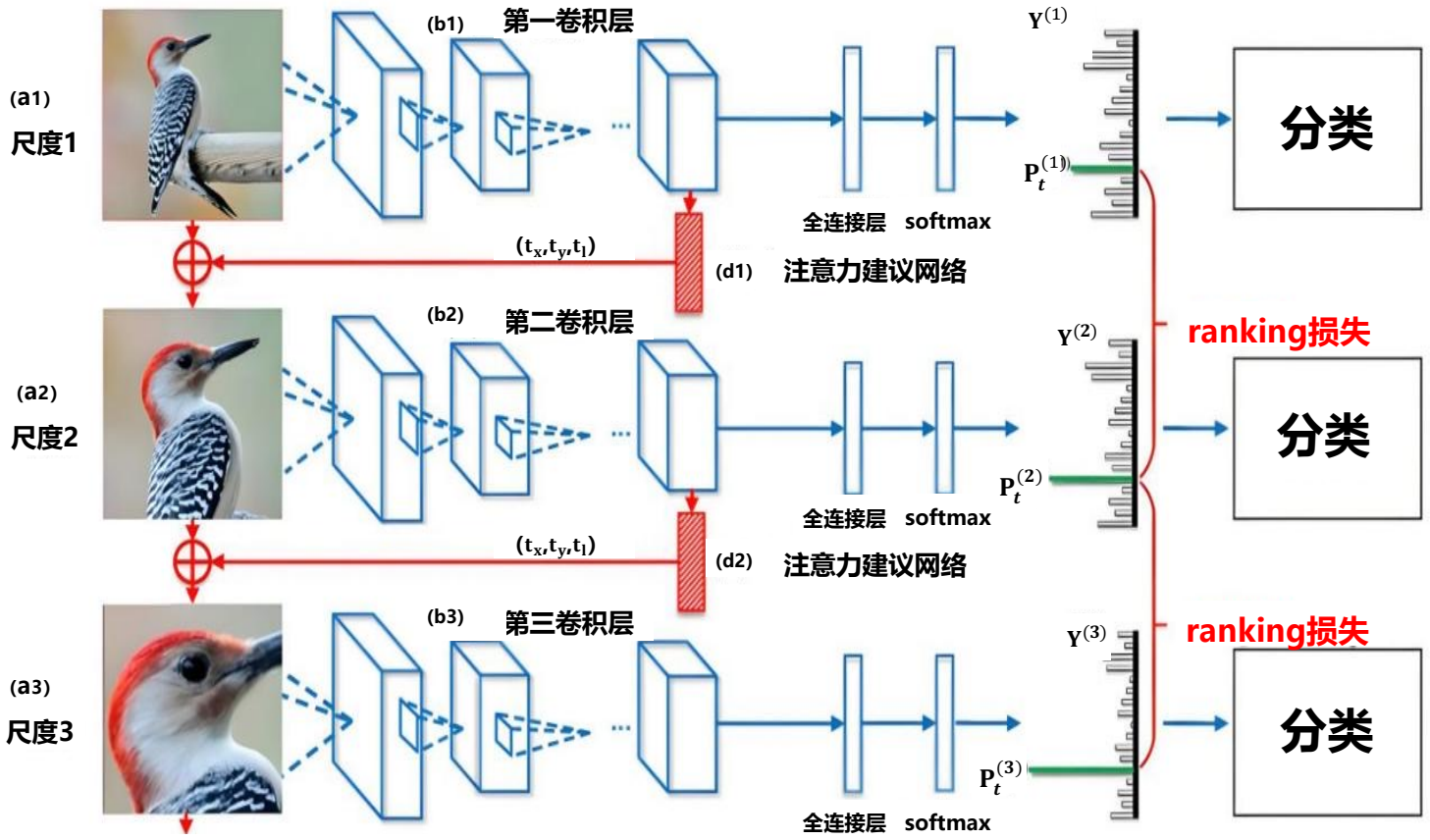
■ 挑战

由于细粒度图像分类**需要对局部细节做出精确判断**，环境因素（如光照、角度、背景等）容易干扰分类结果。此外，如何让模型学会区分细微差别并保持高准确性是一个难点。

■ 应用

生物多样性保护与研究、医疗与生命科学、艺术品与文化保护、电子商务与零售等。

■ 经典方法——循环注意力卷积神经网络（RA-CNN）

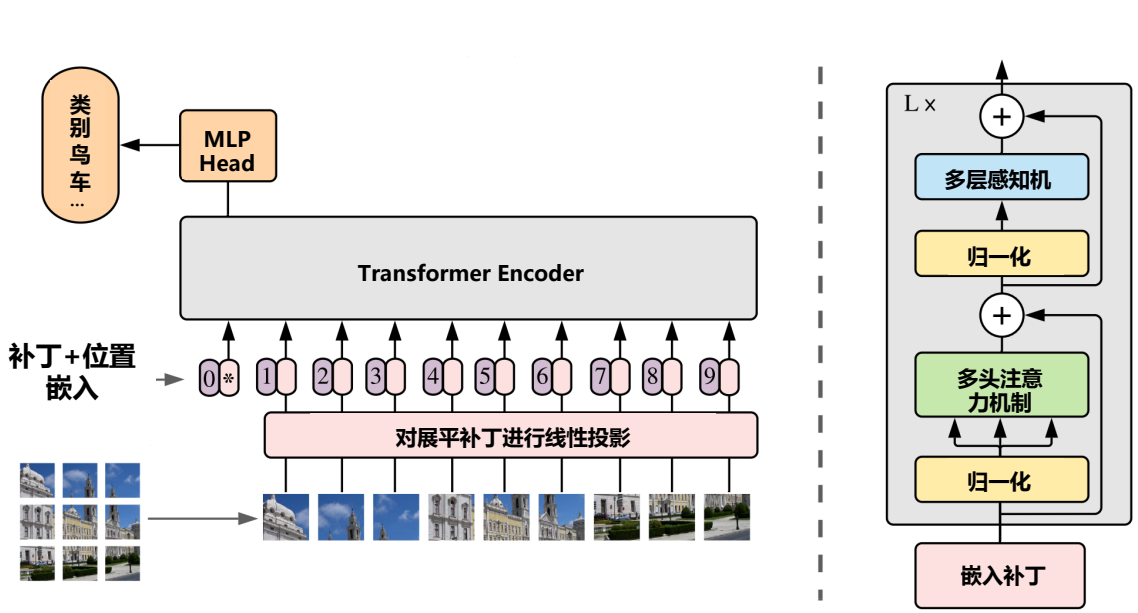


- **结合CNN与注意力机制：** RA-CNN 将卷积神经网络与注意力机制相结合，使模型能够**自动聚焦于图像中的关键信息区域**，避免了对冗余信息的过度处理，提升了计算效率和精度。
- **递归结构：** RA-CNN 引入了递归结构，采用递归计算，**逐步调整关注区域**，增强记忆与动态调整能力。
- **自适应特征提取：** RA-CNN 能够通过反复的**迭代调整对图像特征的关注**，从而实现更精准的特征提取。

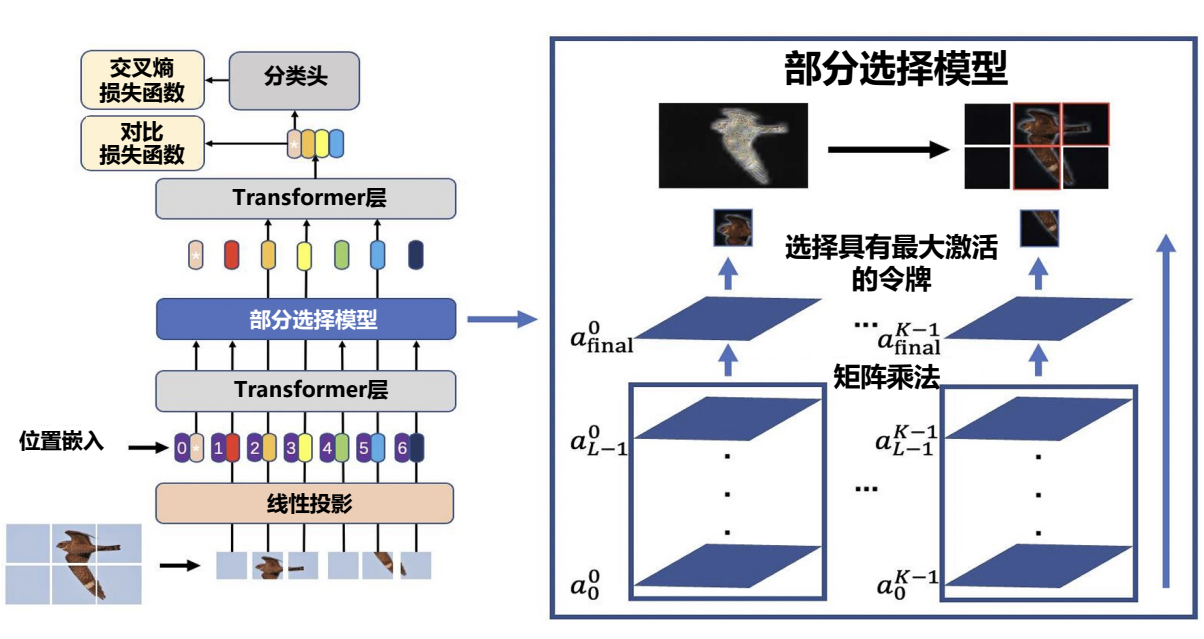
Jianlong Fu, Heliang Zheng, Tao Mei. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. CVPR 2017.

■ ViT模型的引入

➢ **Vision Transformer（ViT）模型**将Transformer架构引入图像分类，通过自注意力机制捕捉图像中的全局信息。不同于传统CNN，ViT将图像分割成若干小块，并将这些块作为序列输入Transformer。通过自注意力，ViT能够建模图像的全局空间关系，尤其适用于大规模数据集。



Vision Transformer（ViT）模型



TransFG模型

Alexey Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.
Junyu He et al. TransFG: A Transformer Architecture for Fine-Grained Visual Recognition. ICCV 2021.

