计算机视觉

从成像原理到多模态视觉-语言模型 • 博士进阶课程

MarkZZZ WeChat: MarkZZZ20XX

课程简介

计算机视觉（Computer Vision）是人工智能与信息科学的核心交叉领域，研究如何使计算机从图像与视频中理解视觉世界。本课程从图像形成的物理原理出发，经由经典几何视觉（针孔模型、对极几何、三维重建）和深度学习视觉（CNN、目标检测、图像分割、视觉 Transformer）进阶至 多模态与生成视觉（CLIP、Stable Diffusion、NeRF）的前沿研究方向。

课程强调数学严格性与工程实践的紧密结合：每一主题均给出完整的数学推导（如 8 点算法、NeRF 体渲染积分、InfoNCE 损失），同时讨论其在工业界与学术界的最新应用。全部 12 讲以博士水平呈现，帮助学生掌握计算机视觉的理论本质，理解从经典算法到大模型的技术演化脉络，并具备独立开展视觉方向研究的基础能力。

适合人群

- 已掌握线性代数、概率论与深度学习基础，希望深入学习计算机视觉理论与前沿的博士生/研究者
- 从事自动驾驶、医学图像、机器人感知等方向，需要扎实视觉理论基础的工程师与算法工程师
- 对视觉几何、深度学习视觉与多模态模型感兴趣的计算机/数学专业学生
- 希望了解 NeRF、Stable Diffusion、CLIP、LLaVA 等前沿系统核心机制的研究者

前置知识

- 线性代数：矩阵运算、特征值分解、奇异值分解（SVD）、张量积
- 概率论与统计：随机变量、贝叶斯推断、信息论基础（熵、KL 散度）
- 微积分与优化：梯度、链式法则、凸优化基础、随机梯度下降
- 深度学习基础：神经网络前向/反向传播、CNN 基本结构、注意力机制
- 编程能力：Python/PyTorch 或 TensorFlow 使用经验有助于理解算法实现

1 课程内容

讲次	主题	内容概要
1	图像形成	针孔相机模型（射影几何、焦距、主点）；透视投影变换（齐次坐标、3×4 投影矩阵）；图像传感器（CCD/CMOS、拜耳阵列、去马赛克）；图像的线性滤波（卷积、高斯滤波、频域分析、奈奎斯特采样定理）；色彩空间（RGB、HSV、Lab）
2	图像特征	边缘检测：Marr-Hildreth（LoG）与 Canny 算法（非最大值抑制、双阈值）；Harris 角点：二阶矩矩阵、响应函数 $R=\det(M)-k\,\mathrm{tr}(M)^2$、旋转不变性证明；SIFT 特征：尺度空间（高斯差分 DoG）、极值检测、梯度方向直方图、128 维描述子；特征匹配与比率测试（Lowe's ratio test）
3	相机几何	两视图几何：本质矩阵（Essential Matrix）与基础矩阵（Fundamental Matrix）的定义与关系；对极约束 $x'^{\top}Fx=0$ 的推导；8 点算法的完整矩阵推导；RANSAC：鲁棒估计框架、内点模型、迭代次数分析；相机标定（Zhang's 方法、径向畸变校正）
4	立体视觉与三维重建	双目立体视觉：三角测量（Triangulation）原理与线性解；视差图（Disparity Map）；稠密匹配：代价聚合、半全局匹配（SGM）；运动恢复结构（Structure from Motion, SfM）：增量式 SfM 流水线（特征匹配 $\rightarrow$ 两视图初始化 $\rightarrow$ 增量注册 $\rightarrow$ BA）；光束法平差（Bundle Adjustment）的非线性最小二乘形式
5	CNN 理论	卷积的表示能力：通用近似定理的卷积形式；感受野的递归分析；Batch Normalization：算法推导、训练/推断差异、梯度流分析；残差网络（ResNet）：恒等映射的梯度传播、跳连接的数学作用；深度可分离卷积（MobileNet）；神经架构搜索（NAS）概述
6	目标检测	两阶段检测器：R-CNN $\rightarrow$ Fast R-CNN $\rightarrow$ Faster R-CNN（RPN 锚框、分类 + 回归双头、端对端训练）；单阶段检测器：YOLO 系列（YOLOv1 到 YOLOv8 的演化）、SSD、RetinaNet（Focal Loss）；非极大值抑制（NMS）及其变体；检测评估指标（mAP、IoU、AR）

