人工智能与计算思维·生物医学图像处理

裴玉茹

2025 年 10 月 22 日

目录

前言

本书面向生物医学领域的专业需求，针对生物医学迈入了人工智能赋能的新时代面临的机遇与挑战，探讨人工智能与生物医学领域的交叉应用研究。内容涵盖传统机器学习算法、计算机视觉核心概念以及深度学习先进技术，具体包括经典机器学习方法与现代深度学习模型，并重点讲解这些技术在生物医学影像处理中的应用。课程还将深入介绍解剖结构组织建模与评估方法，阐述如何从多模态生物医学影像中提取结构信息、识别异常病变，为疾病诊断、手术模拟规划与术中干预等环节提供关键技术支撑。通过系统学习，学生将探索生物医学影像处理领域从人工设计到数据驱动、从单模态适配到多模态通用的发展路径，特别关注多模态大模型与生物医学知识的深度融合，了解大模型如何重塑“数据依赖”的传统生物医学影像处理范式，提升模型泛化能力，并实现跨模态协同的生物医学影像理解。

裴玉茹

2025 年 *10* 月 *22* 日

第一章　引言

本讲义以人工智能与计算思维在生物医学图像处理中应用为主题。在传统机器学习方法领域，将深入讲解回归分析、分类算法及其正则化问题，同时涉及决策树、随机森林等经典模型，以及人工神经网络的基础理论与应用。在当前人工智能研究的主流领域深度学习方面，将系统阐述各类神经网络架构，包括当前广泛应用的生成模型体系，诸如自编码器、变分自编码器、生成对抗网络及扩散模型等。内容还延伸至计算机视觉领域，聚焦于生物医学图像处理这一交叉方向，包括视觉任务中的图像分割、目标检测、图像重建与图像配准等技术。将探讨大规模语言模型与生物医学图像处理的关联机制，特别是当代大型语言模型已具备处理多模态输入的能力，能够实现对图像内容的解析，为医学图像处理提供了新的技术支撑。

在应用案例方面，课程将介绍医学图像分割领域的主流模型，如可支持多模态、多器官部位的分割任务预训练大模型，同时涵盖高通量成像技术的分析方法，针对电子显微镜与光场显微镜获取的纳米级结构图像，课程将阐述如何通过图像处理算法，包括深度学习出现之前的传统方法，实现细胞及亚细胞结构形态的自动化提取。此外，课程还将介绍基于深度学习的预训练模型在显微图像细胞分割任务中的应用。

从学科发展视角出发，将梳理生物医学图像处理的演进脉络。自然图像成像技术的历史可追溯至1839 年，法国科学院公布路易·达盖尔的银版摄影术，标志着现代摄影术的诞生。而医学图像的起源则以1895 年威廉·伦琴发现X 射线开始，该发现推动了X线成像技术的发展，涵盖二维线片如牙片、胸片，以及螺旋CT成像与低剂量CT等。伦琴本人也因此获得首届诺贝尔物理学奖。

随着1946 年电子计算机的问世，人工智能、计算机视觉与数字图像处理等学科逐步形成。人工智能领域通常以1956 年的达特茅斯会议为正式诞生标志，但神经网络的雏形早在20 世纪40 年代已出现。计算机视觉的明确目标，即解析图像内容于1966 年MIT暑期学校中被正式提出。数字图像处理技术则在20 世纪60 年代通过美国NASA 的太空探索项目得到早期应用。

医学图像设备的临床普及始于20世纪80年代后，其中CT技术于1971年在英国医院首次实现临床成功应用，MRI 技术于1977 年完成首个人体图像采集，其发明者后来亦获得诺贝尔物理学奖。生物显微图像技术在近百年间取得快速发展，尤其80年代后电镜与光镜的分辨率实现纳米级突破，可捕捉细胞器形态及动态行为。深度学

习领域的发展以2006 年为重要节点，杰弗里·辛顿等人在Science期刊发表的自编码器论文推动了深度神经网络的蓬勃发展。如今，深度神经网络已成为生物医学图像处理及众多学科领域的主流方法，其发展时间跨度虽短，自20世纪90年代末至今，但已深刻改变了生物医学图像处理的技术生态。

1.1 人工智能的定义

人工智能被定义为计算机科学中专注于构建能够执行需人类智能完成任务的系统，涵盖推理、学习、问题解决、感知、语言理解与决策等能力。人工智能创造能执行需人类智能才能完成任务的机器或系统，这些任务包括推理、学习、问题解决、感知、语言理解和决策等。人类智能源于数十亿年前生命起源，历经漫长进化，从远古生物逐步演化而来。从进化视角看，人类智能的物质基础是经过400 多万年进化形成的大脑，其容量在进化过程中持续提升。在从猿到人的进化过程中，智力与工具使用能力不断提升，最终形成解决问题的智能。教育心理学家霍华德· 加德纳提出多元智能理论，涵盖视觉、语言、交往、自知、逻辑、身体、音乐和自然观察智能，其中视觉智能（空间信息感知与处理）、语言智能（符号使用能力）与逻辑智能（推理与分析能力）均与生物医学图像处理密切相关。

1.1.1 人类智能vs.人工智能

人工智能与人类智能的差异主要体现在以下方面：

学习过程与数据依赖

- AI 模型：以数据驱动为核心，监督学习需大规模人工标注数据优化参数。无监督学习、强化学习虽减少对标注的依赖，但仍需从环境反馈或无标签数据中提取模式，例如训练生物医学图像分割模型时，需大量标注的解剖结构数据，且数据质量直接决定模型性能。

- 人类智能：以经验学习为基础，融合社会互动与环境反馈。基础教育阶段存在“监督学习”，如教师明确知识对错，但脱离该阶段后，更多依赖少量样本实现“举一反三”，如医生通过少量病例掌握疾病特征，且具备抽象推理与直觉能力，无需依赖海量数据。

问题解决能力与场景适应性

- AI 模型：擅长计算密集型、重复性任务，可高速处理海量数据，但泛化能力局限于特定任务，例如，医学图像分割模型可处理多模态、多器官分割，但无法直接迁移至疾病诊断或手术规划任务。在信息不完整的动态环境中，模型性能易受影响。

- 人类智能：擅长通用问题解决，可通过批判性、创造性思维应对复杂场景。医生在影像信息不全时，可结合临床病史与经验做出诊断，具备跨任务泛化能力，如从影像分割自然迁移至疾病判断，且能在模糊、动态环境中快速适应。

速度效率与标注成本

- AI 模型：推理阶段效率极高，训练完成的分割模型可在毫秒级输出结果，但训练阶段需消耗大量计算资源与时间，例如三维医学图像分割模型的预测时间通常小于1 秒，而训练可能需数天至数周。

- 人类智能：数据标注效率低，尤其针对三维、动态医学图像。标注一个简单解剖结构需1-2 分钟，复杂结构，如亚细胞结构需数分钟甚至更久。若需为算法验证标注“真值”，往往需投入数周时间，且标注结果易受主观因素影响。

结果一致性与金标准差异

- AI 模型：输出结果具有高度一致性，只要输入数据与参数不变，结果可重复，但“准确性”依赖于训练数据的质量与标注真值的可靠性。

- 人类智能：结果一致性低，即使经验丰富的专家，对同一医学影像的标注，如病变边界界定也可能存在差异。临床“金标准”通常需通过多专家投票、统计平均等方式确定，而非单一专家的判断。

创造力、意识与伦理维度

- 创造力：AI 的创造力局限于训练数据范畴，如生成图像、文本均基于已有数据模式。人类可产生全新概念与创新成果，如提出新型医学影像分析方法。

- 伦理道德：AI 的伦理取向取决于开发者的编码目标与价值观。人类可基于社会规范、职业伦理，如医学希波克拉底誓言做出判断，尤其在生物医学领域，如患者隐私保护、诊断责任界定，人类的伦理决策不可替代。

- 意识与自我意识：人类具备反思、内省与道德判断能力。AI 无自主意识，仅模拟人类行为，无法理解自身决策的本质。

人工智能发展的物质基础是计算能力的提升，这与计算机技术的进步高度同步。摩尔定律描述了计算能力的增长规律，单位面积电子器件数量每18-24 个月翻倍，推动过去60 余年计算机技术的快速发展。以GPU 为例，4090 显卡采用4-5 纳米工艺，包含760 亿个晶体管，具备16000 个计算核心与24GB 显存，浮点运算能力达82.6 TFLOPS（每秒/万亿次浮点运算）。

1.1.2 人工智能研究领域

人工智能可分为**狭义人工智能**和**广义人工智能**。狭义人工智能专注特定任务，如语音识别、图像分类，能力局限于特定领域，依赖数据和算法，无自主意识与情感。如今生物医学图像处理领域的人工智能模型均属于此类。尽管如此，狭义人工智能已广泛应用于语音助手、推荐系统、图像处理、自动驾驶等领域，展现出强大能力。广义人工智能则是理论上具备人类般广泛智能的系统，能够跨领域自主学习、推理和解决问题，拥有与人类相似的认知和情感，但目前尚未实现，关于其实现路径当前仍存在诸多争议。

人工智能的研究领域广泛，涵盖**计算机视觉、自然语言处理、智能机器人、机器学习、深度学习和专家系统**等。在计算机视觉领域，深度学习成为主流研究方法，涉及多模态数据融合和自监督学习等技术，应用于虚拟现实、人脸识别、自动驾驶等场景。自然语言处理能力不断突破，大语言模型不仅能撰写论文，还能解析生物医学图像并生成报告，未来将在架构改进、强化学习融合和知识图谱应用等方面持续发展。智能机器人领域，AI 结合传感器和计算机视觉技术，提升机器人感知能力，借助自然语言处理理解人类指令，使机器人运动控制更灵活。

深度学习在**计算机视觉**领域的主导地位与关键技术。当前，卷积神经网络及各类深度神经网络模型已成为计算机视觉领域的主流研究方法。尽管计算机视觉学科的发展历程悠久，但深度学习凭借其强大的特征提取与建模能力，逐步占据该领域核心地位。多模态融合是重要发展方向之一, 不仅涵盖传统视觉处理的图像数据，还包括视频序列、深度图像、红外图像，以及各类传感器采集的非视觉数据，如温度、压力传感器数据等。这些多源模态数据通过协同计算机制，可实现信息互补，显著提升计算机视觉系统对复杂场景的感知与理解能力。

与此同时，自监督学习技术在深度神经网络应用中逐渐凸显重要性。在传统有监督学习模式下，深度神经网络的性能高度依赖人工标注数据，因此曾有“人工智能即人工标注智能”的调侃，即模型性能与标注数据量呈强相关。但当前的学习范式已突破这一局限，除自监督学习外，强化学习等技术也能在不依赖大规模人工标注数据的前提下实现模型训练，大幅缓解了对标注数据的依赖，同时降低了人工标注的工作负担，为数据稀缺场景中的模型应用提供了可能。

自然语言处理（NLP）是人工智能领域当前应用最广泛的方向之一，其技术能力已深度融入日常学习与工作场景。各类聊天机器人工具已成为学生完成作业、解决信息查询需求的常用辅助手段。当前自然语言处理技术的能力已达到较高水平，不仅能应对常规信息问答，还可支持复杂任务，如学术论文生成。

2024 年，萨卡马（Sakama）公司推出的AI 系统已实现端到端学术论文生成，被称为“AI 科学家”，可完成从研究主题头脑风暴、实验设计与模拟，到论文撰写与修订的全流程工作，且能针对顶刊顶会论文提供定制化生成服务，相关服务甚至采用明码标价模式。此外，自然语言处理技术还可与计算机视觉结合，应用于生物医学图像分析场景。大型语言模型（LLM）可基于生物显微图像自动生成分析报告，

模拟有经验医师的图像解析过程，辅助判断图像中可能存在的疾病相关特征，实现“图像- 文本”的跨模态信息转换。

从技术发展来看，自然语言处理领域仍在持续突破：一方面，深度神经网络架构不断优化，以Transformer 架构为代表的基础模型持续迭代，衍生出多种改进版本。另一方面，技术融合趋势显著，通过结合强化学习可提升模型的决策与优化能力，融入知识图谱则能增强模型对领域知识的理解与应用，进而提升问答系统的准确性与深度。以典型大语言模型deepseek 为例，其发展历程呈现清晰的性能跃升轨迹。2024 年1 月，其性能已超越Llama 系列模型。2024 年6 月，性能进一步超越GPT 系列部分模型。DeepSeek已推出多代大模型以及对话助手，为用户提供高效、智能且免费的AI服务，持续探索通用人工智能。