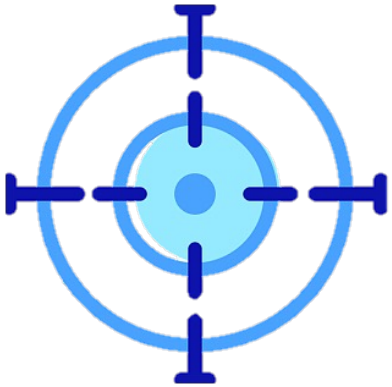
■ 挑战及未来的研究方向

- 细粒度图像分类的未来发展方向将围绕精度提升、效率优化、鲁棒性增强、数据依赖减少、应用场景扩展等方面展开。
- 例如，通过结合Transformer的全局信息捕捉能力与CNN的局部特征提取能力，可以进一步提升细粒度分类的性能。开发更加适合细粒度分类的大规模、高质量数据集将成为重点，数据集的提升对于模型的泛化能力至关重要。

特性	CNN(卷积神经网络)	Transformer
主要应用领域	图像处理、物体检测、图像分类	自然语言处理(NLP)、图像处理、序列任务
核心机制	卷积操作、池化操作	注意力机制(Self-Attention)
上下文捕获能力	局部上下文(通过感受野逐层扩大)	全局上下文(直接捕获任意位置间依赖关系)
并行化处理	卷积操作并行，但需逐层处理	自注意力机制完全并行处理
计算复杂度	通常较低，尤其是浅层网络	随着输入序列长度增加，计算复杂度增加
特征提取	固定卷积核，擅长局部特征提取	自适应特征提取，根据输入动态调整权重
模型参数量	较少(依赖卷积核数量和大小)	较多(依赖注意力头和层数)
训练速度	通常较快，适合硬件加速	并行性好，但因参数量大训练时间可能更长
处理长距离依赖	较弱(需多层卷积扩展感受野)	很强(直接通过自注意力机制处理)
数据结构假设	强(适用于有局部相关性的图像数据)	弱(适用于各种类型的数据)

更大规模的细粒度分类数据集

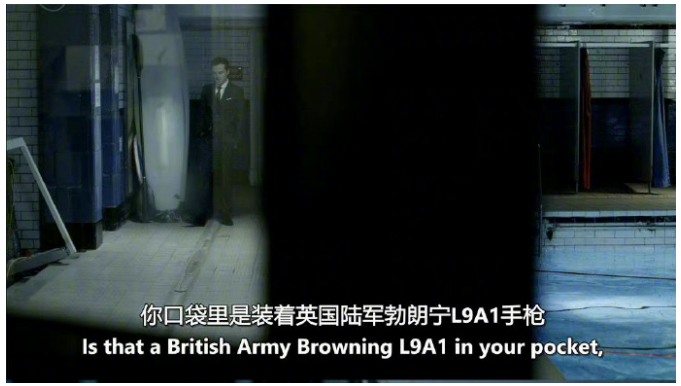
知识点4：AI侦探——目标检测



01 目标检测概述

02 目标检测经典方法

觅迹寻踪的福尔摩斯

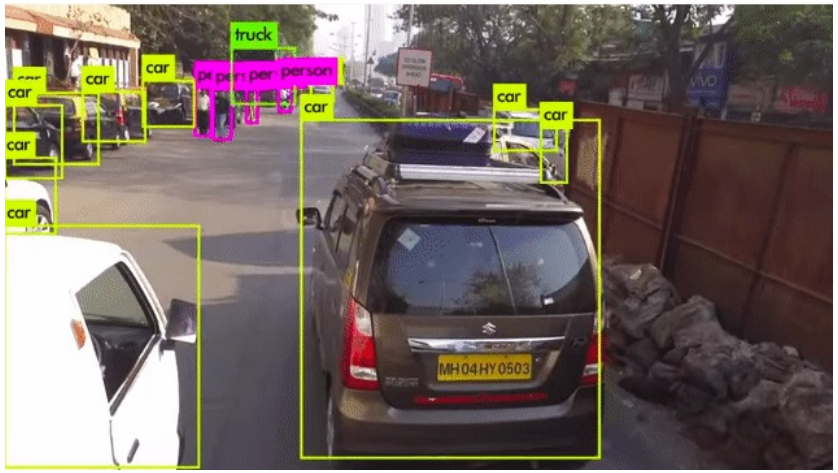


目标检测

目标检测概述

■ 目标检测的定义

目标检测旨在从图像或视频中识别出感兴趣的目标对象，并确定这些目标的**位置和类别**。

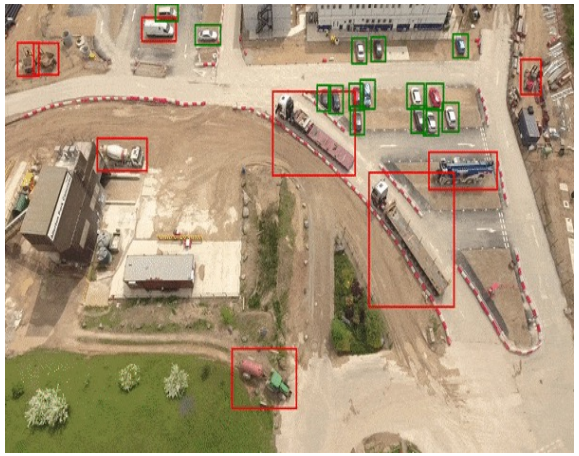


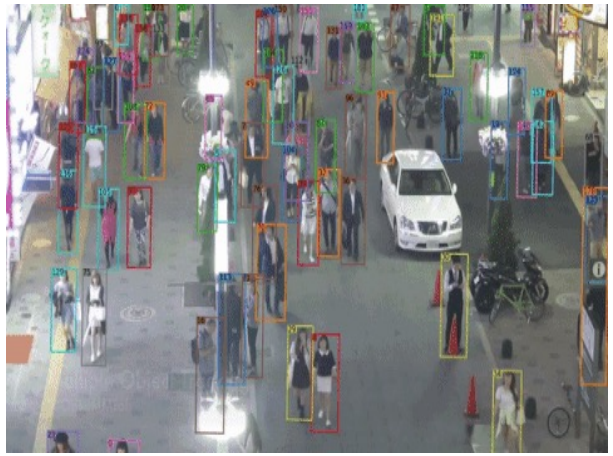
■ 目标检测与图像分类的区别

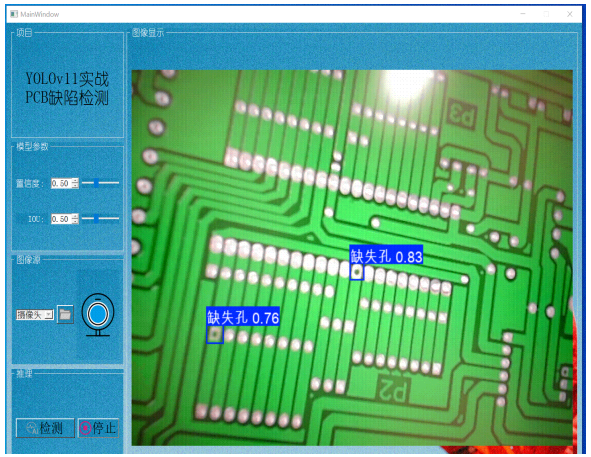
- **处理对象不同：**目标检测定位的是图像视频中的特定目标，图像分类则是对整幅图像进行分类。
- **输出结果不同：**目标检测输出的是目标的位置和类别，而图像分类仅输出图像所属类别。
- **算法复杂度不同：**目标检测需要更复杂的算法和更大的计算量。

■ 广泛的应用场景

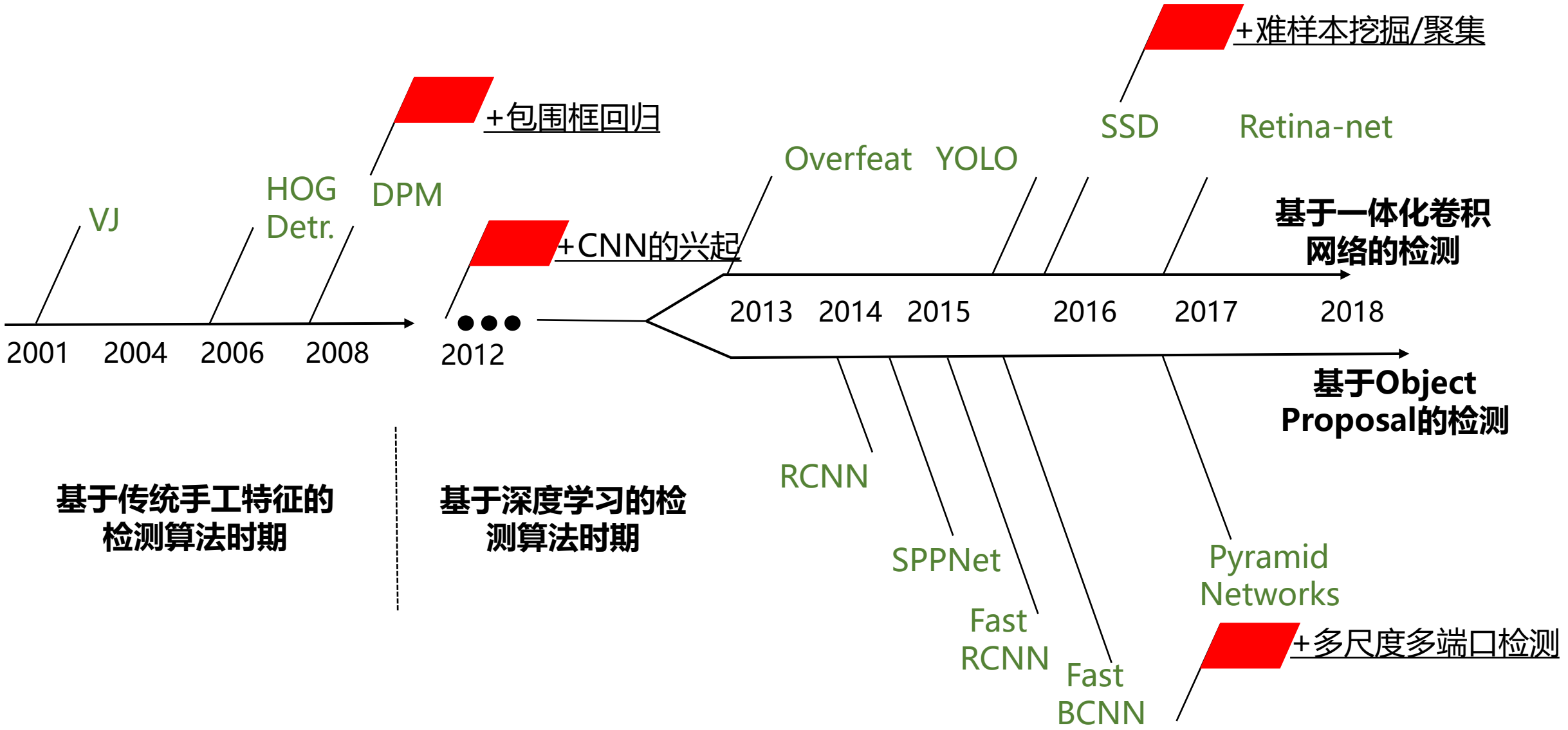
- **智能监控：** 对监控视频中的异常行为或特定目标进行识别和报警。
- **自动驾驶：** 识别道路上的车辆、行人、交通标志等，为自动驾驶提供决策依据。
- **医疗影像分析：** 通过目标检测技术在医学影像中识别病变区域，辅助医生进行诊断。
- **PCB板缺陷检测：** 快速定位缺陷位置，提高生产效率和产品质量。
- **对地观测分析：** 实现对地面目标的快速定位与分析，帮助规划部门决策。





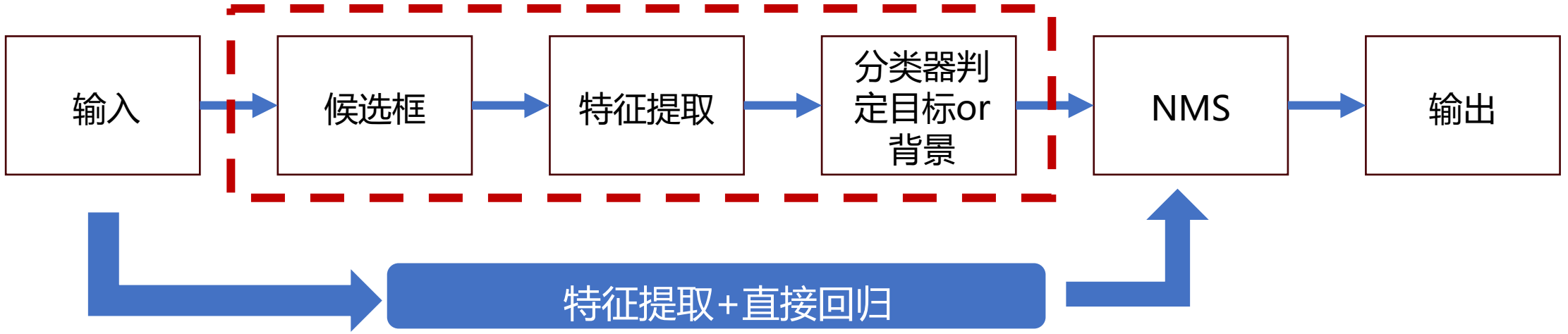


目标检测技术的发展


+难样本挖掘/聚集
+包围框回归
SSD
Retina-net
Overfeat
YOLO
HOG
Detr.
DPM
VJ
+CNN的兴起
基于一体化卷积网络的检测
2001
2004
2006
2008
2012
2013
2014
2015
2016
2017
2018
基于Object Proposal的检测
RCNN
SPPNet
Fast RCNN
Fast BCNN
Pyramid Networks
+多尺度多端口检测
基于传统手工特征的检测算法时期
基于深度学习的检测算法时期