讲次	主题	内容概要
7	图像分割	语义分割：FCN（全卷积网络）、空洞卷积（膨胀率设计）、DeepLab 系列（ASPP）；实例分割：Mask R-CNN（RoIAlign、掩码头）、分割-检测联合训练；U-Net：跳连接的理论分析、编码器-解码器对称结构、医学图像应用；分割一切模型（SAM）：提示机制、无监督预训练策略
8	视觉 Transformer	Vision Transformer（ViT）：Patch Embedding、位置编码、CLS Token 分类；多头自注意力的计算复杂度（$O(n^2d)$）；Swin Transformer：分层结构、窗口注意力、Shifted Window 机制及其移位证明；DeiT（知识蒸馏训练）；ViT vs CNN 的归纳偏置比较
9	生成视觉模型	GAN 理论：DCGAN 架构、训练目标、模式崩溃问题；StyleGAN：映射网络、AdaIN 归一化、风格混合与截断技巧；扩散模型（Diffusion Models）：DDPM（去噪扩散概率模型）数学推导、DDIM 加速采样、条件生成（Classifier-free Guidance）；DALL-E 2 与 Stable Diffusion：文本-图像对齐机制、潜在扩散模型（LDM）
10	视频理解	光流估计：Horn-Schunck 方程（变分优化推导）、Lucas-Kanade 方法、RAFT；视频分类：双流网络（空间流 + 时间流）、3D 卷积（C3D/I3D）、SlowFast 网络（双帧率设计）；Video Transformer：TimeSformer（分离时空注意力）、Video Swin；动作识别与时序动作检测
11	多模态视觉-语言	CLIP（Contrastive Language-Image Pre-training）：InfoNCE 对比损失完整推导、大规模图文对预训练、零样本迁移；ALIGN（噪声数据对齐）；视觉问答（VQA）：早期融合方法到大模型方法；LLaVA：视觉指令微调、多模态大模型架构；视觉接地（Visual Grounding）
12	三维视觉前沿	神经辐射场（NeRF）：体渲染方程完整推导（体积积分、透射率、采样策略）、位置编码的必要性、Instant-NGP 加速；三维高斯散点（3D Gaussian Splatting, 3DGS）：椭球表示、光栅化、梯度推导；点云处理：PointNet（最大池化置换不变性证明）、PointNet++（局部特征学习）；前沿研究展望（动态 NeRF、可编辑 3D、具身视觉）

2 教学方法

- **理论讲义**（Lesson）：每讲 8–16 页，包含完整定理证明、引理推导、典型例子与 7–8 道分级习题（ 基础 / 进阶 / 研究级）
- **幻灯片**（Lecture）：每讲 20–25 帧，覆盖核心概念、算法流程图、可视化示例与研究方向
- **数学严格性**：所有核心结论均给出完整证明，包括：8 点算法秩约束证明、NeRF 体渲染积分推导、InfoNCE 损失信息论解释等

3 主要参考文献

3.1 经典教材

- R. Hartley & A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press, 2004. [相机几何、对极几何、三维重建的权威参考]
- R. Szeliski, *Computer Vision: Algorithms and Applications*, 2nd ed., Springer, 2022. [覆盖面最广的综合性教材，在线免费]
- I. Goodfellow, Y. Bengio & A. Courville, *Deep Learning*, MIT Press, 2016. [深度学习数学基础，在线免费]
- D. Forsyth & J. Ponce, *Computer Vision: A Modern Approach*, 2nd ed., Prentice Hall, 2011. [经典视觉教材]

3.2 核心论文

- K. He et al., "Deep Residual Learning for Image Recognition," *CVPR* 2016. [**ResNet**]
- S. Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *NeurIPS* 2015. [**Faster R-CNN**]
- K. He et al., "Mask R-CNN," *ICCV* 2017. [**Mask R-CNN**]
- A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR* 2021. [**ViT**]
- Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *ICCV* 2021. [**Swin Transformer**]
- A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *ICML* 2021. [**CLIP**]
- B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *ECCV* 2020. [**NeRF**]
- B. Kerbl et al., "3D Gaussian Splatting for Real-Time Novel View Synthesis," *SIGGRAPH* 2023. [**3DGS**]
- J. Ho et al., "Denoising Diffusion Probabilistic Models," *NeurIPS* 2020. [**DDPM**]
- H. Liu et al., "Visual Instruction Tuning," *NeurIPS* 2023. [**LLaVA**]

3.3 补充资源

- Stanford CS231n: Convolutional Neural Networks for Visual Recognition (http://cs231n.stanford.edu) —深度视觉学习最佳入门资源
- CMU 16-385 Computer Vision (http://16385.courses.cs.cmu.edu) —经典视觉几何
- ECCV / CVPR / ICCV / NeurIPS / ICLR 年度论文集—前沿进展