人工智能技术的发展也推动了**机器人领域**的革新。机器人并非局限于类人形态，而是涵盖各类具备自动化执行能力的机械系统，如工业机械臂、医疗手术机器人、类人移动机器人及机器狗等。在医疗领域，达芬奇手术机器人作为典型代表，已能完成高精度、精细化的外科手术操作；在移动机器人领域，类人机器人具备强大的运动控制能力，可实现自主避障、奔跑等复杂动作；机器狗则已应用于军事等特殊场景，执行巡逻、探测等任务。

人工智能对机器人的赋能主要体现在三个层面：

- 感知能力增强。通过多类型传感器，如视觉传感器、深度传感器、红外传感器与计算机视觉算法的结合，机器人可更精准地感知外部环境信息，提升环境适应性。

- 指令理解能力提升。依托自然语言处理技术的进步，尤其是大语言模型的发展，机器人能更准确地理解人类自然语言指令，实现更高效的人机交互。

- 运动控制优化。AI 算法可通过对运动数据的学习，优化机器人的运动轨迹规划与执行精度，使机器人运动更灵活、稳定。

需注意的是，机器人性能的提升与感知技术、计算能力的发展呈协同演进关系，二者共同推动机器人从单一功能设备向智能化系统转变。

1.2 人工智能发展历程

人工智能的发展历史虽相对较短：1956 年达特茅斯会议确立其目标；1965 年出现早期聊天程序；80 年代各领域专家系统涌现；1997 年“深蓝”战胜国际象棋冠军；2002 年出现具备避障功能的扫地机器人；2009 年自动驾驶技术问世；2011 年人工智能在智力竞赛中击败人类；2011 - 2014 年语音助手普及；2014 年GAN 模型诞生；2016 年AI 在围棋领域战胜世界冠军；2018 年AI 成为大学独立课程。

早期概念与技术雏形（20世纪40年代前）

人工智能的思想源头可追溯至更早的计算工具发明：17 世纪帕斯卡（Pascal）发明的机械计算器，以及19 世纪巴贝奇（Babbage）与阿达（Ada Lovelace）提出的可编程计算思想，均为人工智能的诞生提供了技术灵感；20世纪40年代电子计算机问世后，神经网络的雏形开始出现，为后续人工智能的算法研究提供了基础框架。

智能判定标准与学科命名（20 世纪50 年代）

1950 年，图灵（Turing）提出“图灵测试”，为智能的判定提供了可操作的标准：若一台机器能在对话交互中令人信服地模仿人类行为，即可认为其具备智能。当前主流聊天机器人在多个任务场景中已能通过图灵测试，智能判定已不再是稀缺目标。1956 年的达特茅斯会议具有里程碑意义，此次会议首次正式提出**“人工智能（AI）”**这一术语，标志着人工智能作为一门独立学科的诞生，为后续领域发展明确了方向。

早期技术探索（1956 年至今）

自1956 年达特茅斯会议首次提出“人工智能”概念以来，其核心目标始终保持一致，开发可模拟人类智能任意方面的机器系统。此次会议汇聚了一批具有前瞻性的科学家，包括约翰· 麦卡锡、马文·明斯基等，他们提出的概念不仅奠定了AI 学科的基础，更深刻重塑了后续科技发展格局。会议参与者及相关领域研究者中，多人后续获得图灵奖、诺贝尔奖等顶级学术荣誉，印证了该会议对科技领域的深远影响。

AI 早期研究聚焦于符号主义人工智能，代表性成果包括赫伯特·西蒙（Herbert Simon）开发的“逻辑理论家”（Logic Theorist），首个可自动证明数学定理的程序，以及一系列面向特定领域的专家系统。值得注意的是，机器人技术的探索也同步起步。具备自主避障功能的机器人并非现代产物，早在AI 学科诞生初期便已出现，体现了早期研究者对“智能执行物理任务”的探索。

AI 寒冬的成因与同期技术进展

20 世纪60-70 年代，AI 领域进入首次“寒冬”，诱因是马文·明斯基（Marvin Minsky）在《感知机》（Perceptrons）一书中指出：单层神经网络无法解决非线性问题，且当时的技术难以突破这一局限。这一论断导致AI 相关研究的资金支持与学术关注大幅减少。

尽管处于寒冬期，专家系统仍取得显著进展，并在多个领域落地应用：医疗领域，MYCIN 系统可诊断细菌感染，尤其是血液感染并推荐治疗方案；矿产勘探领域，专家系统能分析地质数据以识别潜在矿脉；内科医学领域，决策支持类专家系统可辅助医生制定治疗方案。此外，IBM 开发的问答机器人沃森（Watson）在21 世

纪初展现出超越人类的知识竞赛能力，为后续自然语言处理技术的发展提供了重要参考。

技术范式转型与深度学习革命（20世纪90年代至今）

传统机器学习的崛起（20世纪90年代）20 世纪90 年代起，AI 技术范式从传统符号方法转向统计机器学习，核心逻辑是通过数据驱动学习数据内在模式，而非依赖人工定义规则。1997 年，IBM “深蓝”（Deep Blue）击败国际象棋世界冠军加里·卡斯帕罗夫，标志着AI 在复杂决策任务中的突破。反向传播（Backpropagation）算法的成熟，为深度神经网络的参数优化提供了核心工具，成为后续深度学习模型训练的基础。

深度学习的爆发（21 世纪10 年代至今）

21 世纪10 年代，深度神经网络进入高速发展期，涌现出一系列里程碑式工作：

生成模型体系逐步完善，变分自编码器（VAE）、生成对抗网络（GAN）、扩散模型（Diffusion Model）等成为解决图像生成、修复等任务的核心工具。

Transformer 架构的提出催生了视觉Transformer（ViT）等模型，推动计算机视觉领域从卷积神经网络（CNN）向“注意力机制+ 深度学习”融合范式转型。

大规模数据集的支撑作用凸显，ImageNet 作为首个大规模图像数据库，为深度学习模型的性能验证提供了关键支撑。2012 年，AlexNet基于CNN在ImageNet 图像分类竞赛中取得突破性成绩，其错误率远低于传统方法，彻底改变了计算机视觉领域的技术格局，尽管CNN 的理论框架早有提出，但大规模数据的出现才真正释放了其性能潜力。

2024 年成为AI 领域的重要里程碑，两项AI 相关成果获得诺贝尔奖：约翰·霍普菲尔德（John Hopfield）与杰弗里·辛顿（Geoffrey Hinton）因“通过人工神经网络奠定机器学习基础的发明与发现”获诺贝尔物理学奖。

阿尔法折叠（AlphaFold）团队的华盛顿大学大卫·贝克教授、DeepMind 的德米斯· 哈萨比斯与约翰·江珀因“开发AI 驱动的蛋白质结构预测技术”获诺贝尔化学奖。这标志着AI 技术已从计算机科学领域延伸至基础科学研究，并获得学术界的最高认可。

1.2.1 人工智能发展面临的核心挑战

尽管深度学习/深度神经网络成为人工智能领域的核心技术，但该领域仍面临诸多待解决的挑战，主要集中在数据、模型架构、计算能耗三个方面：

（一）数据依赖与数据质量问题

当前深度神经网络的性能高度依赖数据支撑：在面对特定任务时，首先需解决数据获取问题；若需实现零样本（Zero-Shot）学习，则需依赖大规模预训练模型。

在生物医学图像处理等特定领域，数据相关问题更为突出，不仅需考虑数据量的充足性，还需关注数据标签的可用性，包括标签的有无、标签的置信度等，这些因素直接决定模型能否有效完成任务。

（二）模型架构的持续迭代需求

过去十余年间，深度神经网络领域涌现出多个里程碑式模型，这些模型的成功均依赖于架构层面的创新。当前，模型架构的演进仍未停止。以大型语言模型的底层架构Transformer 为例，其设计仍在持续优化，衍生出适应不同任务的变体架构。同时，针对特定领域的模型架构定制化需求也日益凸显，需通过架构改进提升模型对领域数据的适配性。

（三）计算资源与能耗问题

深度神经网络的训练与部署需消耗大量计算资源与能源。训练阶段通常需配备高性能计算设备，如高端GPU，且运行过程中能源消耗显著。据统计，GPT 系列模型每日耗电量超过50 万千瓦时，相当于1.7 万个美国家庭的日均耗电量。微软公司为支撑其AI 业务，宣布与能源企业合作重启核电站，以满足巨大的能源需求。从能耗效率对比来看，人类智能与人工智能存在显著差异：人类处理单张图像的能耗约为0.00002 kWh/次，而GPT 系列模型处理单张图像的能耗为0.1-1 千瓦时。在基础能耗方面，人类大脑的基础功耗约为20 瓦，而单块GPU 的功耗通常在250-400 瓦。此外，二者的环境影响也存在差距。人类智能的运行过程碳排放极低，而人工智能系统因依赖化石能源发电，碳排放水平较高。需注意的是，尽管人类智能的能耗效率优势显著，但人类智能的习得过程，从基础教育到高等教育需消耗大量资源。一旦完成知识积累，其后续应用的能耗成本则极低，这一特性与人工智能“前期训练高能耗、后期推理相对低能耗”的模式形成鲜明对比。

1.3 生物医学图像处理概述

1.3.1 医学图像vs. 自然图像

医学图像的起源可追溯至1895 年威廉·伦琴发现X 射线，此后逐步发展出CT、MRI、PET、超声等多模态技术，如今，生物医学图像处理涉及多种模态和部位的图像，包括二维X线片、超声、CT、MRI、PET 等。尽管医学图像与自然图像的成像原理不同，医学图像多为物理量的空间映射，而自然图像源于光线对感光介质的作用，相关领域已有数十位科学家获诺贝尔奖，体现了其在基础科学与临床中的重要地位。

与自然图像的共性为从数字图像处理的角度来看，二者都可采用张量表示，一维、二维、三维甚至更高维的图像都能以相应形式的矩阵或向量进行表达。二维图像为2D 张量（HxWxC，H 为高度、W 为宽度、C 为通道数），三维图像为3D 张量（DxHxWxC，D 为深度），动态图像可扩展为4D 张量（TxDxHxWxC，T 为时间维

度），这为跨模态图像处理提供了统一基础。

其差异体现在自然图像依赖物体反射/发射光线成像，侧重外观特征。医学图像多为物理量映射，如X 射线的组织密度、MRI 的质子共振信号，侧重解剖与功能特征。X 射线可穿透人体显示内部结构，而自然图像无法直接呈现物体内部信息。虽然自然图像和医学图像在数据表示上相似，但处理方法存在差异。医学图像处理中的微分同胚概念用于计算图像间双向可逆的点对应关系，这一方法也适用于自然图像三维网格的对应计算。X射线在医学领域广泛应用，同时也被引入自然图像三维重建，以解决物体内部结构不可见的问题。

1.3.2 生物医学图像处理任务

图像配准与对齐

图像配准与对齐寻找图像间具有语义相关性的对应点，计算形变场以实现多模态、多时间点图像的空间对齐。在临床上可用于分析治疗前后解剖结构的形态变化，如肿瘤大小变化、手术中器械与器官的位置匹配。通过二维X 射线定位三维器械位置。生物医学图像配准与自然图像处理中的“光流计算”、“场景流分析”任务一致，均需通过对应关系捕捉动态变化，但医学图像需满足“微分同胚”、双向可逆、无结构坍塌以确保解剖结构完整性。

三维重建

从二维医学图像，例如X 线片构造三维CT图像，从稀疏数据恢复完整三维结构。可减少辐射风险，用低剂量稀疏数据重建三维CT图像、避免治疗中重复数据采集，在手术中通过二维图像实时重建三维器官形态。融合AI 模型，如自编码器、扩散模型、隐式神经网络与传统数值优化方法。基于扩散模型的技术可从单张二维图像重建三维医学结构。

分割

生物医学图像处理是结合计算机技术与医学影像的重要交叉学科，旨在通过对医学图像进行增强、分割、配准及特征提取，辅助疾病诊断与研究。该技术广泛应用于CT、MRI、显微镜图像及超声等影像分析中，可自动识别病变区域、量化生物标记物，并支持手术导航与治疗方案规划。随着人工智能与深度学习的发展，图像处理的精度与效率显著提升，成为推动精准医疗和转化医学发展的关键工具，为临床决策提供客观、可靠的依据。