➢ 传统目标检测方法

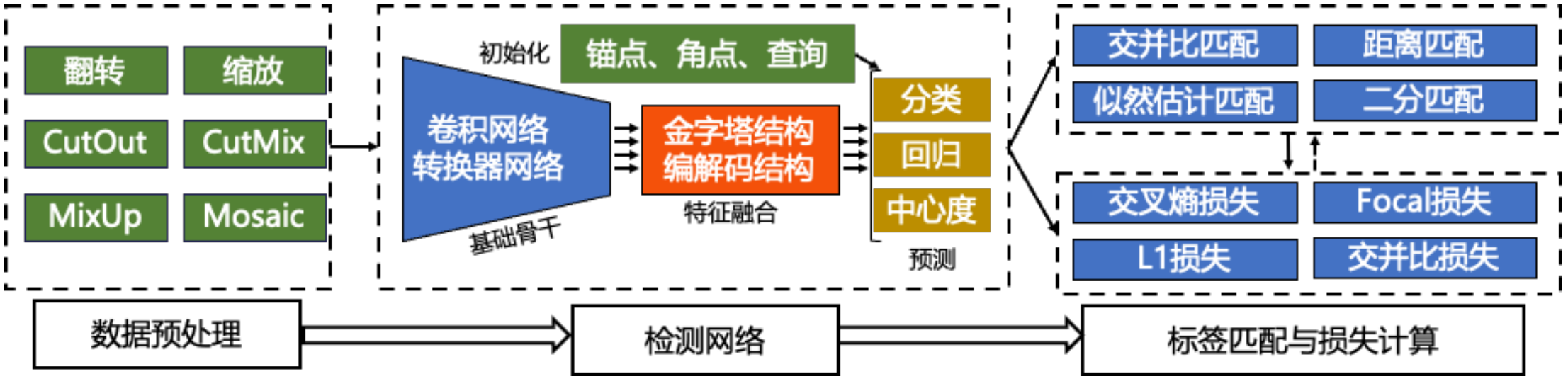
- 传统目标检测流程主要包括三个步骤：首先，通过某种策略**选取**可能包含目标的**候选区域**；其次，从这些区域中**提取特征**；最后，利用**分类器**对特征进行**识别与分类**。
- 传统方法依赖手工设计的特征，例如VJ检测器和HOG检测器。这些方法通过滑动窗口或计算图像的梯度方向直方图来提取特征，但受限于手工设计的特征，检测性能有限。



目标检测技术的发展

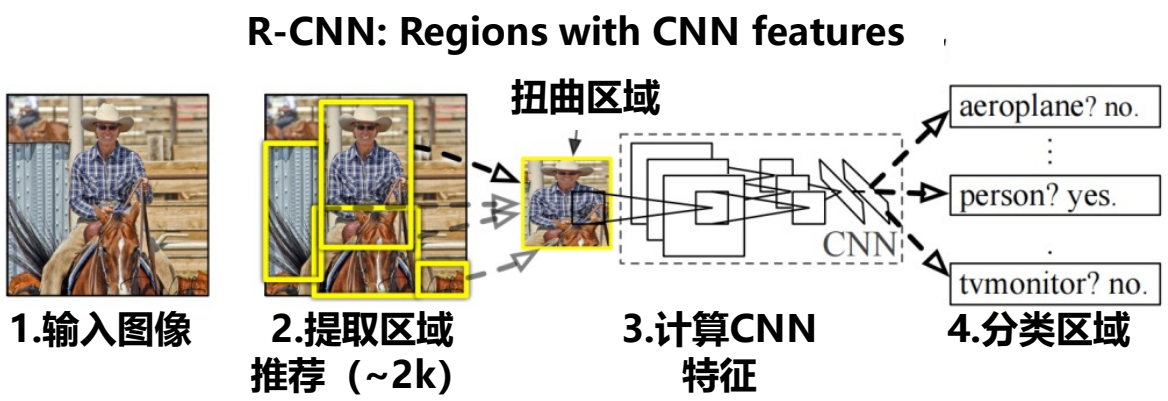
➢ 深度学习的兴起

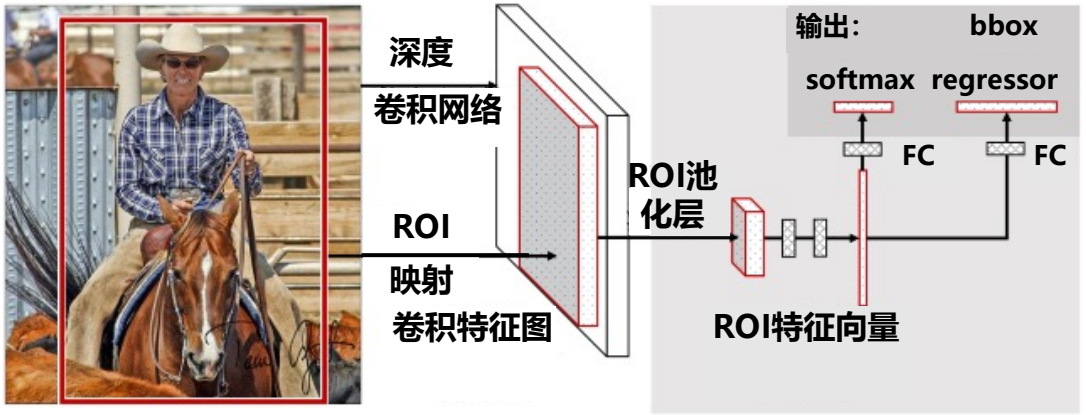
- 深度学习带来新突破，特别是R-CNN和YOLO等模型的出现。R-CNN采用**双阶段检测**方法，首先生成候选区域，然后对这些区域进行分类。YOLO系列则通过**单阶段**的方式提高检测速度。
- 随着Transformer技术的崛起，DETR等新型检测模型将强大的**序列建模能力**引入目标检测领域，展现了超越传统CNN方法的检测精度。



目标检测经典方法

■ 两阶段算法——R-CNN与Fast R-CNN





- **首次将卷积神经网络（CNN）引入目标检测任务，**引入了**区域建议机制**（如Selective Search），通过生成候选区域（Region Proposals）来减少搜索空间，从而在保证检测精度的同时提高了效率。
- **推动端到端训练：**虽然R-CNN本身非完全端到端训练框架，但它为目标检测的端到端训练方法奠定了基础。“**CNN特征提取+支持向量机分类**”模式开启了以深度学习为基础的目标检测新时代。

- **端到端训练：**通过统一的网络架构实现**端到端训练**，提高了训练效率。
- **共享卷积特征：**通过**一次卷积提取图像特征**，避免了R-CNN中每个区域独立计算特征，提升了效率。
- **ROI池化层：**引入**ROI池化层**，从不同大小的候选区域提取固定大小的特征图，简化计算。

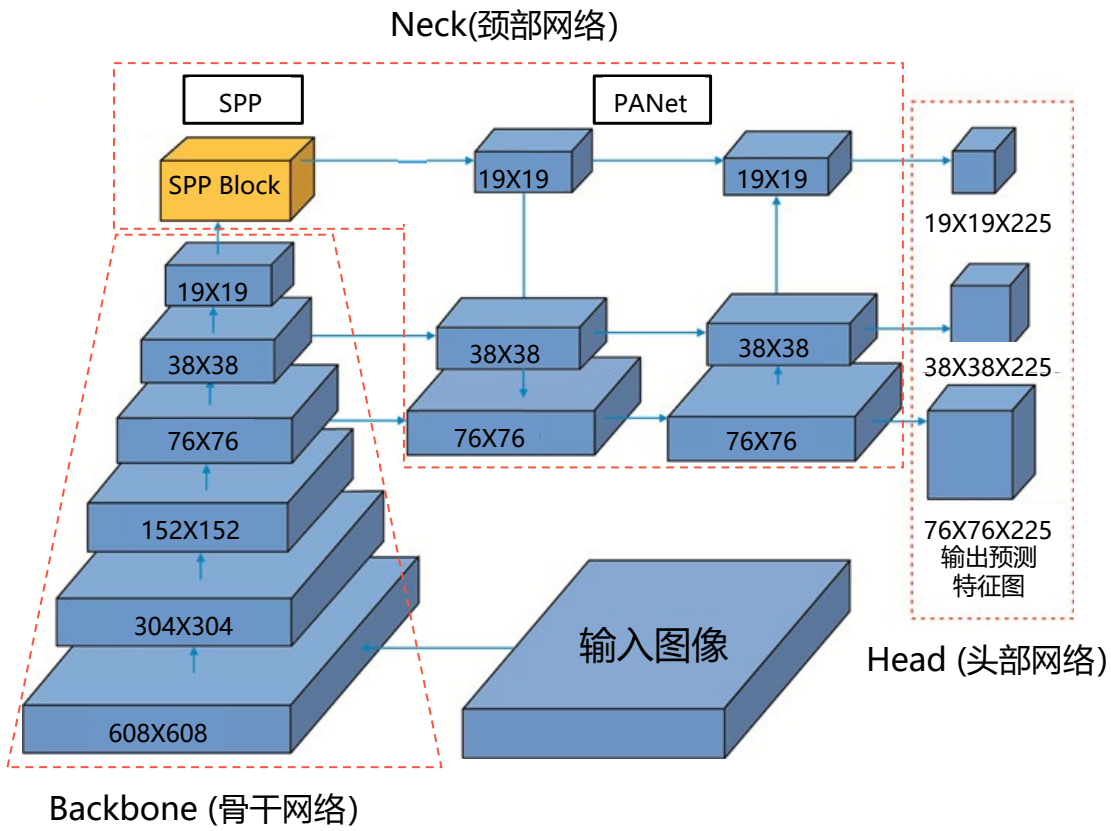
Ross Girshick. Fast R-CNN. ICCV 2015.

Ross Girshick et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR 2014.

■ 单阶段算法——YOLO（You Only Look Once）系列

YOLO是一种单阶段目标检测模型，结合了高效的检测速度和较高的精度。其主要贡献包括：

- **实时目标检测**：YOLO系列通过将目标检测任务**转化为回归问题**，显著提升了检测速度，使得目标检测成为**实时**任务。这一创新使得YOLO在视频处理和实时监控中表现出色。
- **端到端训练**：YOLO提出了**端到端训练**的概念，模型从原始图像输入到最终的目标检测输出，所有步骤都在一个统一的网络中完成。

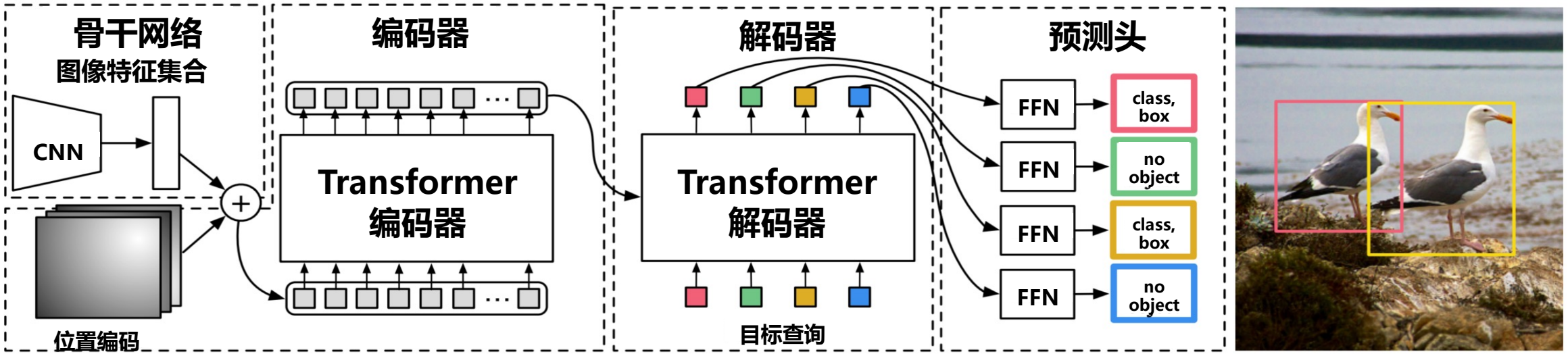


- **全局信息利用信息**：YOLO**一次性处理整张图像**，利用全图信息进行预测，增强了空间关系理解。
- **多尺度建模**：通过不同层级的特征图进行**多尺度检测**，能够有效识别不同大小的目标，提升检测的鲁棒性。
- **平衡精度与速度**：成功**平衡了检测精度和速度**，适用于高效处理大量图像或视频。

Joseph Redmon et al. You Only Look Once: Unified, Real-Time Object Detection. CVPR 2016.

目标检测经典方法

■ 基于Transformer的检测——DETR（Detection Transformer）



- **首个将Transformer应用于目标检测任务的模型**：突破了传统方法（如R-CNN系列）对卷积神经网络的依赖，**使用自注意力机制进行全局上下文建模**，从而在目标检测中提高了长距离依赖的捕捉能力。
- **端到端训练**：通过自注意力和匹配损失的设计，DETR能够**直接从原始图像预测最终的目标位置和类别**。
- **简化目标检测流程**：DETR避免了复杂的区域提议生成、非极大抑制等步骤，**采用了直接预测框架**，使得目标检测任务的处理更加简洁和高效。
- **局限性**：DETR也存在**训练时间较长**、对**计算资源要求高**以及在**小目标检测上性能有待提升**等缺点。

Nicolas Carion et al. End-to-End Object Detection with Transformers. ECCV 2020.

目标检测的挑战与展望

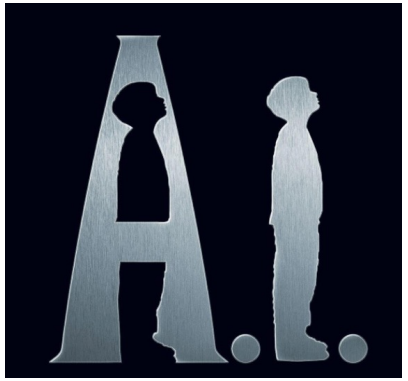
■ 目标检测的挑战

- **目标的多样性：**小目标在图像中占的像素较少，特征信息不足导致检测困难；而遮挡目标会丢失关键特征，增加检测难度。低对比度目标由于与背景区分度低，也容易漏检或误检。
- **环境的多变性：**光照变化、动态背景、恶劣天气等因素会干扰检测效果。目标检测算法需要在这些复杂环境中保持稳定性和鲁棒性。
- **计算资源的高要求：**深度学习模型通常需要大量的计算资源，这限制了它们在资源受限环境中的应用。如何平衡实时性和高精度也是当前研究的重点。

■ 未来发展方向

- **算法的轻量化与高效性：**随着嵌入式设备和移动设备的普及，研究如何在资源有限的情况下实现高效目标检测将成为重要方向。未来的检测算法应更加轻量化，以满足实际应用需求。
- **数据标注自动化：**大规模、高质量的数据集是训练高精度目标检测模型的基础，然而数据标注的高成本是一个主要问题。研究自动化标注工具将有助于降低数据标注的成本，促进目标检测技术的发展。

知识点5：剪影艺术家——图像分割



01 图像分割概述

02 图像分割经典方法

剪影艺术

剪影艺术是一种通过简洁轮廓展现物体形态的艺术形式，它以最简单的线条勾勒出主体的外形，舍弃细节和色彩，展现出一种纯粹而强烈的视觉效果。这种艺术形式历史悠久，在中国和西方都有深厚的文化根基，体现了极简美学的魅力。

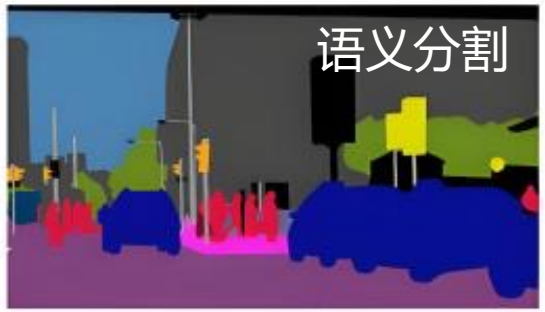
剪影艺术在现代社会中依然具有广泛的应用和深远的影响。它不仅被广泛应用于传统的民间艺术、装饰艺术和节日庆典中，还在现代设计、摄影、数字艺术等领域展现出强大的生命力。

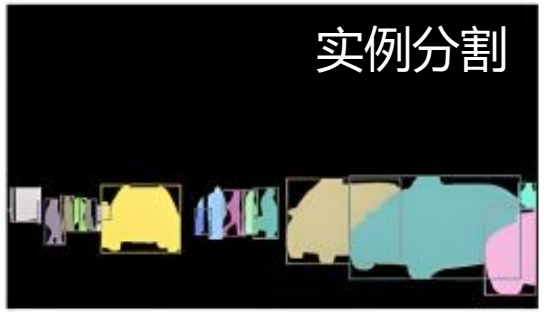
在计算机视觉中，剪影艺术的原理也可被借鉴用于**图像分割技术**，通过对物体轮廓的提取和分析，实现图像中不同区域的划分和识别。

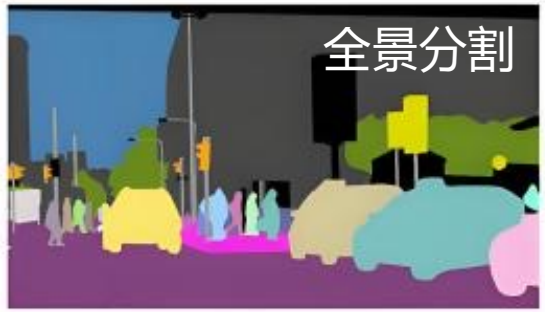
■ 图像分割的定义与分类

图像分割旨在将图像划分为多个具有特定意义的区域，**实现像素级别的分类和定位**。与图像分类和目标检测相比，图像分割更加细致。

- **语义分割**聚焦于识别图像中的各类物体或背景，并统一划分至相应类别，但不区分同一类别中的不同个体。
- **实例分割**不仅识别不同类别的物体，还能区分同一类别下的不同实例，为每个实例赋予独特标记，但通常不处理背景区域。
- **全景分割**则是语义分割与实例分割的完美结合，既识别并分割出不可数的背景元素，也精确区分并标记出可数的前景物体实例，实现了对整个场景的全景式解析。

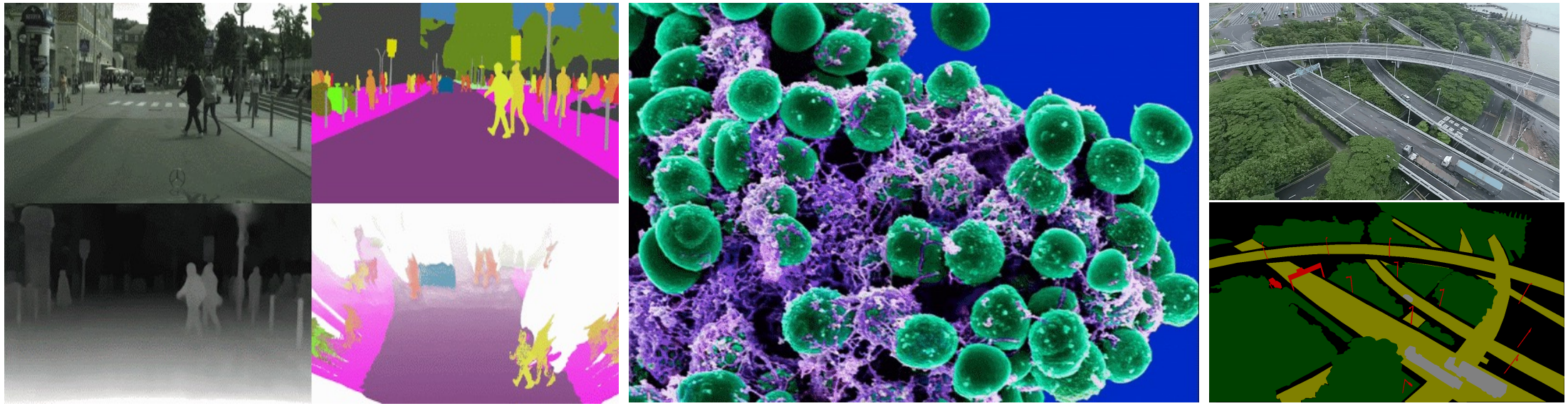






■ 图像分割的应用

- **自动驾驶系统：**图像分割技术可以将摄像头捕捉到的图像划分为不同的语义区域，如道路、车辆、行人等，为自动驾驶系统准确地提供这些环境信息。
- **医学领域：**图像分割技术可以识别和分析病变组织，辅助医生诊断决策。
- **遥感影像处理领域：**通过图像分割技术，可以将遥感图像中的不同地物类型（如水体、森林、城市等）区分开来，为环境监测、城市规划、灾害预警等提供科学依据。

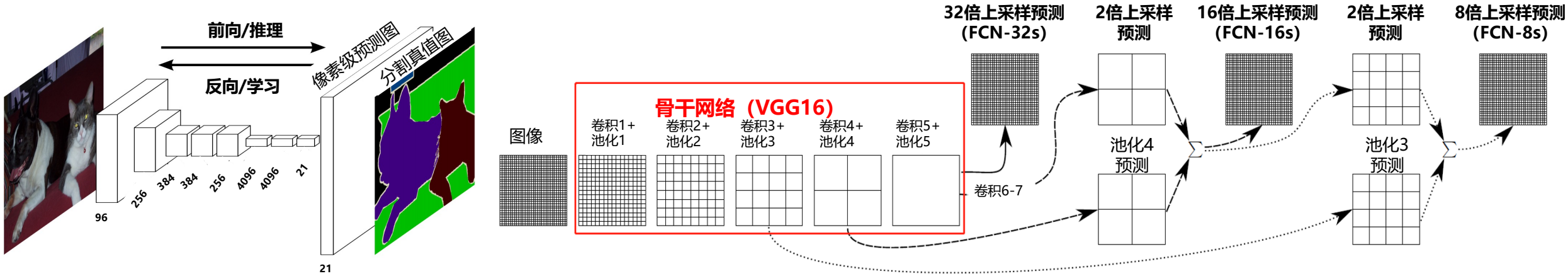


经典算法介绍

■ 算法分类

- 算法分类
  - 基于全卷积网络的模型
    - **特点**：将传统卷积神经网络中的全连接层替换为卷积层，实现端到端的像素级预测。通过上采样（如反卷积）和跳跃连接来恢复空间分辨率和细节信息。
    - **代表模型**：FCN（Fully Convolutional Networks）、U-Net等。
  - 编码器-解码器结构
    - **特点**：编码器部分通过卷积和池化操作提取图像特征，降低空间分辨率；解码器部分通过上采样和卷积操作逐步恢复空间分辨率和细节信息。
    - **代表模型**：SegNet、PSPNet等。
  - 空洞卷积模型
    - **特点**：在不增加参数数量和计算量的前提下，通过空洞卷积增加感受野，保留更多的空间信息，有助于分割细节。
    - **代表模型**：DeepLab系列（如DeepLabv3+、DeepLabv3）等。
  - Transformer模型
    - **特点**：通过自注意力机制有效捕捉了全局上下文信息，提高了分割精度。
    - **代表模型**：SETR、 SegFormer等。