1.3.3 AI 在临床全流程中的角色

AI 已渗透至生物医学图像处理的临床全流程，而非局限于影像科诊断：

- 前端，数据采集：通过AI 优化数据采集策略，如低剂量CT、少视角MRI 的重建，在降低辐射/ 扫描时间的同时保证图像质量。

- 中端，影像分析：辅助影像科医生完成分割、检测、量化分析，如病变体积计算，提升诊断效率与准确性。

- 后端，临床决策：支持手术规划、预后预测，但最终诊断仍需人类医生确认，当前AI 医院、AI 医生尚未能完全替代人类，而是作为“增强工具”提升医疗服务质量。

1.3.4 研究范式与技术融合

生物医学图像处理领域当前存在四种主流研究范式，且均处于活跃状态：

- 传统数值优化：无需训练，通过在线求解优化问题实现任务，如基于光流的图像配准，适用于数据稀缺场景；

- 小规模机器学习模型：基于少量数据训练端到端模型，在线推理时无需优化计算，平衡效率与数据需求；

- 大规模预训练模型（基础模型）：利用海量跨领域数据预训练特征提取器，在具体任务中通过微调或零样本学习实现应用，如基于CLIP、MedCLIP 的医学图像问答；

- 隐式计算与传统方法融合：结合隐式神经网络与数值优化，降低模型复杂度，实现小数据高效学习。

这些范式并非互斥，而是根据任务需求，考虑数据量、精度要求、计算资源灵活选择，体现了领域内“技术融合”的发展趋势，例如医学图像配准可同时采用Transformer 特征提取与微分同胚约束，兼顾精度与解剖合理性。

在算法层面，过去十余年深度神经网络的发展为生物医学图像处理带来了众多强大模型，如CNN、Transformer、各类生成模型和图神经网络等，以及近年来备受关注的大模型。这些模型在图像配准和三维重建等任务中发挥着重要作用。在图像配准任务中，早期可通过传统数值计算方法在线优化求解形变场。机器学习方法出现后，利用小规模数据训练端到端模型直接预测形变场。深度学习时代，使用大规模数据训练复杂模型，如3D U-Net。如今基于大模型，可利用预训练特征提取结合其他方法计算形变场，同时也存在结合传统数值计算的轻量化模型，多种方法在当前研究中并存。在三维重建任务中，自编码器、隐式神经网络、扩散模型等都可用于从二维图像恢复三维模型，新的三维生成模型也不断涌现。

1.4 AI生物医学图像处理小结

深度神经网络已广泛应用于生物医学图像处理的各个方面，涵盖图像生成、分割、诊断、治疗、三维重建和图像问答等任务。大语言模型能够基于医学图像回答手术阶段、器械识别等问题。在显微图像处理领域，人工智能同样可实现分割检测、图像质量增强和诊断等功能。基础模型凭借海量数据训练的优势，在各类任务中展现出强大性能，“零训练”（train-free）模型也日益常见。AI如同古代炼丹术士精心炼制丹药一般，通过对数据和算法的“混合提炼”，优化模型参数，期望获得能够有效解决生物医学图像处理问题的“神奇丹药”。

当前生物医学图像处理的核心特征是“AI 深度赋能”，但领域仍存在明确的研究价值。在任务创新方面，生物医学与临床医学的发展持续催生新任务，例如单细胞影像分析、动态功能成像量化，需针对性开发AI 模型。在技术方法方面，考虑弱监督、自监督学习解决数据稀缺问题。跨中心、跨模态迁移泛化能力不足、AI 诊断的因果推理中可解释性差等问题仍是研究重点。在人机协同方面，AI 与人类医生的“协同模式”而非“替代关系”成为共识，例如AI 辅助标注、人类确认诊断的workflow 已在临床逐步落地。

基础模型的轻量化适配，面向移动设备的医学AI 模型、多模态融合，如影像-文本- 基因组数据的联合分析，伦理规范构建，如患者隐私保护、AI 决策可追溯性等，这些方向将推动生物医学图像处理向更高效可靠的目标发展。

第二章　机器学习基础I

机器学习是人工智能的核心，而深度学习/深度神经网络是机器学习的重要内容。机器学习的目标是开发算法并构建模型，使计算机能够从数据中自主学习规律，进而实现基于数据的预测与决策。在基于机器学习生物医学图像处理领域，通过数据驱动的方式，自动化完成图像分割、病灶检测、病理诊断等传统依赖人工的任务，提升效率与准确性。

机器学习的核心要素

机器学习的本质是从数据中学习模型，进而利用模型解决预测、决策等实际任务。机器学习的核心要素包括数据、模型、算法与预测。本课程聚焦于模型与算法两部分。

- 数据：生物医学领域的核心数据为医学影像，如CT、MRI、电镜图像，数据准备阶段需完成采集，进行诸如临床病例影像收集，并基于专家知识进行标注，例如标注解剖结构掩膜，进行图像预处理操作。数据一般具有明确的任务与领域指向性，需针对特定任务在特定领域内进行准备。

- 模型：用于刻画数据规律的数学/计算框架，如传统机器学习中的决策树、支持向量机，深度学习中的卷积神经网络、Transformer。

- 算法：实现模型训练与优化的方法，如监督学习中的梯度下降、无监督学习中的K 均值聚类。

- 预测：模型训练完成后对新数据的推断过程，如输入新CT图像输出肿瘤分割结果，需通过“预测结果与真值的一致性”评价模型性能。预测则涉及模型在实际任务中的应用。

值得强调的是，在机器学习中数据的数量与质量直接影响模型性能。通常而言，数据量越充足、质量越高，越有可能训练出性能更优的模型。这也正是当前研究中高度关注数据问题的原因。在不同应用场景下，如何获取覆盖范围更广、变化特征更全面的数据，以及如何得到更可靠、准确的标注信息，是亟待解决的重要课题。在生物医学图像处理中，数据的“量”与“质”直接影响模型性能。通常数据规模

越大、标注越精准，模型泛化能力越强，这也印证了“人工智能的智能水平依赖人工提供的数据量”的行业共识。

2.1 机器学习类型

机器学习按学习方式分类，可划分为监督学习、无监督学习与强化学习三类。监督学习与无监督学习的核心区别在于训练数据是否包含目标域标签。以图像分割任务为例，若训练数据中的图像附带标签，如CT 数据中对咀嚼肌的标注，则该学习过程属于监督学习。以Mask R-CNN 模型为例，其作为多头图像分割模型，在生物医学图像处理中应用广泛。值得思考的是，当前深度学习模型层出不穷，研究者仍在持续开发新模型的原因何在？以CT 图像分割为例，CT 图像中常因口腔金属种植体产生射线硬化伪影，导致图像局部区域出现异常条纹，软组织成像信息完全丢失。若不对模型进行特殊处理或引入额外信息，分割模型在这些区域的表现会显著下降。针对此问题相关研究引入知识先验，如人体组织的附着对称关系，肌肉组织的对称性尤为明显，有效克服了伪影导致的图像退化问题，这也体现了模型优化与改进的现实需求。

无监督学习无需依赖数据标签，聚类算法是其典型应用，通过像素表观的相似性对近邻像素聚类，可实现亚细胞结构等目标的分割。同类亚细胞结构通常具有相同或相似的表观特征，该类算法在生物图像与自然图像的处理中均较为常见。需注意的是，不同模态的图像存在差异，即便针对相同的分割任务，这种差异也可能对处理过程产生影响。以细胞图像中亚细胞结构分割为例，有标签时的学习为监督学习，无标签时则为无监督学习。此外，不同模态图像的差异，如细胞图像中感兴趣结构尺寸小、重复出现的特点，与自然图像中物体的差异，对图像处理算法与分割模型会提出特殊要求。当前常用的预训练基础模型，如基于Transformer 的视觉模型ViT 需对图像进行Patch 划分，若目标结构尺寸极小，则对Patch 划分的精度要求会显著提高，这与自然图像的处理需求存在明显差异。

2.1.1 有/无/半/自监督学习

有监督学习

有监督学习是基于“输入-标签”配对数据训练模型的学习范式，要求每个输入数据均对应明确的真值标签，如医学图像对应解剖结构分割掩膜。模型通过学习“输入到标签的映射关系”，实现对新数据的预测。其流程为：定义损失函数衡量预测与真值的差异。通过优化算法，如梯度下降最小化损失函数，求解模型最优参数。

典型模型与生物医学应用

医学图像分割 考虑从腹部CT/MRI图像中自动分割肝脏、肾脏、脾脏等多类器官，输入为三维医学影像或二维切片序列，标签为对应器官的二值掩膜，其中属于目标器官的像素记为1，否则记为0；多器官分割需多通道掩膜，每通道对应一类器官。基于编码器-解码器（Encoder-Decoder）结构的CNN，例如U-Net，通过跳跃连接保留低层次空间信息，提升分割精度；

常用Dice相似系数（Dice Similarity Coefficient, DSC）、交叉熵（Cross-Entropy）或二者结合，其中DSC 更适用于医学分割，可以缓解类别不平衡问题，如小器官分割；通过梯度下降算法，如Adam、SGD迭代优化CNN的卷积核、偏置等参数，使模型输出的分割掩膜与人工标注真值尽可能一致。

线性回归

线性回归构建输入特征与输出变量的线性映射关系：

$$y = Ax + b$$

其中A为权重矩阵，b为偏置，通过最小二乘法最小化预测值与真实值的平方误差：

$$\min \sum_{i=1}^{n} (y_i - (Ax_i + b))^2$$

基于肿瘤体积的历史测量数据，其输入特征x预测未来生长趋势输出y，或通过患者年龄、血压等指标等输入预测器官体积的输出。

对于参数求解，对损失函数求偏导并令其为0，转化为线性方程组求解最优A与b，确保解的唯一性。最小二乘问题为凸优化问题，存在全局最优解。

逻辑回归

逻辑回归通过Sigmoid激活函数，$\sigma(z) = \frac{1}{1+e^{-z}}$将线性输出映射到[0,1] 区间，实现二分类任务，如“病灶存在/不存在”判断。

考虑输入为肺部CT图像的纹理特征，如灰度共生矩阵，输出为“肺结节良性/恶性”的概率；对于参数优化，采用对数损失函数：

$$\min -\sum_{i=1}^{n} [y_i \ln \sigma(z_i) + (1 - y_i) \ln(1 - \sigma(z_i))]$$

通过梯度下降求解权重参数，Sigmoid函数的非线性特性使其可拟合复杂分类边界。

决策树与随机森林

决策树为树状结构模型，通过根节点，即全数据集、分支节点包含特征判断条件，如“像素灰度> 128？”、叶节点包含分类/回归结果，实现数据划分，分裂标准为“最大信息增益”或“最小熵”。

随机森林为集成学习模型，通过“随机选择数据子集”与“随机选择特征子集”构建多棵独立决策树，最终通过投票（分类）或均值（回归）输出结果，提升模型鲁棒性。

考虑基于病理图像的细胞类型分类，其中决策树通过细胞核大小、染色强度等特征划分细胞，或通过多模态影像特征，CT密度、MRI信号，预测肿瘤分级，随机森林降低单棵决策树的过拟合风险。

支持向量机

支持向量机在特征空间中寻找最优超平面，使两类样本的间隔最大化，“最大间隔超平面”，通过核函数，如RBF 核处理非线性数据。

考虑脑MRI图像的组织分割，区分灰质、白质、脑室，输入为像素的纹理与强度特征，输出为组织类别标签，SVM 的强泛化能力适用于小样本医学数据场景。

无监督学习

无监督学习无需人工标注标签，通过挖掘数据自身的结构、模式与分布实现学习，可显著缓解生物医学领域“标注成本高、数据稀缺”的问题，如三维医学图像标注需专家耗时数小时/例，其目标从无标签数据中提取有效特征、划分数据簇或学习数据生成规律。

K均值聚类

K均值聚类在医学图像分割中的有效性高度依赖聚类中心的优化，其核心是通过“迭代优化聚类中心- 样本分配”实现“同类像素特征相似、异类像素特征差异显著”的分割目标。将数据划分为K个簇，使簇内样本方差最小化。迭代步骤为“初始化簇中心，分配样本，更新簇中心”。