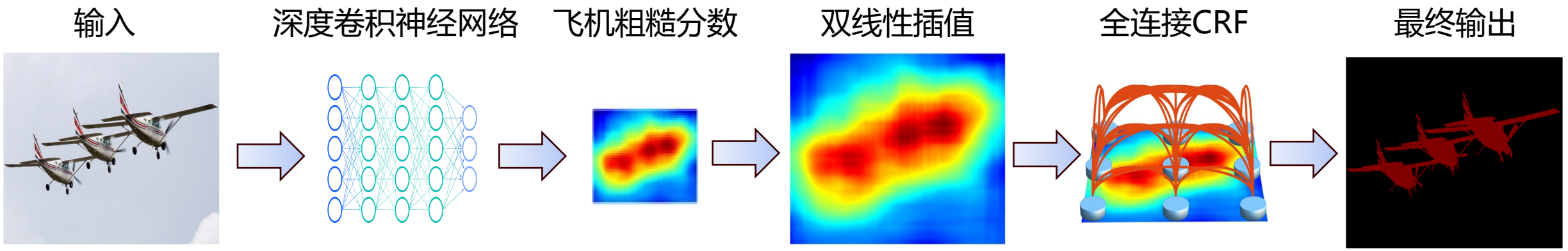
■ 开山之作——全卷积网络（Fully Convolutional Network, FCN）



- **首个将CNN应用于像素级预测的模型**：它通过**卷积操作替代了传统全连接层**，使得网络能够处理图像的每个像素，进行精确的图像分割。
- **端到端训练：允许模型从输入图像直接输出像素级的标签**，提高了训练效率。
- **引入上采样和跳跃连接：**通过**结合低层次的特征和高层次的语义信息**，提高了分割精度。
- **无需手工特征提取：**无需人工设计特征，**自动学习图像中每个像素的特征**，提高了分割精度。
- **局限性**：对图像细节信息的捕捉较弱，分割边界模糊。

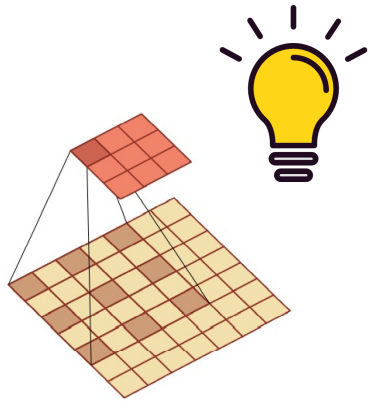
Jonathan Long, Evan Shelhamer, Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015: 3431–3440.

■ 经典方法——DeepLab系列模型



- **首次引入空洞卷积：**增加了卷积核的感受野，提升了对上下文信息的捕捉能力，同时避免了计算量的急剧增加，显著提高了图像分割精度。
- **条件随机场（CRF）后处理：**使用CRF对分割结果进行后处理，细化边界，使得分割更加精确，特别是在物体边缘的处理上效果显著。

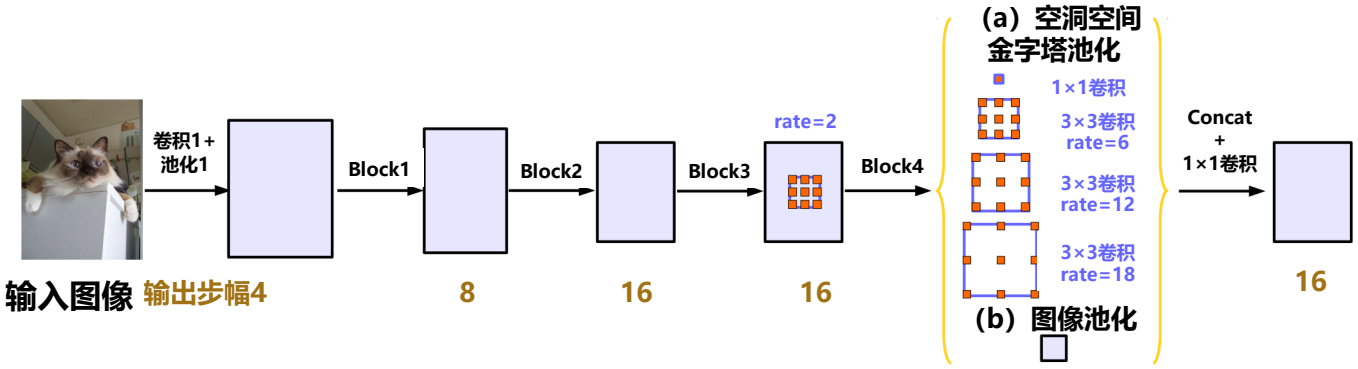
空洞卷积（也称扩张卷积）是一种特殊的卷积操作，通过在卷积核中插入空洞（间隔），扩大卷积核的感受野，从而捕获更大范围的上下文信息，而无需增加额外的计算量或参数。

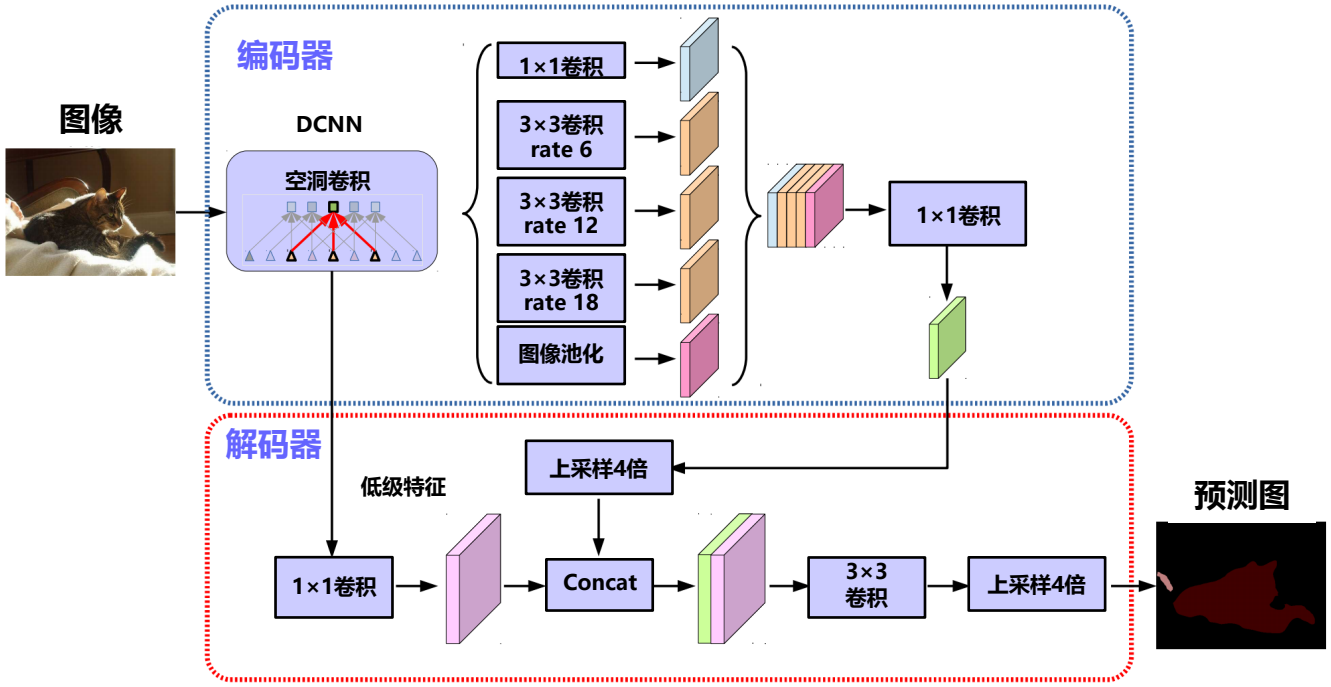


空洞卷积示意图

Liang-Chieh Chen, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE TPAMI 40(4): 834-848, 2018.

■ DeepLab v2、v3及v3+的进阶



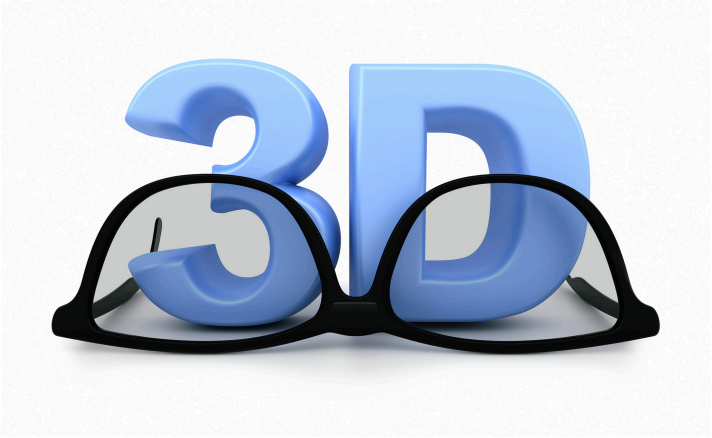


DeepLab v2：改进的空洞卷积：引入了多个空洞卷积层构成**空间金字塔池化模块（ASPP）**，增强了不同尺度的信息提取能力，进一步提升了图像分割的精度和鲁棒性。

DeepLab v3：空洞空间金字塔池化（ASPP）的优化：引入**并行多分支结构**，同时引入全局平均池化分支，直接提取图像级全局上下文特征。**简化后处理流程：**通过端到端的深度网络直接输出高质量分割结果，**减少了对CRF的依赖**。

DeepLab v3+：编码器-解码器结构的融合：新增的解码器模块通过上采样逐步恢复空间细节，结合编码器中的浅层特征，显著提升边界精度。**深度可分离卷积的引入：**将标准卷积分解为深度卷积和点卷积，大幅减少参数量和计算量。