流程如下：初始化阶段：随机选择K个像素或超像素的特征作为初始聚类中心，K需根据分割目标设定，如电镜图像分割中K=4对应细胞核、线粒体、微粒体、背景；

样本分配规则：采用“最近邻准则”，计算每个像素或超像素的特征与K个聚类中心的距离，如欧氏距离、余弦距离，将其分配至距离最近的聚类中心所属类别；

聚类中心更新：计算每个聚类内所有样本的特征均值，将其作为新的聚类中心；

迭代终止条件：当聚类中心的变化量小于预设阈值，或迭代次数达到上限时停止，此时聚类中心即为“最优聚类中心”，满足“类内方差最小化”目标函数：

$$\min \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2$$

μ_k为第k个簇的中心。

生物医学图像分割案例（电镜细胞结构分割）

特征提取：将高维电镜图像像素特征通过谱嵌入降维，使同类结构如细胞核的特征在低维空间中聚集；

聚类分割：设定K=4，通过K-Means将特征相似的像素划分为4类，生成与真实细胞结构一致的分割掩膜，无需人工标注，仅依赖数据自身特征分布即可实现亚细胞结构的自动分割；

可用于解决“医学图像标注成本高”问题，尤其适用于电镜、光镜等难以批量标注的显微图像分割场景。

k近邻（K-NN）

K-NN是典型的“惰性学习”模型，无显式训练过程，仅在预测阶段计算距离，其分类逻辑基于“近邻相似性”：

近邻检索：对于待分类样本如未知类别的病理图像块，在训练集中检索K个距离最近的样本；

类别投票：统计K个近邻的类别分布，将待分类样本归为“占比最高的类别”。考虑K=5 时，3 个近邻为“癌细胞”、2 个为“正常细胞”，则待分类样本归为“癌细胞”。

常用距离度量包括：**欧氏距离**：适用于低维特征，如病理图像的灰度均值。**余弦距离**：适用于高维特征，如基于CNN提取的医学图像特征向量。

应用案例，病理图像细胞分类： 输入乳腺病理切片的细胞图像块，特征为“细胞核面积、染色强度、圆形度”。通过K-NN（K=7）将细胞分为“良性上皮细胞”“恶性癌细胞”“炎症细胞”，准确率可达90% 以上，在小样本场景下优于传统参数模型。

医学图像检索： 输入待检索的肺部CT图像，含疑似结节。通过K-NN在CT数据库中查找K=10个最相似的CT图像，辅助医生参考历史诊断结果，提升诊断一致性。

自编码器

自编码器具有对称的编码器-解码器结构，编码器将输入数据压缩为低维特征，解码器从低维特征重构输入，通过“重构误差”优化模型，实现无监督特征学习；自编码器是对称的“编码器-解码器”结构，核心目标是“从数据中自主学习有效特征，实现输入数据的重构”，属于无监督学习模型。

编码器（Encoder）：通过卷积/全连接层将高维输入压缩为低维隐向量（Latent Vector），捕捉数据的核心特征。解码器（Decoder）：通过反卷积/全连接层将低维隐向量重构为与输入维度一致的输出。

优化目标为最小化“重构误差”，如MSE、L1损失，使输出与输入尽可能一致，重构误差越小，说明隐向量捕捉的特征越完整。

去噪自编码器

去噪自编码器（Denoising Autoencoder, DAE）输入为含噪MRI图像，编码器提取鲁棒特征并过滤噪声信息；如MRI运动伪影、CT噪声，通过编码器过滤噪声信息，解码器输出无噪图像；

考虑MRI图像去噪，输入含运动伪影的脑部MRI图像，DAE通过学习“噪声与信号的差异”，重构出无伪影的MRI 图像，无需“含噪-无噪”配对标签。

掩码自编码器

掩码自编码器（Masked Autoencoder, MAE）随机掩码输入图像的部分区域，如50%像素，模型需预测掩码区域的像素值，强制学习全局图像结构；

考虑医学图像预训练，对胸部CT图像进行随机掩码，MAE预训练后可作为特征提取器，迁移至“肺结节检测”任务，在小样本场景，如少量标注数据下也可实现很好的性能提升。

变分自编码器

变分自编码器（VAE）可控生成与医学图像合成，在隐向量中引入概率分布约束，如高斯分布，使隐向量具有可解释性，支持可控图像生成；

考虑肿瘤图像合成，通过VAE学习正常/病变肝脏CT图像的隐空间分布，调整隐向量可生成不同阶段的肿瘤图像，辅助医生理解肿瘤生长规律。

自监督学习

自监督学习是无监督学习的延伸，通过数据自身的内在关联生成“伪标签”作为监督信号，兼具“无监督，即无需人工标注”与“监督“即有明确学习目标的优势。其核心是：设计“预训练任务”，如图像裁剪拼接、像素预测，让模型从数据中自动生成标签，实现特征学习。例如对胸部CT图像进行随机裁剪，将裁剪块打乱后让模型预测原始图像的裁剪块位置，其中伪标签为“裁剪块坐标”；预训练后的模型可作为特征提取器，迁移到“肺结节检测”等下游任务中，在小样本场景下，自监督预训练模型的性能显著优于随机初始化模型，降低对标注数据的依赖。

自监督学习的核心是无需人工标注的目标域标签，仅依赖输入数据自身构建监督信号。除图像配准外，隐式神经网络NeRF模型中，多层感知机（MLP）用于表示三维场，如形变场、辐照场，在CT 重建中可表示X 射线衰减场，学习过程中不依赖场的真值，仅通过输入图像提供监督信号。

对比学习是另一类无监督学习方法，与多年前左右的测度学习相关，其核心是学习图像的良好表征。深度神经网络提取的图像特征需能支撑分割、分类、检测等下游任务。对比学习的核心是：相似数据对的特征应具有相似性，不相似数据对（如负样本）的特征应具有差异性，这一目标通过特定损失函数实现。常见的对比学

习损失函数包括InfoNCE 与动量对比损失（MoCo）。InfoNCE 损失函数的分子部分约束正样本对的特征相似性，分母部分约束负样本的特征差异性。MoCo 损失函数形式与之类似，核心差异在于维护一个可更新的负样本队列，无需每次计算时重新选取负样本，提升了计算效率。

自监督学习通过“数据自身生成伪标签”，无需人工标注，与无监督学习的差异在于“有明确的预训练任务目标”：

- 图像补全：随机掩码医学图像的局部区域，模型预测掩码区域像素，其中伪标签为原始像素；

- 对比学习：对同一医学图像进行不同增强，如旋转、翻转，模型学习“同一图像的增强版本为正样本，不同图像为负样本”的特征区分。

生物医学应用：神经辐射场（NeRF）的自监督训练

基于神经辐射场（NeRF）的三维重建任务目标通过二维RGB图像重建三维场景，如器官、细胞的三维结构。

自监督框架包括：

模型：用MLP表示“神经辐射场”描述三维场景的辐照度分布；

监督信号：从MLP生成的二维合成图像与原始输入图像的“重建误差”，无需真实三维结构标签；

基于单张二维X线片重建三维骨骼结构，减少患者辐射暴露，无需采集三维CT。

半监督学习

半监督学习利用“少量有标签数据+大量无标签数据”训练模型，适用于“医学数据标注成本高、标签稀缺”的场景，如罕见病影像、三维医学图像。通过无标签数据挖掘数据分布，辅助有标签数据优化模型。半监督与自监督学习是数据稀缺场景的解决方案。

半监督学习适用于部分数据有标签、多数数据无标签的场景。相较于仅使用少量有标签数据进行监督学习，引入大量无标签数据的半监督学习通常能提升模型性能。图像配准/对齐任务可作为理解无监督与半监督学习的典型案例。图像配准处理存在形变差异的输入图像，考虑对同一对象的CT 与MRI 多模态数据进行配准，可实现数据融合；对治疗过程中多次采集的数据进行配准，可分析组织结构的形变，为治疗效果评估提供依据。在深度学习框架下，图像配准常采用无监督方式实现，典型模型为对称编解码器，编码器提取输入图像对的特征，解码器预测二者间的形变场。训练时，将形变场施加于浮动图像，通过约束形变后的浮动图像与固定图像的一致性实现模型优化。由于未依赖真实形变场的标签信息，此类模型属于无监督

模型。因利用数据自身，形变后图像与目标图像的一致性作为监督信号，有时也被称为自监督模型，在于学习过程中无需目标域标签。

基于图像配准建立图像间的稠密对应关系后，可实现像素属性的跨图像迁移，这一特性可应用于基于标签迁移的图像分割，此类任务属于半监督学习。数据集中少量图像/图集有标签，通过建立所有图像间的稠密对应，将标签从图集迁移至无标签图像。从分割任务视角看，该过程因借助少量标签数据，属于半监督过程，有时也被称为“少样本分割”。

教师-学生模型

教师-学生模型（Teacher-Student Model）是一种知识蒸馏框架，常用于机器学习领域。该模型通过将一个庞大而复杂的“教师模型”所学到的知识，转移至一个轻量级的“学生模型”中。教师模型通常具有强大性能但计算成本高，它通过提供软标签或隐藏层特征，作为额外的监督信号来指导学生模型的训练过程。学生模型则通过模仿教师模型的行为，在显著减少参数量和计算开销的同时，尽可能保持较高的预测性能。该方法广泛应用于模型压缩、加速推理以及隐私保护等场景，是实现高效模型部署的重要技术之一。

- 教师模型：用少量有标签数据训练基础模型，生成无标签数据的“伪标签”，如肿瘤区域的预测掩膜；
- 学生模型：用“有标签数据+无标签数据+伪标签”训练，同时最小化“监督损失（利用有标签数据）”与“一致性损失（伪标签与学生模型预测的差异）”。

应用案例：脑部MRI肿瘤分割

数据：仅适用少量标签数据，例如10例有标签MRI，其含肿瘤掩膜+ 100例无标签MRI；

性能：半监督模型的精度显著优于仅用10例有标签数据的全监督模型，具有更好的泛化能力。

弱监督学习

弱监督学习使用“弱标签”，如图像级标签“含肿瘤/无肿瘤”，而非像素级分割掩膜，用于训练模型；

应用案例：肺结节检测：

输入：仅含“有结节/无结节”图像级标签的胸部CT。

模型：通过弱监督学习定位结节候选区域，再通过精细化模型实现结节边界检测，无需像素级标注。

2.1.2 强化学习：交互学习范式与生物医学机器人应用

强化学习（Reinforcement Learning, RL）通过“智能体（Agent）与环境（Environment）的交互”学习最优策略，其要素包括：

- 状态：环境的当前状态，如手术机器人的位置、医学图像中的器械姿态；
- 动作：智能体的决策，如机器人的机械臂运动、图像中病灶区域的标注；
- 奖励：环境对动作的反馈，如“成功抓取器械”给予正奖励，“碰撞器官”给予负奖励；
- 目标：学习使“累积奖励最大化”的策略，实现长期最优决策。

生物医学领域典型应用

手术机器人的精准控制 考虑达芬奇手术机器人的器械抓取优化，强化学习无需人工编程控制路径，适应复杂多变的手术环境。：

环境：手术场景的RGBD图像，含器械、器官位置信息；

智能体：手术机器人的机械臂；

学习过程：通过强化学习训练机械臂。“成功抓取器械”获正奖励，“触碰血管”获负奖励，最终学习到避障且精准的抓取策略；

医学影像分析的动态决策 考虑肺结节检测的自适应窗宽窗位调整：

状态：CT图像的当前窗宽窗位、结节候选区域特征；

动作：调整窗宽窗位，如增大窗宽显示更多组织；

奖励：调整后“结节清晰度评分”，高清晰度获正奖励；

目标：通过强化学习使系统自动调整窗宽窗位，最大化结节检测准确率。

2.1.3 判别模型与生成模型

机器学习模型还可分为判别模型、生成模型与描述模型。**判别模型**通常对应分类器或回归模型，聚焦于刻画数据到标签的映射，用于分类、回归、分割等任务，如判断样本类别、预测标签。分割任务本质上是对像素的二分类，即属于或不属于目标结构。

生成模型的核心是学习数据的潜在分布，如高斯分布、混合高斯分布，深度神经网络可对更复杂的分布建模，掌握分布后可生成新样本。常见的深度生成模型包括GAN、VAE 与扩散模型。以扩散模型为例，其通过正向加噪与逆向去噪过程，可从高斯噪声生成图像；若基于CT 图像训练，可生成CT 图像，甚至能从病人图像生成对应健康图像，可用于无监督病灶分割，通过计算病灶图像与生成的健康图像的残差实现。

描述模型旨在发现数据中的模式与相关性，聚类、降维等任务均属于此类。主成分分析（PCA）是典型的描述模型。通过对协方差矩阵进行特征值分解得到主成分，主成分对应数据变化最大的方向。借助PCA 可实现降维，仅保留有限个主成分，即数据变化最显著的维度，大幅降低数据维度，如人脸三维网格数据的降维。聚类算法同样属于描述模型，通过聚集相似数据，挖掘数据中隐藏的关系，如细胞图像

中基于像素相似性的结构分割。三类模型对应不同任务：判别模型适用于分类、分割；生成模型可用于异常检测等；描述模型则用于聚类、降维。在生物医学图像处理中，需根据具体任务选择模型，无需严格区分模型类型。

以深度图像去噪为例，可将其转化为生成任务：利用GAN 的生成器从含噪深度图像生成无噪图像。GAN 包含生成器与判别器，二者构成对抗关系。生成器生成无噪图像，判别器判断图像的“真假”，判断生成的无噪图像与真实无噪图像是否可区分，在此过程中生成模型与判别模型协同作用。

表 2.1: 判别模型与生成模型对比

	判别模型	生成模型
学习目标	学习输入x到标签y的映射关系$P(y\|x)$	学习数据的联合分布$P(x,y)$或边缘分布$P(x)$
能力	分类、回归、分割（直接推断标签）	数据生成、重构、补全（生成符合真实分布的新数据）
典型模型	CNN、SVM、决策树、逻辑回归	GAN、VAE、扩散模型、自编码器

判别模型聚焦“任务求解”

考虑基于CNN的乳腺癌病理图像分类，输入病理切片图像，模型直接输出“阳性/阴性”标签，学习P(标签— 病理图像)；

基于Transformer的医学图像分割中，输入三维MRI图像，模型输出器官分割掩膜，学习P(掩膜—MRI 图像)。

生成模型聚焦“数据生成与增强”

生成对抗网络 通过生成器与判别器的对抗训练生成逼真医学图像，如生成病理切片图像补充稀缺数据集，或生成“正常器官-病变器官”的对比图像辅助医生培训。

扩散模型属于生成模型，从随机噪声中逐步生成高质量医学图像，如基于单张二维X线片生成三维CT 结构，减少患者辐射暴露，无需直接采集三维CT。

变分自编码器也是常见的生成模型，生成具有可控性的医学图像，如调整隐空间向量生成不同阶段的肿瘤图像，模拟肿瘤生长过程。

2.1.4 参数模型与非参数模型

参数模型与非参数模型的核心区别在于是否对数据的底层分布做预设假设，二者均包含参数，仅参数的确定方式不同。

参数模型需预先假设数据分布的形式，并通过训练数据估计分布的参数。例如线性回归假设数据符合线性关系，逻辑回归假设数据服从伯努利分布，神经网络需预先定义层数、神经元数等结构参数。这类模型的优势在于简单可控，训练前即可明确模型复杂度，避免参数规模无限制增长；但缺点是灵活性有限：若预设分布与实际数据不符，模型无法有效捕捉数据规律，导致预测性能下降。因此，使用参数模型需对数据分布有基本认知，确保模型假设与任务匹配。

非参数模型不对数据分布做预设，完全根据观测数据建模，参数规模通常随数据复杂度增加而增长。例如，决策树的深度与节点数由数据分布决定，数据越复杂，树结构越复杂；随机森林通过组合多棵决策树进一步适应复杂数据；核密度估计通过核函数，如高斯核、均匀核叠加各数据点的贡献来估计概率密度，无需假设分布形式。非参数模型的优势在于适应性强，可捕捉数据中的复杂关系，无需严格假设；但缺点是参数量可能随数据规模激增，且需大量数据才能稳定学习底层结构。例如，拟合线性分布时，参数模型仅需2 个数据点，而非参数模型需更多观测才能推断分布形式。

表 2.2: 参数模型与非参数模型对比

类型	特点	典型模型	生物医学应用
参数模型	参数数量固定，与数据量无关	线性回归、逻辑回归、SVM	基于患者年龄、血糖等5个参数预测糖尿病风险
非参数模型	参数数量随数据量增加而增加，灵活性高	K近邻（K-NN）、决策树	基于100例小样本病理图像的细胞分类（K-NN）

数据量充足、任务规律明确时选参数模型，如用线性回归预测肿瘤体积，参数少、可解释性强。数据量少、任务规律复杂时选非参数模型，如用K-NN处理罕见病的病理图像分类，无需假设数据分布。

2.2 凸优化基础

机器学习，尤其是深度学习的核心任务之一是参数优化，通过调整模型参数最小化目标函数/损失函数。凸优化是参数求解的基础，其特性对优化效率与结果有显著影响。

凸优化需满足两个核心条件：目标函数为凸函数，且优化可行域为凸集。

定义2.1 (**凸函数**). 对定义域内任意两点，函数在两点连线上的函数值不大于两点函数值的线性组合，即满足不等式$f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y)\Phi 0 \leq \lambda \leq 1\Psi$。

凸函数的优势在于局部最优解即为全局最优解，便于求解。

定义2.2 (**凸集**). 对集合内任意两点，其连线上的所有点均属于该集合。

凸集确保优化过程中不会因可行域的非凸性导致求解陷入局部最优。

凸优化问题的一般形式为“在凸约束条件下最小化凸目标函数”，约束条件需对应凸函数。凸优化具有优点：存在唯一全局最优解；KKT 条件是最优解的必要且充分条件；对偶性显著，对偶问题与原始问题等价，且更易求解。

常用优化方法

实际模型训练中，优化方法的选择需权衡效率与复杂度：凸优化是指在“凸函数”与“凸集”上寻找最小值的优化问题，其核心优势是“存在全局最优解”，避免陷入局部最优。机器学习中，许多经典模型的损失函数为凸函数，包括：线性回归的最小二乘损失函数；逻辑回归的对数损失函数。

梯度下降（Gradient Descent）

梯度下降法：通过计算目标函数的梯度，沿梯度反方向迭代更新参数，是深度学习中最常用的方法，常采用随机梯度下降SGD。其核心参数为学习率/步长，需设置较小值以避免跨过极值点。沿损失函数的负梯度方向迭代更新参数，逐步逼近最小值；

应用：训练医学图像分割模型，通过批量梯度下降（BGD）、随机梯度下降（SGD）优化数百万个卷积核参数。

牛顿法

牛顿法：利用海森矩阵构建目标函数的二次近似，收敛速度通常快于梯度下降，但需计算海森矩阵的逆，计算复杂度高，尤其高维参数场景。利用损失函数的二阶导数加速收敛，适用于凸函数；可以用于求解支持向量机的最优超平面，在小样本医学数据场景下快速收敛。

非凸问题的近似求解

深度神经网络的损失函数为非凸函数，存在多个局部最优，需通过以下方法近似求解：

优化器改进：使用Adam等自适应学习率优化器，平衡收敛速度与稳定性；

初始化策略：采用预训练模型初始化参数，如用ImageNet预训练的CNN初始化医学图像分割模型，避免陷入较差的局部最优。

内点法

通过障碍函数将不等式约束融入目标函数，将约束优化转化为无约束优化，适

用于处理复杂约束。近端方法：通过近端算子处理非光滑目标函数，如使用L1 正则项的函数，在稀疏优化中应用广泛。

凸共轭方法

通过求解目标函数的共轭函数转化对偶问题，利用对偶性简化求解。当前深度学习框架，如PyTorch、TensorFlow已封装了多种优化算法，实际应用中无需手动实现，但理解优化的本质，对目标函数求解最优参数，对模型调优至关重要。

2.3 流形学习降维方法：从线性到非线性的医学数据表征优化

高维数据的可视化需借助流形学习、降维等算法技术，因人类视觉系统仅能感知三维及以下维度的信息。在数据可视化领域，降维是核心问题之一；在图像处理中，降维同样常用于高维特征的可视化呈现。例如，利用t-SNE 算法对Mask R-CNN模型中三组对称咀嚼肌的特征进行二维可视化。

线性降维：主成分分析（PCA）原理与生物医学应用

PCA通过“线性投影”将高维数据映射到低维子空间，保留数据的主成分，步骤为：

数据预处理：对高维特征如医学图像像素、患者生理指标进行“去中心化”减去均值与“标准化”，除以标准差，消除量纲影响；

协方差矩阵计算：通过协方差矩阵刻画特征间的线性相关性；

特征值分解：对协方差矩阵进行特征值分解，特征值越大对应“主成分”的信息贡献度越高；

维度选择：选取前d个最大特征值对应的特征向量/主成分，将高维数据投影至d维子空间，$d \ll$ 原始维度。

高维医学数据的压缩与可视化

案例：三维人脸图像降维

原始数据：1024×1024分辨率的三维人脸深度图像（维度10^6）；

降维效果：通过PCA选取数十个主成分，可保留95%以上的人脸结构信息，维度大大降低，显著降低后续模型的计算复杂度；

案例：多模态医学数据融合

输入：CT灰度特征、MRI T1加权信号特征的高维特征向量；

降维目标：通过PCA将多模态特征融合为低维特征，用于肿瘤分级预测，避免“维数灾难”。

非线性降维：流形学习的方法

流形学习假设“高维医学数据位于低维流形上”，通过非线性映射保留数据的局部/全局几何结构，适用于线性方法难以处理的复杂数据，如瑞士卷、细胞图像特征。

等距映射（Isomap）

保持原始高维空间中的“测地距离”（流形表面的最短路径距离）在低维嵌入空间中不变。

计算方法：

1. 构建近邻图：将每个数据点与其K个近邻连接，边权重为欧氏距离；

2. 测地距离计算：通过Dijkstra算法计算任意两点间的最短路径，考虑测地距离；

3. 多维缩放（MDS）：将测地距离矩阵映射到低维空间，保留全局距离结构；

电镜图像的亚细胞结构特征可视化，将1000维的亚细胞特征通过Isomap 降至2维，清晰区分细胞核、线粒体等结构的特征簇。

t分布随机邻域嵌入（t-SNE）

通过“高维空间高斯分布-低维空间t分布”的KL散度最小化，保留数据的局部近邻关系，适用于高维医学数据的可视化。

计算方法：

1. 高维空间：通过高斯核函数计算样本间的相似性；

2. 低维空间：通过t分布/柯西分布计算嵌入后样本的相似性；

3. 优化目标：最小化高维与低维相似性分布的KL散度，实现局部结构保持；

医学应用：病理图像的特征可视化，将基于CNN提取的1000维病理特征通过t-SNE 降维，不同类型的肿瘤特征在低维空间中形成明显分离的簇，辅助医生直观判断肿瘤亚型。

局部线性嵌入（LLE）

假设高维流形的局部区域近似线性，通过“局部近邻的线性组合”表示数据点，嵌入低维空间时保持该线性关系；

计算方法：

1. 近邻选择：为每个数据点选择K个近邻；

2. 权重计算：通过最小化重构误差求解“近邻线性组合的权重”；

3. 低维嵌入：在低维空间中找到数据点位置，使近邻权重与高维空间一致，通过特征值分解求解；

医学应用：脑MRI图像的功能连接分析，将脑区功能连接的高维特征通过LLE降维，用于阿尔茨海默病的早期诊断，保留脑区局部功能关联结构。

拉普拉斯特征映射（Laplacian Eigenmaps）：图结构保持

将数据点构建为图，节点为样本，边权重为相似度，通过图拉普拉斯矩阵的特征值分解实现降维，保留“近邻样本在低维空间中仍近邻”的性质；

计算方法：

1. 图构建：以像素/超像素为节点，通过高斯函数计算边权重；

2. 图拉普拉斯矩阵计算：$L = D - W$，D为度矩阵，W为权重矩阵；

3. 特征值分解：对L进行特征值分解，选取最小特征值对应的特征向量作为低维嵌入；

医学应用： 细胞图像的分割特征提取，将细胞图像的高维像素特征通过拉普拉斯特征映射降至20 维，嵌入特征可直接用于K-Means聚类，生成分割结果，且保留细胞的局部形态结构。