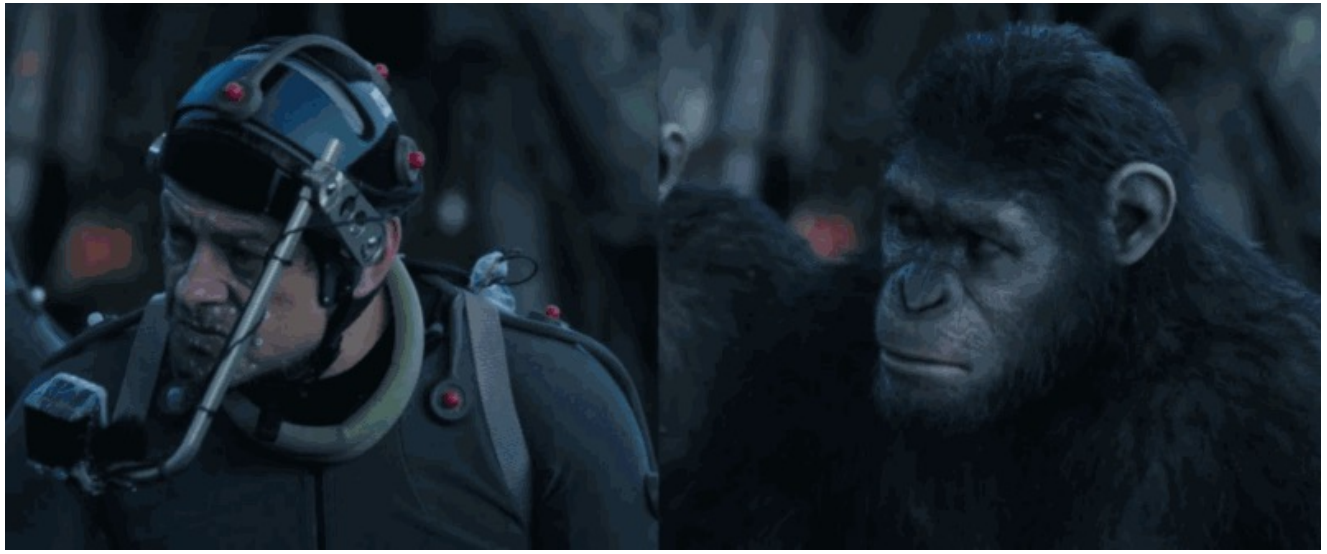
知识点6：
揭秘阿凡达中的三维视觉



01 三维视觉概述

02 典型三维视觉任务

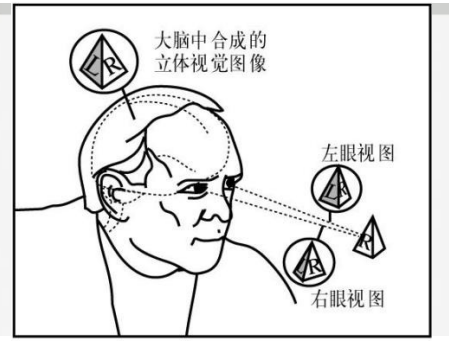
电影中的3D特效



- 2009年上映的《阿凡达》凭借革命性的**3D技术**成为全球票房冠军。影片60%的镜头采用计算机生成图像（CGI），结合“协同工作摄像机”和动作捕捉技术，将演员表演与虚拟角色无缝融合，增强了画面的立体感和沉浸感。此外，先进的面部捕捉技术通过头戴设备记录演员的细微表情，并映射到虚拟角色上，使CG角色的表情和动作更加逼真。
- 在计算机视觉领域，三维视觉技术同样致力于提取三维信息，通过多个视角的图像融合来感知深度和立体感。

■ 三维视觉的目标

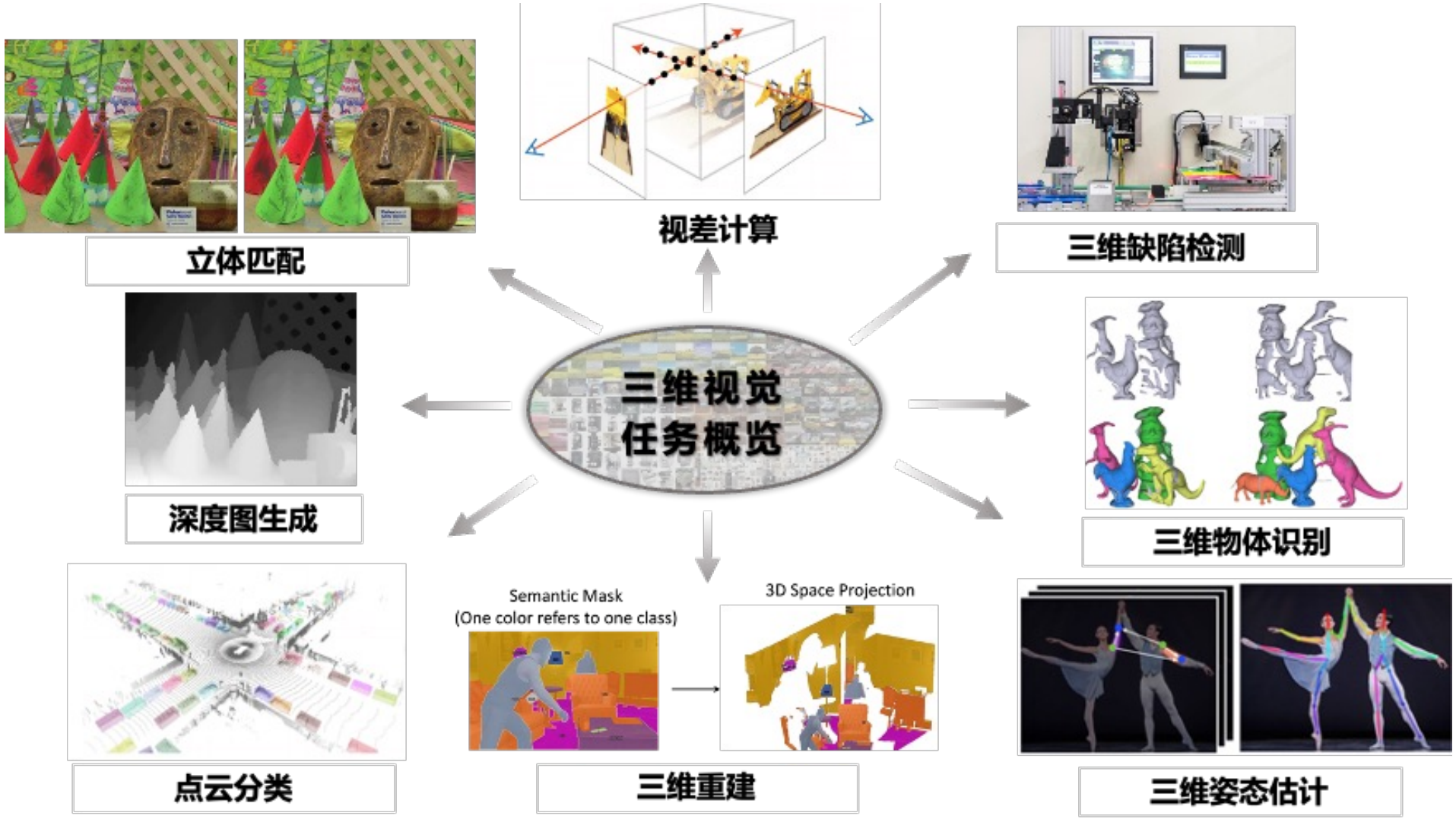
三维视觉通过摄像头从不同角度获取同一场景的多幅图像，利用这些图像之间的**视差信息**重建出场景的三维结构或深度信息。这模仿了人类双眼的视觉原理，使机器能够像人类一样感知三维空间中的物体和距离。



■ 三维视觉技术任务

- 三维视觉技术的实现涉及多个关键任务，包括图像采集、预处理、特征提取、立体匹配、视差计算、深度图生成和三维重建。
- 立体匹配是确定图像间像素点对应关系的基础，视差计算和深度图生成则将这些匹配结果转化为空间深度信息，直接影响三维重建的精度。
- 三维重建用于建立物体或场景的数学模型，广泛应用于自动驾驶、机器人导航、智能制造等领域。

■ 三维视觉任务概览


三维视觉
任务概览
立体匹配
视差计算
三维缺陷检测
深度图生成
三维物体识别
点云分类
Semantic Mask
(One color refers to one class)
3D Space Projection
三维重建
三维姿态估计

典型三维视觉任务——深度估计

■ 深度估计

- 深度估计是从二维图像中推测出场景中物体的三维深度信息。传统方法依赖激光扫描或结构光技术，而基于图像的深度估计无需高昂设备，仅需普通相机即可实现。
- 深度估计技术在增强现实、自动驾驶和机器人导航等领域广泛应用，支持虚拟对象融合、环境理解、路径规划等任务，助力各行业的创新与发展。