2.4 机器学习范式在生物医学图像处理中的选择策略

在实际生物医学图像处理任务中，需根据“数据量、标签可用性、任务目标”选择合适的学习范式：若标签充足优先选择有监督学习，如U-Net分割；若标签稀缺则采用无监督/半监督/自监督学习，如K-Means聚类、MAE预训练；若涉及动态决策则适用强化学习，如手术机器人控制。随着基础模型，如MedCLIP、SAM 的发展，“预训练+ 微调”的自监督范式将成为生物医学图像处理的主流，进一步降低对标注数据的依赖，推动精准医疗的落地。

2.5 实例：BiomedParse

生物医学图像分析是研究中的重要工具，广泛应用于从细胞器到器官的多尺度研究中。该领域通常包括三个核心任务：分割、检测和识别。传统方法通常为每个任务设计独立工具，忽视了任务间的内在关联。当前研究多集中于分割任务，而忽略了如元数据和目标类型名称等语义信息，不仅限制了模型性能，多数分割工具仍依赖用户提供精确边界框。手动绘制框需专业知识；矩形框难以准确描述不规则目标；在处理目标密集的图像时效率极低。BiomedParse 是一种生物医学图像分析基础模型 [54]，将三大任务统一为一个图像解析框架，实现联合学习与推理。通过文本提示中的语义标签即可完成高质量分割，无需用户提供边界框，显著降低了交互负担。

训练过程中的主要挑战是数据缺失。BiomedParseData数据集包含340 万组“图像-掩码-标签”三元组和680 万组“图像-掩码-描述”三元组，涵盖超过100 万张图

表 2.3: 机器学习不同类型的特点与应用对比

学习类型	优势	适用场景	典型模型/方法	生物医学应用案例
有监督学习	任务精度高，可解释性强	标签充足，如公开医学数据集	U-Net、SVM、随机森林	腹部多器官分割、肺结节良恶性分类
无监督学习	无需标注，降低成本	标签稀缺，如罕见病影像	K-Means、自编码器、Isomap	亚细胞结构分割、医学图像去噪
半监督/弱监督	平衡标签成本与模型性能	少量标签+大量无标签数据	教师-学生模型、弱监督检测	脑部肿瘤分割、罕见病影像诊断
强化学习	适应动态环境，自主优化策略	机器人控制、动态决策任务	DQN	手术机器人器械抓取、CT 窗宽窗位自适应调整
自监督学习	利用数据内在关联，泛化能力强	预训练任务，如特征提取、模型初始化	MAE、对比学习、NeRF	医学图像预训练、三维器官重建

像、9 种成像模态和82 类生物医学目标。BiomedParse 能够学习不同目标类别的典型形态，其感知机制更接近人类视觉认知。仅通过输入文本提示，即可实现精准分割，无需任何位置先验。该模型特别擅长处理形状不规则或结构复杂的目标，并具备零样本识别能力，可在无提示情况下自动识别图像中的所有目标。在多项实验评估中，BiomedParse 在分割任务上达到了最先进水平，性能显著优于MedSAM 和SAM 等已有方法。仅凭文本提示，它就展现出卓越的可扩展性，而传统方法需大量人工操作才能达到类似效果。该模型还能准确拒绝无效提示，并在不规则目标分割中表现出明显的准确性提升。

为提升语言输入的鲁棒性，利用GPT-4 构建了生物医学对象本体，包括器官、异常和组织学三个顶层类别，15 个元对象类型和82个细粒度类别。通过人工审核和标准化工具对类别名称进行映射和校验，确保本体结构的一致性和可扩展性。利用GPT-4 为每个语义标签生成同义文本描述，显著增强了模型对多样化提示的适应能力。通过设计统一的任务表述模板并进行人工质量管控，有效减少了错误和幻象描述。训练过程中随机选择提示文本，使模型能够学习多样化的语义表达。

模型结构

BiomedParse 采用模块化结构，包含图像编码器、文本编码器、掩码解码器和元对象分类器四大核心组件。模型以SEEM 框架为基础，支持多种骨干网络，SAM-ViT和预训练语言模型，如PubMedBERT。图像与文本提示分别通过编码器处理，特征嵌入经掩码解码器生成分割结果，元对象分类器输出对应的语义类别，最终实现端到端的图像解析功能。

模型训练细节

BiomedParse 的训练以文本引导的分割任务为核心。训练过程中最小化以下加权损失函数：

$$\mathcal{L} = a\mathcal{L}_{c\_CE\_text} + b\mathcal{L}_{m\_BCE\_text} + c\mathcal{L}_{m\_Dice\_text},$$

其中，c 表示基于交叉熵（CE）的元概念分类损失，m 表示基于二元交叉熵（BCE）与Dice 损失的掩码预测损失。定义如下：

$$\begin{aligned}
\mathcal{L}_{c_CE_text} &= -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \\
\mathcal{L}_{m_BCE_text} &= -\frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \left[m_p \log(\hat{m}_p) + (1-m_p)\log(1-\hat{m}_p)\right], \\
\mathcal{L}_{m_Dice_text} &= 1 - \frac{2\sum_{p\in\mathcal{P}} m_p \hat{m}_p}{\sum_{p\in\mathcal{P}} m_p + \sum_{p\in\mathcal{P}} \hat{m}_p},
\end{aligned}$$

y 表示真实元类别的one-hot 向量（$c = 1, \ldots, C$），$\hat{y}$ 为预测的元概念概率分布。m_p 表示像素$p \in \mathcal{P}$ 的真实二值掩码，$\hat{m}_p$ 为模型预测的像素概率。训练过程中还引入了视觉采样器损失及其他辅助损失（SEEM [61]）以支持交互式空间优化。

2.6 小结：机器学习在生物医学图像处理中的意义

机器学习通过“数据驱动”的范式，将生物医学图像处理从“人工主导”转向“人机协同”：

有监督学习在标注充足场景下，如大规模公开医学数据集实现高精度任务；

无监督/自监督学习在数据稀缺场景下，如罕见病影像突破标注限制，降低应用门槛；

生成模型通过数据增强与图像生成，缓解医学数据“量少质差”的痛点；

针对维数灾难、过拟合等问题的解决方案，确保模型在临床场景中的可靠性与泛化能力。

随着基础模型的发展，机器学习将进一步实现“跨模态、跨任务”的生物医学图像分析，为精准医疗提供更强大的技术支撑。

回顾机器学习的发展历程，图像处理的范式已发生显著变化：从早期的浅层神经网络，如自编码器到如今的深度神经网络，其含海量参数，再到近年的基础模型/大模型，技术迭代持续推动问题解决能力的提升。这种变化不仅改变了学术研究生态，也逐步影响临床实践，如AI 辅助诊断。实际应用中，模型选择需结合任务特性：即使CNN 已发展十余年，生成模型、大模型等新兴方法与传统模型仍并存，需根据数据规模、任务复杂度如多模态图像、小样本场景选择合适模型。模型性能的评价同样关键：例如生成模型的图像生成质量、分割模型的边界精度等，需通过客观指标与临床需求结合的方式评估，这也是模型落地的核心挑战之一。

第三章　机器学习基础II

3.1　机器学习中的关键问题

3.1.1　维数灾难

维数灾难（Curse of Dimensionality）由数学家理查德贝尔曼于20 世纪50 年代提出，指高维数据带来的数据处理难度激增问题，在当前机器学习中极为常见。

实际数据的**高维性体现**在多方面：二维图像的维数为长×宽×通道数；三维图像，如CT、MRI 的维数更高；考虑患者治疗过程中的多次采集数据的时序数据，因时间维度的引入进一步提升维数；此外，通过深度神经网络等方法提取的特征也通常具有高维性。高维数据的核心问题在于信息获取效率下降，具体表现为：

数据稀疏性：当数据集规模固定时，随维数升高，数据点在高维空间中分布愈发稀疏。例如1000 个数据点在二维平面上可能较密集，但在高维空间中几乎彼此孤立，导致模型难以捕捉数据分布。

计算复杂度激增：以欧式距离计算为例，高维数据的距离计算涉及更多维度的平方和，不仅计算量增大，还可能因所有样本间距离差异缩小而失去区分度。此外，高维空间的体积随维数呈指数增长（若每一维长度为L，则d 维空间的超体积为L^d），进一步加剧了数据稀疏性与计算负担。当数据维度，如医学图像的像素数、多模态特征数过高时，会出现“样本稀疏”，高维空间中样本间距离增大，模型难以学习规律与“计算复杂度剧增”的问题。例如1024×1024分辨率的CT图像像素维度达10^6，直接输入模型会导致参数爆炸。解决方案包括：

降维：通过线性方法或非线性方法将高维数据映射到低维空间，在保留关键信息的同时降低处理难度。例如，图像分割任务中可先对数据降维，再在低维空间中进行后续处理，如将电镜图像的高维像素特征降维；

特征选择：从高维特征中筛选出关键特征，减少冗余维度。决策树的特征选择过程，选择能最大化信息增益的特征，通过贪婪搜索选择最优特征，降低模型输入的维数，如在肺结节检测中，仅保留“灰度均值、纹理熵”等有效特征，剔除冗余的像素位置特征；

模型优化：采用CNN的卷积操作自动提取低维局部特征，或使用Transformer 的注意力机制聚焦关键区域，减少无关维度的干扰。

正则化：通过约束参数降低模型复杂度，间接缓解高维数据导致的过拟合问题。

3.1.2 过拟合与欠拟合

过拟合：模型在训练集上表现优异，但在测试集上性能显著下降，成因包括“训练数据量不足”“模型复杂度过高”“标注噪声”，如专家标注的解剖边界不一致。

考虑基于小样本训练的医学图像分割模型，在训练集上Dice系数达0.95，但在新测试集上仅0.7，因模型过度拟合训练集的标注偏差。

欠拟合：模型无法捕捉数据规律，训练集与测试集性能均差，成因包括“模型复杂度不足”“特征质量差”，如用线性模型分割形状复杂的肿瘤。

解决方案：

表 3.1: 常见模型问题与解决措施

问题	解决措施
过拟合	**数据增强**：对医学图像进行旋转、翻转、缩放、加噪，扩大训练集； **正则化**：在损失函数中加入L1/L2正则项，限制模型参数规模； **早停（Early Stopping）**：当验证集性能下降时停止训练； **模型轻量化**：使用MobileNet等轻量模型，减少参数数量。
欠拟合	**提升模型复杂度**：将线性模型替换为CNN、Transformer等非线性模型； **优化特征**：增加“纹理、形状”等有效特征； **补充数据**：收集更多标注数据，提升数据多样性。

过拟合与欠拟合是模型训练中常见的问题。欠拟合源于模型复杂度不足，无法捕捉数据潜在模式，表现为训练集与测试集性能均较差，损失函数值难以下降；过拟合则因模型过于复杂，对训练数据拟合过度，表现为训练集性能优异但测试集性能显著下降。从偏差- 方差权衡视角看：过拟合模型通常偏差低、方差高，对训练数据扰动敏感，泛化能力差；欠拟合模型则偏差高、方差低，与真实情况差异大，无论对训练集还是测试集，性能均不理想。

训练中需关注偏差-方差的平衡：模型训练时，训练集误差会随迭代逐渐减小，但存在一个拐点，超过该点后，验证集或测试集性能会逐步下降，即进入过拟合阶段。因此，训练中需关注模型在验证集上的性能，避免过度迭代。

避免过拟合有诸多经验准则：若数据集较小，交叉验证是常用方法，如K 折交

叉验证；在极端情况下，如数据量极少，可采用留一法，每次留一个样本作为测试集，其余用于训练。交叉验证可充分利用数据，最终通过所有K 折的结果综合评价模型性能。

过拟合的避免策略

在实验过程中，过拟合的避免是模型训练的核心问题之一。除交叉验证外，还有多种常用策略。

交叉验证在数据集规模有限时应用广泛。若数据集规模较小，通常需提供交叉验证结果，否则审稿人会要求补充。这是因为当数据集较小时，单一的数据划分可能导致模型性能评估偏差，例如某一划分下模型性能优异仅源于对该划分的特殊拟合；而交叉验证通过多次数据划分并对模型性能取平均，可实现对模型的公平评价。同时，交叉验证使数据集中的所有样本均参与模型的训练与评价，一定程度上降低了过拟合风险。