典型三维视觉任务——深度估计

■ 深度估计分类

➢ 基于单目视觉的深度估计

从单张二维图像中推断深度，由于缺乏双眼视差等几何线索，任务难度较大。

➢ 基于双目立体视觉的深度估计

通过双目摄像头获取不同视角的图像，利用视差原理推断深度。

左相机图　　右相机图

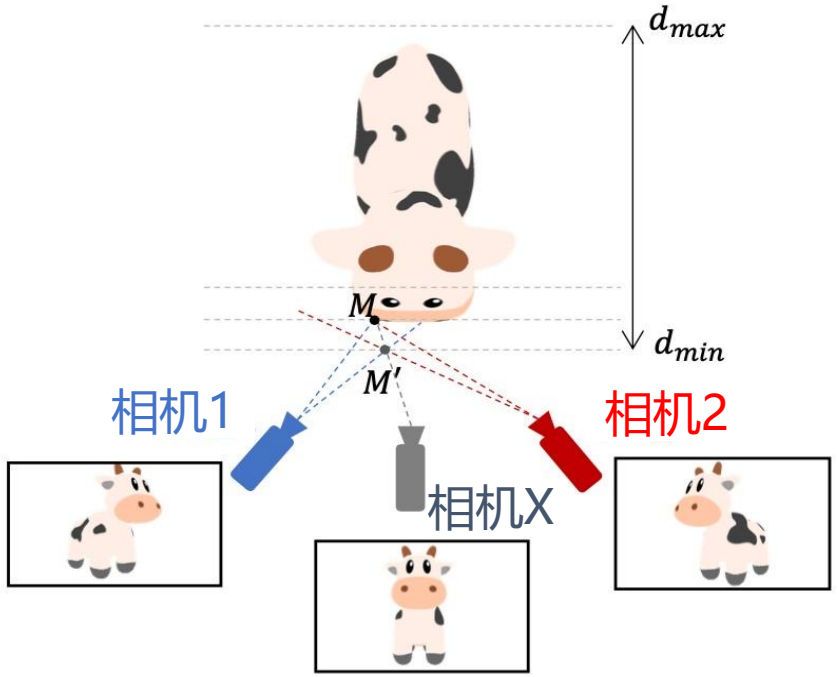
视差图

典型三维视觉任务——深度估计

■ 深度估计分类

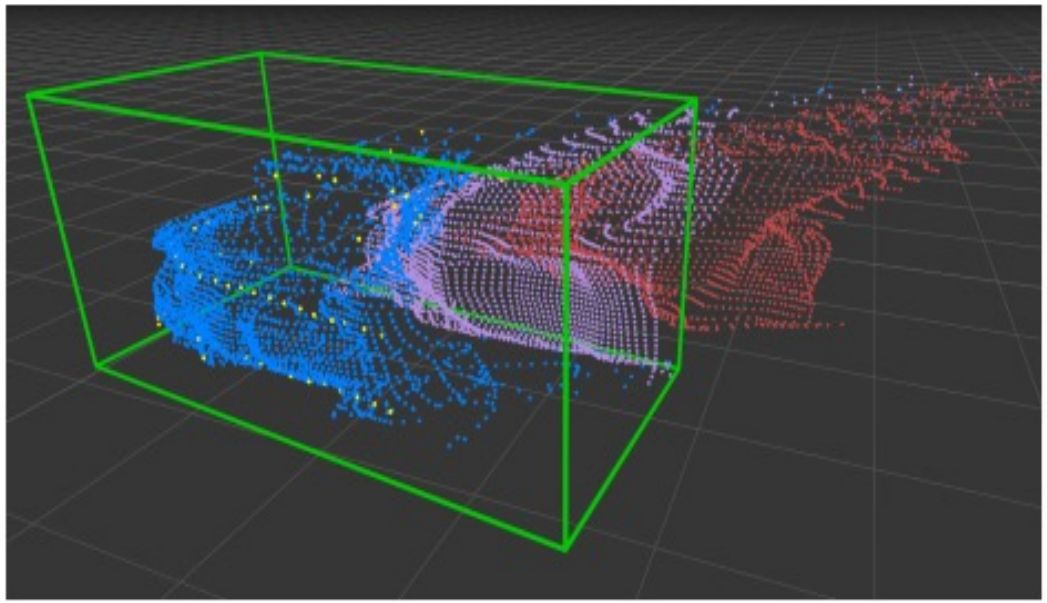
➢ 基于多视图的深度估计

利用多个不同角度的图像进行三维深度推断，提供比双目系统更丰富的几何信息。



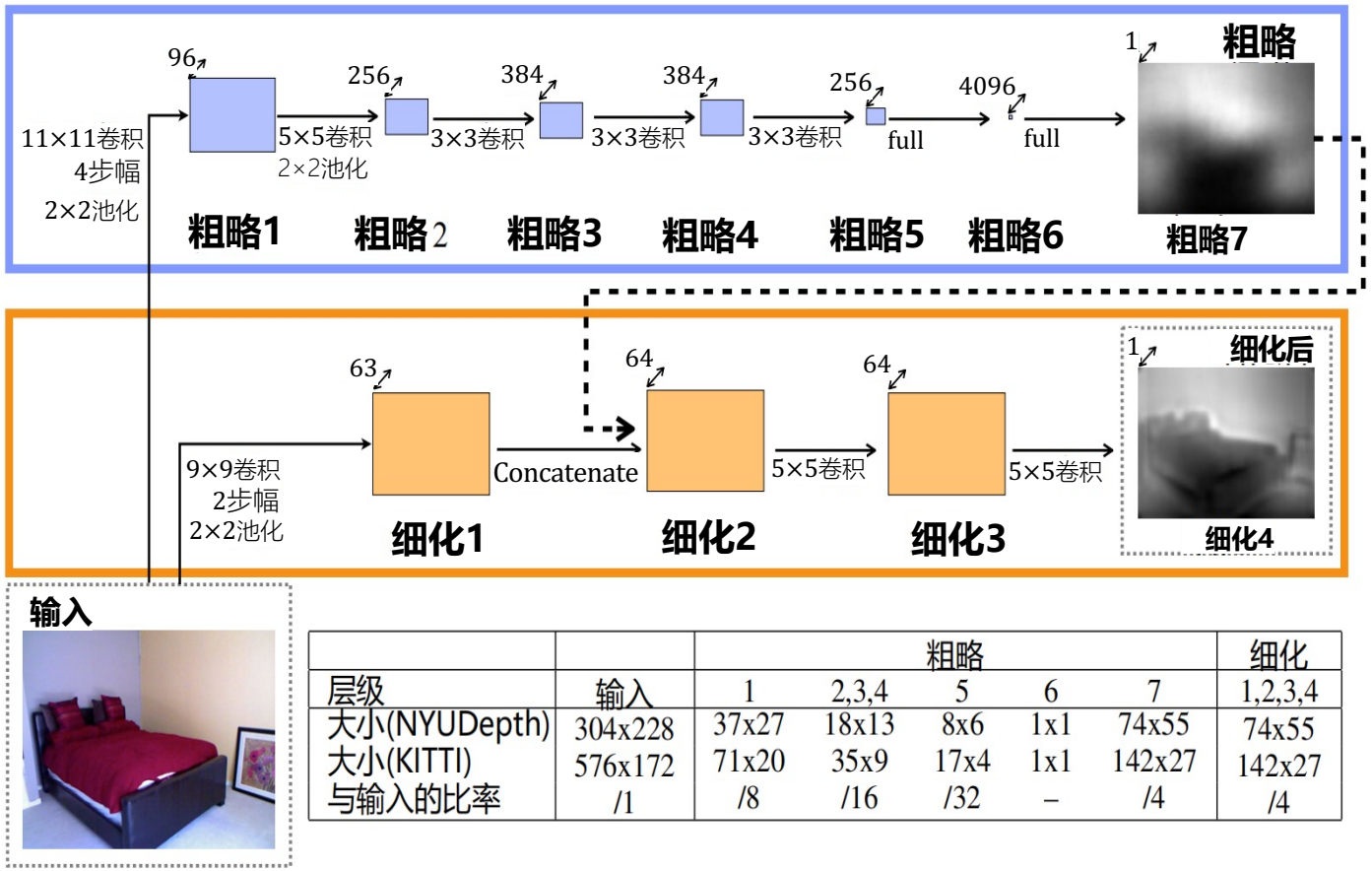
➢ 基于激光雷达的深度估计

通过激光测距直接提供精确的深度信息，但由于激光雷达数据稀疏，通常结合图像数据进一步优化。



典型三维视觉任务——深度估计

■ 经典方法——基于深度学习的单目深度估计网络



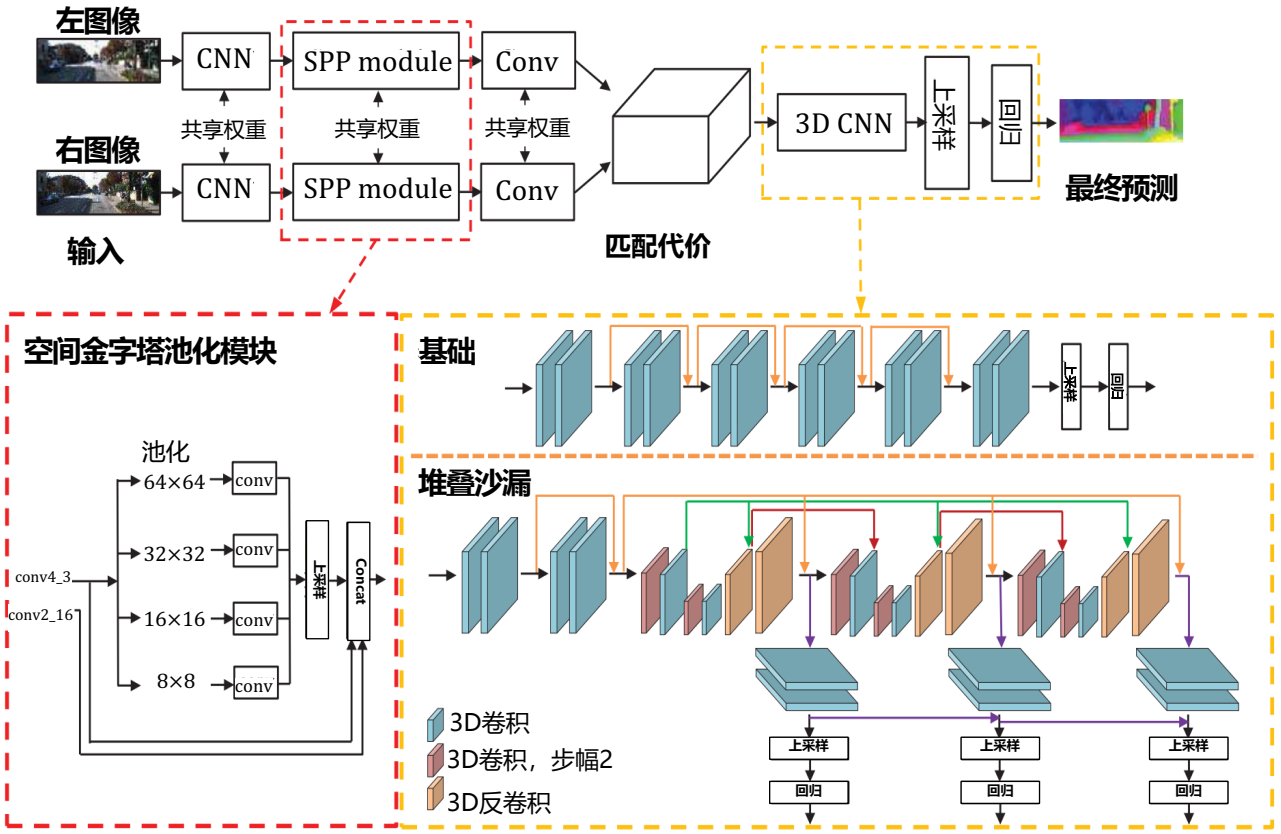
		粗略					细化
层级	输入	1	2,3,4	5	6	7	1,2,3,4
大小(NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
大小(KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
与输入的比率	/1	/8	/16	/32	–	/4	/4

- **首个深度学习单目深度估计模型：**首次利用CNN实现直接从单张RGB图像预测深度图，开创了**数据驱动**的深度估计新范式。
- **多尺度架构：**结合全局粗尺度网络及局部细尺度网络，解决了**全局与局部**深度信息的融合问题。
- **损失函数：**使用**尺度不变误差**作为损失函数，减少深度值的全局尺度不确定性。通过**梯度匹配**优化深度图的平滑性和边界清晰度。

David Eigen, et al: Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. NIPS 2014: 2366-2374.

典型三维视觉任务——深度估计

■ 经典方法——金字塔双目深度估计网络



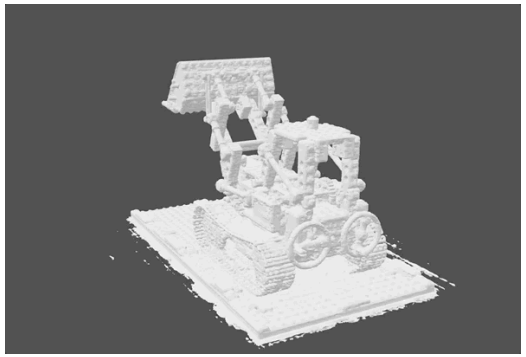
- **空间金字塔池化模块：**通过不同尺度的池化操作**获取丰富的上下文信息**，使模型能更好地理解全局和局部的深度信息，提高深度估计的准确性。
- **3D代价体卷积：**通过构建 3D 代价体并在其上应用3D卷积神经网络，**高效地聚合双目立体匹配的信息**，增强了深度估计的鲁棒性和精度。
- **端到端训练：**实现了从双目图像输入到最终深度预测的端到端学习，**减少了传统立体匹配方法中的繁琐后处理**，提高了计算效率和易用性。

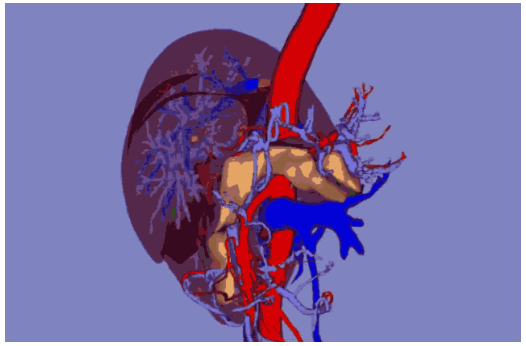
Jia-Ren Chang, Yong-Sheng Chen: Pyramid Stereo Matching Network. CVPR 2018: 5410-5418.