正则化是另一类重要的过拟合抑制方法，例如线性拟合中求解法线方程时，为改善方程条件数，会引入吉洪诺夫正则项。在模型参数优化中，常见方式是为目标函数添加L1 或L2 正则项，对应岭回归（Ridge Regression）与拉索回归（Lasso Regression）。需注意的是，添加正则项会降低目标函数中数据项的占比，可能导致模型对训练数据的拟合精度略有下降，但从抑制过拟合、提升泛化能力的角度来看，其作用显著。L1 与L2 正则化可通过约束参数范数降低模型复杂度，进而增强模型的泛化能力。实际应用中，正则项的形式多样，例如图像配准任务中，损失函数常包含平滑项以约束形变场的平滑性：尽管平滑性无法直接保障形变后图像与目标图像的一致性，本质上是在削弱数据项的权重，但对稳定模型训练、避免过拟合具有积极作用。

模型剪枝同样可有效抑制过拟合。以决策树为例，其构建过程通过不断依据最优判决条件分裂节点直至达到预设深度终止；而剪枝操作通过降低树的复杂度，如减少深度或节点数，可直接降低过拟合风险。

提前终止是模型训练中常用的实践策略。模型训练的迭代次数或epoch 数需结合学习曲线确定：若仅关注训练集性能，持续迭代会使训练集误差不断减小，但验证集性能可能在某一拐点后开始下降。该拐点意味着模型即将进入过拟合阶段，此时需及时终止训练。通常做法是在训练集中划分验证集，当验证集性能连续多次未提升或开始下降时，停止迭代。

数据增强通过扩充数据量缓解过拟合。过拟合的根本原因之一是数据量有限而模型复杂度较高，模型易对有限数据过度拟合，却未真正捕捉数据的内在分布，导致对数据微小扰动的鲁棒性极差。数据增强通过对原始数据进行变换，如图像的翻转、旋转、裁剪，或对CT 图像施加刚性/ 非刚性形变等，生成新样本，间接增加数据量，使模型更易学习数据的本质规律而非噪声。

过拟合的直观表现可通过分类任务的**决策边界观察**：若两类数据的决策边界过

于复杂，远超数据内在分布所需的复杂度，则模型在未来预测中易出现误差。实际应用中，过拟合的核心问题是模型在验证集或测试集上的性能显著下降，且在多种评价指标下均表现不佳。

欠拟合的应对方法

欠拟合与过拟合相对，源于模型复杂度不足，无法捕捉数据的潜在模式。应对欠拟合的方法是提升模型对数据的拟合能力：

增加模型复杂度：例如若线性模型无法满足需求，可改用非线性模型；在深度学习中，可增加神经网络的层数、神经元数量或引入更复杂的网络结构，如注意力机制。需注意的是，当前深度学习中欠拟合相对少见，深度神经网络通常具有庞大的参数量，动辄数十万、数百万甚至基于基础模型的亿级参数，模型本身的复杂度已足以应对多数任务。

调整正则化强度：若过度使用正则化，如强L1/L2 约束，会强制简化模型导致欠拟合，此时需降低正则化强度或移除不必要的正则项。L1 正则化通过约束参数一范数促使部分参数为0，L2 正则化通过约束参数二范数限制参数幅值，二者均会降低模型复杂度，因此需根据拟合情况动态调整。

延长训练时间：若模型尚未收敛，如损失函数仍在缓慢下降，需增加迭代次数，使模型有充足时间学习数据模式。此外，改进表征学习，即提升特征质量也能缓解欠拟合，良好的特征可更清晰地刻画数据规律，降低模型的学习难度。在传统图像处理中，特征工程/手动设计特征是关键；而在深度学习中，通常通过深度神经网络自动提取特征，若特征质量不足，可通过调整网络结构，例如增加特征提取层引入预训练特征来优化。

3.1.3 偏差-方差权衡与模型泛化

从偏差-方差视角可更清晰地理解过拟合与欠拟合：过拟合模型通常具有低偏差、高方差，对训练数据拟合充分/偏差小，但对数据扰动敏感/方差大，泛化能力差；欠拟合模型则表现为高偏差、低方差，无法有效拟合数据/偏差大，且对训练集与测试集的拟合效果均较差/方差小。

模型训练的核心目标是实现**偏差- 方差的平衡**：需使模型既能充分捕捉数据规律/偏差适中，又对未见过的数据有稳定表现/方差适中。这类模型在训练集上能学到符合任务需求的函数，如分类器、回归模型或生成模型，且在验证集与测试集上均能保持优异性能，即具备良好的泛化能力。

3.1.4 经验风险最小化与模型学习

机器学习的核心要素，数据、算法、模型、预测中，数据的规模与质量直接决定模型性能，对深度神经网络而言，充足且高质量的数据通常能支撑更优模型的训

练，进而提升在线预测效果。模型的学习方式，有监督、无监督、强化学习需与任务特性匹配，在生物医学图像处理中体现得尤为明显：有监督学习需依赖图像标签，如分割任务的金标准掩码、标记点检测的坐标标注，通过标签与预测结果的差异优化模型；无监督学习则无需人工标注，依托像素/超像素的相似性，如灰度分布、纹理特征自主挖掘图像结构，典型如聚类算法，在电镜细胞图像中，聚类可通过像素特征相似性分配标签，实现线粒体、微粒等亚细胞结构的自动分割，全程无需人工介入。如何训练上述模型？

模型训练的本质是**经验风险最小化**：通过构造与任务匹配的损失函数，即经验风险的量化形式，优化模型参数以最小化损失。这里的“经验风险”是对真实风险的近似，真实风险指模型对所有可能数据的平均预测误差，但实际训练仅能接触有限样本，即使数万张医学图像，在高维图像空间中仍属稀疏采样，因此需通过有限样本定义损失函数。

以训练曲线为例：随模型复杂度增加，训练集误差持续下降，但验证集误差会出现拐点，拐点前模型泛化能力提升，拐点后进入过拟合/验证集误差上升。训练需在验证集性能最优时终止，而非依赖测试集，这一过程本质是在偏差与方差间找平衡：过拟合对应低偏差、高方差模型对训练数据拟合过好，对扰动敏感；欠拟合对应高偏差、低方差/模型过于简单，无法捕捉数据规律。

生物医学图像任务中的损失函数设计

不同任务需定制化损失函数，生物医学图像因数据特性，如模态多样性、标注稀缺性，损失函数设计更需结合领域特点。

无监督图像分割：以深度聚类模型为例，损失函数需度量“模型输出的聚类分配概率”与“数值聚类算法生成的伪标签概率”的一致性，常用交叉熵实现。例如电镜图像分割中，伪标签由传统聚类，如K-means生成，模型通过最小化交叉熵学习像素聚类规律，同时可引入“谱嵌入项”（将图像映射至谱域保留全局结构）与“平滑项”约束相邻像素标签一致性，进一步优化分割连续性。

二/三维图像配准：在从二维X 光片推断三维形变场的任务中，采用自监督学习：损失函数的数据项定义为“形变后三维图像的投影”与“输入X 光片”的差异，如使用均方误差，无需真实形变场标注。模型通过最小化该差异，自主学习二维投影与三维结构的关联。

经典模型示例： 线性回归：损失函数为均方差MSE，即预测值与真实值的平方和，适用于剂量预测、器官体积回归等任务，但对异常值敏感，如CT 图像中的金属伪影会显著放大MSE。

逻辑回归/分类任务：常用交叉熵损失，二分类时度量预测概率与0/1 标签的差异，多分类时扩展为多类交叉熵，如肿瘤亚型分类；SVM 则采用合页损失，通过最大化分类间隔优化模型。

“经验风险最小化”：面对一个可能含标签或不含标签的数据集，需根据任务定

义能量函数/损失函数以构造经验风险，通过最小化经验风险优化模型参数，最终用训练好的模型在测试或推理阶段对未知数据进行预测与决策。

那么如何实现“经验风险最小化”？核心在于“优化”，只要定义好模型架构与损失函数，便可通过工具包，如PyTorch、TensorFlow完成参数优化，常用的优化方法便是梯度下降。

梯度下降

梯度下降是当前最基础也最常用的优化方法。其针对定义好的损失函数L，沿着其“负梯度方向”，以一定的“学习率”/步长逐步迭代，最终实现函数值的最小化。

为什么要选择负梯度方向？从数学角度看，任意函数在局部都可通过一阶泰勒展开进行线性近似，若要使函数值下降，需让“梯度与参数变化量的乘积”为负值，即参数变化方向需与梯度方向相反，即负梯度方向。因此，只要能计算损失函数的负梯度方向，并合理选择步长，即训练中定义的“学习率”，通常取值较小，就能通过迭代逐步降低函数值。

梯度下降具体可分为三种形式：**全梯度下降、随机梯度下降（SGD）与小批量梯度下降**。

损失函数的优化难点：以CIFAR-10 数据集上的损失函数为例，其函数值曲面极为复杂，优化过程中易陷入局部极值。若函数是凸函数，可保证找到全局最优，但深度神经网络的目标函数往往非凸，因此优化需更谨慎。以生物医学图像处理中常用的UNet为例：用二维UNet 完成分割任务时，若在监督学习中定义了基于MSE或Dice 系数的损失函数，优化的核心便是卷积核参数，从输入图像到输出分割结果的过程中，需经过多次卷积操作，而梯度下降优化的正是这些卷积核的参数。此外，损失函数中不同项的组合权重等超参数，也需通过优化实现最优。

三种梯度下降方法的核心差异在于“每步迭代使用的数据量”：

全梯度下降需用整个训练集，如按8:2 划分数据集后，80% 的训练数据需全部参与每步参数更新。其优势是“准确性高”，因使用所有数据计算梯度，能对函数做出最准确的局部近似，每步参数更新更可靠；但缺点也很明显：每步迭代需计算所有数据的损失，求和运算导致计算量极大。

随机梯度下降则是“单次迭代仅用单个训练样本”：与全梯度下降形成鲜明对比，其单步计算复杂度极低，能大幅加快训练进程，相同时间内可完成更多迭代，尤其适用于大规模数据集与复杂模型。不过，单样本计算的梯度方向未必是“最优方向”，精确性有限；但也正因更新具有随机性，反而可能避免陷入局部极值。训练中观察学习曲线会发现，随机梯度下降的损失函数值可能出现震荡，但只要整体呈下降趋势并趋于稳定，即为可接受状态。

小批量梯度下降则是“折中方案”：每步迭代使用一小部分样本，样本数量m 远小于总样本量n，计算代价与精度均介于前两者之间。

三者的对比可总结为：全梯度下降用全量数据，计算量大、存储需求高，但更新噪声最小、精度最高，收敛速度最慢；随机梯度下降用单样本，计算量与存储需求低，更新噪声最大，收敛速度最快但易震荡；小批量梯度下降则在各方面均处于中间位置。实际训练中，随机梯度下降因“计算效率高”最为常用，选择时需在“精度”与“速度”间权衡，结合任务需求与算力条件确定最优方案。深度学习工具包均支持多种梯度下降方法，可灵活选择。

二分类器的性能评价

模型训练完成后，需通过客观指标评价性能。二分类任务，如病灶有无检测、前景背景分割的评价体系以混淆矩阵为基础，衍生出多项指标，需根据生物医学任务特点选择：

评价指标

- 准确率（Accuracy）：整体预测正确比例

$$(TP + TN)/(TP + TN + FP + FN)$$

  适用于类别均衡场景，如正常/病变组织占比接近的切片分类，但在类别不平衡时失效，如病灶仅占图像1% 时，全预测为背景也能获高准确率。

- 精确率（Precision）：预测为正的结果中真实为正的比例

$$TP/(TP + FP)$$

  适用于误报代价高的场景，如肿瘤筛查中需减少“良性判为恶性”的焦虑。

- 召回率（Recall）：真实为正的样本中被正确预测的比例

$$TP/(TP + FN)$$

  适用于漏报代价高的场景，如病灶分割需尽可能找出所有病灶，避免遗漏。

- F1 分数：精确率与召回率的调和平均

$$2 \times PrecisionxRecall/(Precision + Recall)$$

  平衡二者矛盾，如手术导航中需同时减少误报与漏报。

- ROC 曲线与AUC：ROC 曲线以假正率

$$FPR = FP/(FP + TN)$$

  为横轴、真正率（TPR=Recall）为纵轴，直观反映不同阈值下的分类性能；AUC（曲线下面积）量化整体性能，值越接近1 说明分类器越稳健，如用于比较不同模态图像的病灶检测模型。