典型三维视觉任务——三维重建

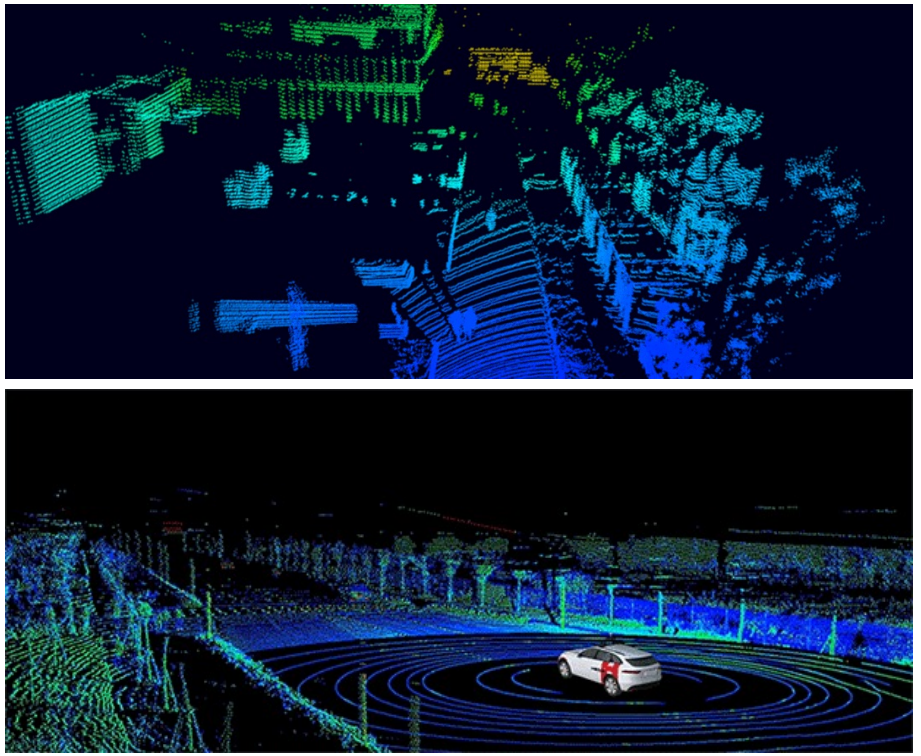
■ 三维重建

- **目标** 捕捉环境中物体的立体信息，从而确定不同物体的空间位置，并构建相应的空间模型。
- **分类**
  - 主动式三维重建方法
  - 被动式三维重建方法
- **应用领域**
  - 建筑和工程领域——创建建筑模型、施工前进行规划和模拟、建筑的修复和保护工作
  - 电影和视频游戏产业——创建逼真的三维角色和环境、通过扫描真实场景快速生成三维模型
  - 工业产品质量控制——自动化质量控制





■ 主动式三维重建方法



➢ 基于激光雷达的三维重建

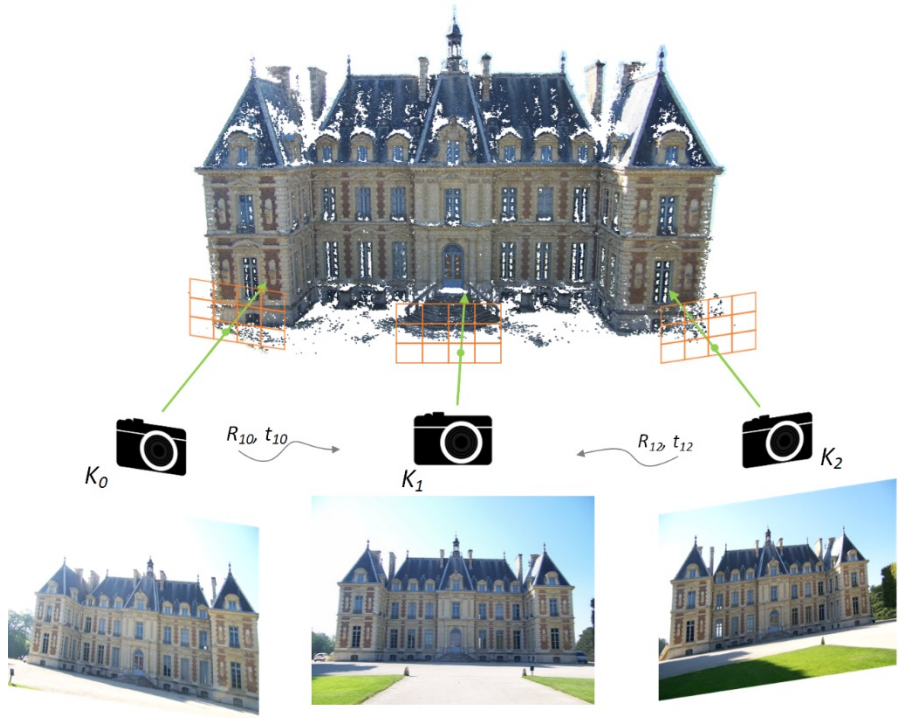
□ 通过**发射激光束并测量其反射回传感器的时间差**，计算物体与传感器的距离，生成三维点云数据。这些数据精度高、范围广，适用于大范围户外场景。然而，激光雷达在恶劣天气下性能可能受限，且生成的点云数据较为稀疏，需要后续处理。

➢ 基于结构光的三维重建

□ 通过**投射特定光栅图案（如条纹）至物体表面**，观察并分析图案变形来恢复三维形状。数据精度高，尤其适合重建表面细节丰富、几何形状复杂的物体。然而，光栅变形对大场景或远物体不明显，可能导致深度估计误差，强光环境也可能干扰捕捉。

典型三维视觉任务——三维重建

■ 被动式三维重建方法



➢ 单目图像法

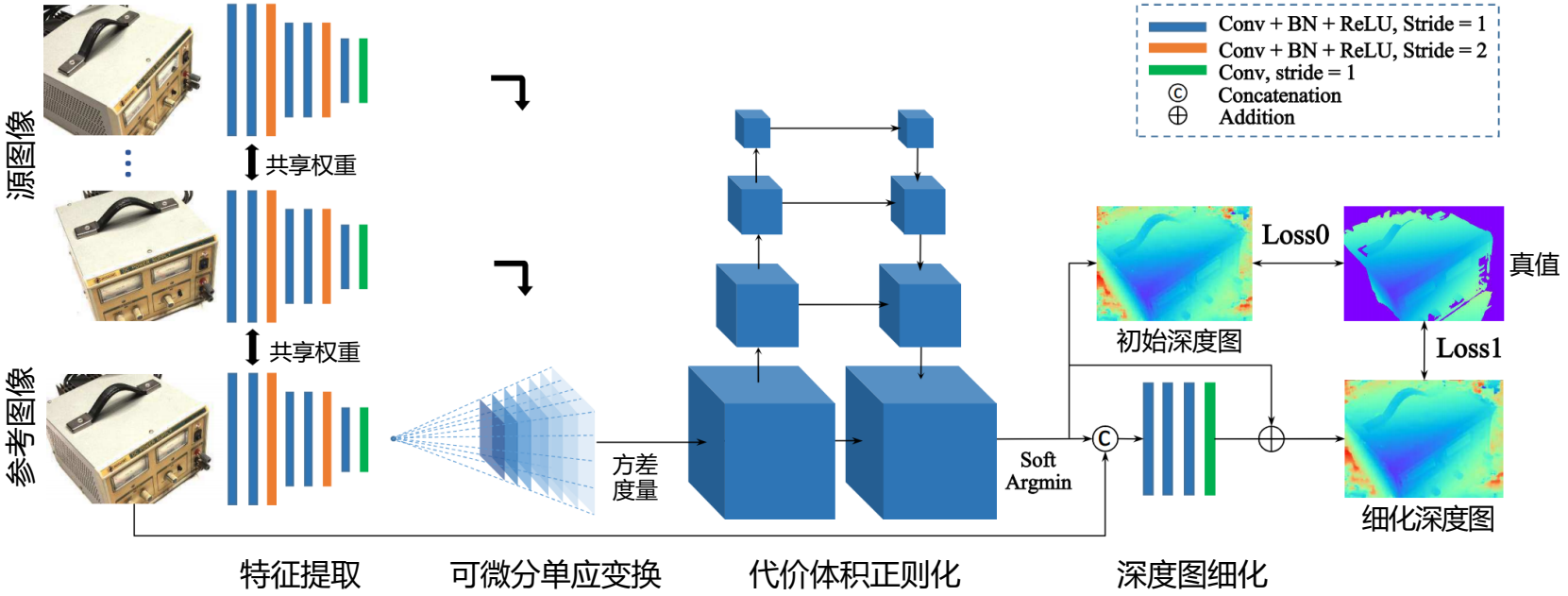
- 通过从**单一图像**中提取特征点，结合相机的内外参信息，恢复物体的三维结构。特点是计算效率高、成本较低，但精度有限，容易受到视角变化和噪声干扰，适用于简单场景的三维重建。

➢ 多视图法

- 通过从**多个视角拍摄同一场景的图像**，利用图像之间的几何关系（如视差和匹配点）来恢复三维信息。该方法精度较高，适用于复杂场景的重建，尤其在细节和深度估计上表现优越。然而，处理多视角图像需要较高的计算量，并且对匹配点的准确性要求较高。

典型三维视觉任务——三维重建

■ 经典方法——多视图立体匹配网络MVSNet

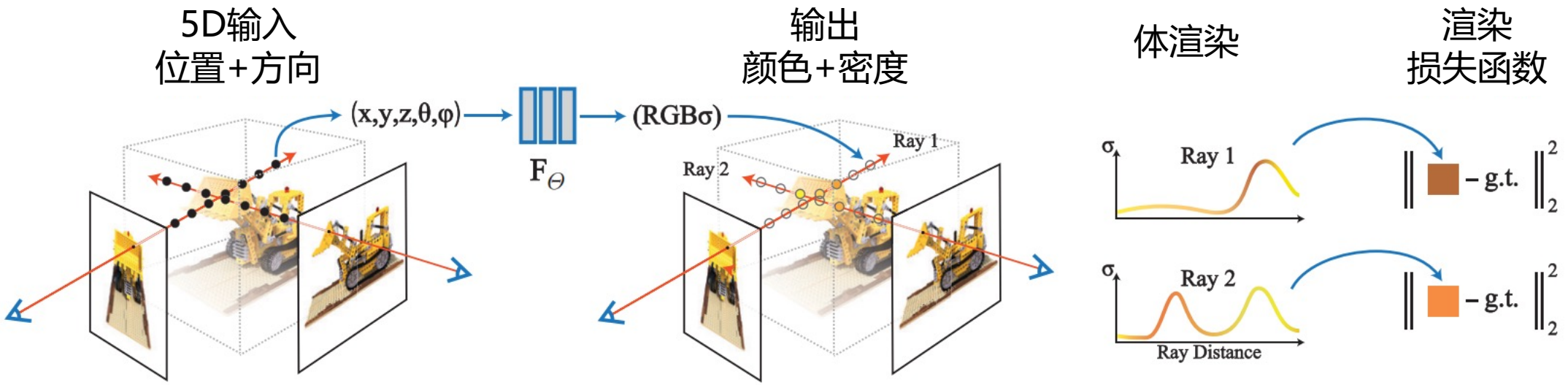


- **首个端到端的多视图立体匹配模型：**无需手工设计匹配代价和后处理步骤，极大简化了流程。
- **3D代价体构建与正交视差回归：**将多个视角图像合成到一个统一的视角中，然后像拼图般组合这些信息以构建**三维“成本模型”**，并利用深度学习方法估算出物体的深度。
- **变分深度采样策略：**MVSNet 提出了**自适应深度范围**的采样策略，能够**动态调整深度搜索空间**，使得模型在不同场景下都能获得更精准的深度估计。

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, Long Quan. MVSNet: Depth Inference for Unstructured Multi-View Stereo. ECCV 2018: 767–783.

典型三维视觉任务——三维重建

■ 经典方法——神经辐射场（Neural Radiance Fields，NeRF）重建



- **隐式表示3D场景：**NeRF不再使用传统的点云或体素，而是通过一个**多层感知机**学习场景的**隐式连续表示**，从而实现高质量的3D重建和新视角合成。
- **基于体渲染的神经辐射场建模：**NeRF使用**体渲染技术**，通过射线投射计算不同深度上的颜色和密度，从而逼真地模拟光线在3D场景中的传播，实现高质量的光照、遮挡和视角变化。
- **视角依赖建模：**NeRF通过引入**视角方向编码**，可以精确建模材质的反射特性，实现逼真的视角变化和光照效果，特别适用于合成真实感极强的光滑表面和半透明材质。

Ben Mildenhall et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Communications of the ACM, 65(1), 99–106, 2021.

山东大学
SHANDONG UNIVERSITY

《人工智能通识》AI For Everyone

智慧之眼——视觉感知

学无止境 气有洪range

教育部-华为“智能基座”课程

人脸识别技术