生物医学图像中，阈值选择直接影响结果：分割任务中常用0.5 作为概率阈值，大于则判为前景，但需根据临床需求调整，如肿瘤靶区勾画可降低阈值以提高召回率，减少漏诊风险。

3.1.5 最大似然估计与参数推断

最大似然估计（MLE）是模型参数推断的核心方法，与经验风险最小化存在内在关联：通过最大化“模型参数生成观测数据的概率”（似然函数）求解最优参数，加负号后可转化为损失函数最小化问题。

设观测数据为D，模型参数为θ，似然函数为$P(D|\theta)$。MLE 的目标是找到

$$\theta^* = \arg\max_{\theta} P(D|\theta)$$

若数据独立同分布，似然函数为各样本似然的乘积；取对数后转化为求和，对数似然函数

$$L(\theta) = \sum \log P(x_i|\theta)$$

简化计算的同时不改变极值位置。生物医学数据示例对CT 值分布建模时，假设数据服从正态分布$N(\mu, \sigma^2)$，MLE 需估计μ 与σ^2：

对数似然函数为

$$L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i - \mu)^2$$

求导并令导数为0，得样本均值$\mu = \frac{1}{n}\sum x_i$l，样本方差为

$$\sigma^2 = \frac{1}{n}\sum(x_i - \mu)^2$$

MLE 的核心优势是一致性：随样本量增加，估计值收敛于真实参数，如大样本CT 数据中，MLE 估计的器官CT 值分布参数更接近真实生理状态。其与经验风险最小化的转换关系，如将负对数似然作为损失函数，使其成为生物医学模型训练的基础工具。例如在病灶概率预测中，交叉熵损失本质就是负对数似然，通过MLE求解最优参数。

传统编程通过固定指令控制机器执行，而机器学习编程的核心是“定义损失函数+ 优化参数”：无需手动编写图像分割的规则，如“病灶为高密度区域”，只需构造损失函数，如分割结果与金标准的交叉熵，借助优化算法，如SGD让模型自主学习参数。这种范式在生物医学图像处理中尤为高效。面对复杂模态，如MRI 的多序列图像，模型可通过数据驱动挖掘特征，远超人工设计规则的能力。

最大似然估计（MLE）的核心价值不仅在于与经验风险最小化的转换关系，更在于其统计理论支撑：

收敛性与最优性：随着样本量增加，MLE 估计值会收敛于真实参数，且在所有无偏估计中具有最小方差。这由克拉莫- 拉奥下界（Cramer-Rao Lower Bound）保

障：任何参数估计的方差都无法小于由费舍尔信息（Fisher Information）决定的下界，而MLE 恰好能达到该下界。最终估计值会围绕真实值形成正态分布，这为模型收敛性提供了明确的理论依据，例如在CT 值分布建模中，大样本下MLE 估计的均值与方差会稳定接近真实生理状态下的CT 值分布参数。

局限性：MLE 需预先确定模型参数形式，如假设数据服从正态分布时需明确估计均值与方差，且对模型假设高度敏感，若实际数据分布与假设不符，如将非正态的MRI 信号强度强行按正态分布建模，估计结果会严重偏差。

尽管存在局限，MLE 仍是损失函数设计的核心：许多常见损失函数，如交叉熵本质上是负对数似然，通过MLE 框架将参数估计转化为经验风险最小化问题，这在生物医学图像的病灶概率预测、器官体积回归等任务中被广泛应用。

3.2 正则化：平衡经验风险与模型复杂度

正则化是解决过拟合的核心手段，通过在损失函数中引入正则化项，实现“经验风险最小化”与“模型复杂度控制”的平衡，在生物医学图像处理中尤为关键。

L2 正则化（岭回归）：

在损失函数中加入参数的L2 范数平方项，例如线性回归中为$\lambda\|\boldsymbol{w}\|_2^2$，通过惩罚大系数迫使参数值整体缩小。从几何角度看，L2 正则化在参数空间中形成圆形约束，最优解为经验风险等值线与圆形约束的切点，而非经验风险的全局最小值。这种方式能抑制过拟合，如在剂量预测任务中，避免模型过度依赖少数异常样本的CT 值，同时改善数值计算条件，降低矩阵不可逆风险。

L1 正则化（Lasso 回归）：

引入参数的L1 范数项（$\lambda\|\boldsymbol{w}\|_1$），在参数空间中形成菱形约束。由于L1 范数在坐标轴处的“棱角”特性，最优解更可能落在坐标轴上，部分参数为0，实现参数稀疏化。这在生物医学图像的特征选择中极具价值。

组合正则化：

结合L1 与L2 正则化，如弹性网络，可同时实现参数缩小与稀疏化。例如在三维形变场估计中，既通过L2 范数约束形变场整体平滑，避免局部畸变，又通过L1 范数稀疏化部分无关体素的位移参数，减少计算冗余。

除了损失函数中的**显式正则化项**，实际训练中还有多种**隐性正则化手段**：

随机失活（Dropout）：训练时随机将部分神经元设为0，防止模型过度依赖特定神经元，如在病灶检测网络中，避免网络仅关注某一固定位置的假阳性区域。

提前停止：当验证集误差不再下降时终止训练，本质是通过控制训练轮次限制模型复杂度，如在眼底图像分割中，避免模型为拟合训练集中的血管噪声而过度复杂化。

任务特定正则化：在生物医学任务中，常结合领域知识设计正则化项。例如在人体器官分割中加入“对称性约束”，如左右肺叶形变需对称，或在三维形变场估计中加入“平滑项”，约束相邻体素的位移梯度，避免生理上不合理的突变。

交叉验证：有限数据下的模型稳健性评估。当数据集规模有限，如罕见病的医学图像样本不足，单次训练- 测试集划分可能导致模型评价偏差，交叉验证通过多轮数据重划分实现更可靠的性能评估。

k 折交叉验证：将数据集均匀划分为k 份，每次取k-1 份训练、1 份验证，重复k 次后取性能均值。例如在脑肿瘤切片分类中，5 折交叉验证可让每个切片都参与一次验证，既避免数据泄露，即验证集不参与训练，又充分利用有限样本，所有数据均用于训练与验证。其优势在于：降低因单次划分偶然导致的性能波动。某轮划分中训练集恰好包含大量易分类样本；更全面反映模型在不同数据子集上的泛化能力。不同医院采集的MRI 图像可能存在设备差异，交叉验证可暴露模型对这种差异的敏感性。

留一交叉验证： 极端情况下的k 折验证，k = 样本数，每次仅留1 个样本作为验证集。适用于数据极度稀缺的场景，但代价高昂，需训练n 次模型，n 为样本数，在深度神经网络等复杂模型中几乎不可行。

需要考虑**数据分布一致性**：交叉验证需假设各折数据分布一致，若数据存在时序性，例如不同治疗阶段的CT 图像或批次差异，如不同扫描仪采集的图像，需按时间/ 批次分层划分，确保每折包含各阶段的样本，避免分布偏移导致的评价失真。

类别平衡：在病灶分割等类别不平衡任务中，需采用“分层k 折”，各折中前景/背景比例与原数据集一致，避免某折验证集因病灶样本过少而低估模型性能。

3.3 PAC 学习：机器学习的理论可学习性保障

PAC（概率近似正确）学习理论从数学上回答了“模型能否从有限数据中学习”的问题，为机器学习提供了理论基础。PAC理论指出若存在学习算法，能以高概率（1- δ ）学到泛化误差不超过 ε 的模型，则该任务是PAC 可学习的。

假设空间：模型可能的函数集合，如线性回归的所有直线方程，其规模直接影响样本需求。假设空间越大，如深度神经网络的复杂函数族，需更多样本才能保证PAC 可学习。

样本复杂度：满足PAC 学习要求的最小样本量，与误差容限 ε 、置信度 δ 及假设空间规模相关。例如在图像分类中，若要求泛化误差 ε ≤0.01、置信度1- δ ≥0.95，且假设空间规模为1000，由PAC 不等式估算样本量需至少990。

PAC 理论的价值在于为生物医学任务的样本采集提供指导。例如在设计新的病

灶检测系统时，可根据允许的误差与置信度，通过PAC 不等式估算所需的最小标注样本量，避免盲目收集数据。

3.4 调参：超参数优化的实践策略

机器学习模型的性能高度依赖超参数，如正则化权重λ、学习率等，调参的本质是在超参数空间中寻找最优组合，以最大化模型泛化能力。

格点搜索： 在超参数的预设范围内均匀取点，逐点训练模型并评价性能。例如在岭回归中，若λ 的候选值为0.1,1,10，可分别训练3 个模型，选择验证集性能最优的λ。若涉及多参数，如同时调λ 与学习率η，则需在二维网格上遍历，如λ 取3 个值、η 取3 个值，共9 种组合。

经验准则： 当超参数较多时，考虑损失函数包含多个正则化项，可通过“数量级对齐”简化调参。让数据项与正则化项的初始值处于同一数量级，例如均为10^{-z}，再微调优化。这在生物医学任务中尤为实用，如形变场估计中，平滑项与数据项的权重常按1:10 初步设置，再根据分割结果调整。

多参数调参面临“维度灾难”，超参数数量增加时，候选组合呈指数增长。实践中可采用“分层调参”：先固定其他参数，单独优化每个参数的大致范围，再缩小范围精细搜索；或结合领域知识减少参数维度。在医学图像配准中，根据器官弹性特性固定平滑项的数量级。

3.5 实例：三维细胞内结构表征学习

随着显微成像技术的飞速发展，细胞生物学已步入大数据时代。海量的图像数据为系统解析细胞内部空间组织及其功能关联提供了可能。然而，“细胞组织”的多元性与复杂性对传统图像分析方法提出了挑战。空间蛋白质分布和多组分细胞内结构的形态分析仍缺乏统一、可解释的计算框架。在蛋白质分布分析中，诸如增殖细胞核抗原（PCNA）所形成的点状模式具有高度动态性，传统方法纹理特征虽可提取某些统计特性，但其生物学意义往往不明确，限制了结果的解释能力。另一方面，对于由多个离散组分构成的细胞器，如核仁或高尔基体，现有形状描述方法，如球谐函数展开通常仅适用于连续形状，难以有效表征其不连续、多形态的特性。即使通过图像分割提取出每个组分，如何整合这些片段以形成整体结构的有效表征，仍是一个开放性问题。

理解亚细胞组织结构的一个挑战在于如何以客观、稳健且可推广的方式，对具有复杂多组分形态的细胞内结构进行可解释的定量测量。传统方法往往依赖于人工设计的特征或局限于连续形状的分析，难以应对细胞内多态性、多组分形态的复杂性。针对这一问题，Vasan等人 [46]提出一种融合三维旋转不变自编码器与点云技术的表征学习框架，学习与方向无关、紧凑且可解释的复杂形状表征。将三维显微图

像中的生物结构编码为点云数据，进而通过旋转不变自编码器学习其低维、可解释的特征表示。该框架不仅适用于点状结构，还可推广至多态性结构，如核仁和高尔基体。

方向敏感性是许多深度学习模型在生物图像分析中的局限之一。尽管细胞方向在某些语境下具有生物学意义，如细胞迁移或发育过程，但在多数情况下方向差异仅反映样品制备或成像过程中的技术变异。利用三维旋转不变自编码器，通过几何深度学习机制，确保同一结构在不同旋转状态下被映射至同一特征向量的等价表示。不仅消除了方向偏差，还提高了特征的一致性及其在跨实验条件下的泛化能力。

基于三维点云与旋转不变自编码器的表征学习框架，为亚细胞结构的形态分析提供了客观可推广的计算方案。该框架不仅克服了传统方法在方向敏感性和形态复杂性方面的限制，还显著提升了特征的可解释性与生物学相关性，为细胞生物学中的大数据分析提供了新范式。

点云模型的输入数据预处理

cellPACK 合成数据集。 通过向每个输入点云添加少量抖动八次，将由cellPACK打包的球的$N = 256$ 个质心列表扩展至2,048 个点。扰动裁剪至0.2，XYZ 坐标的典型范围为-10 到10。随后将其用作3D 点云输入。为提高重建质量，对每个输入重复此增强过程十次。

DNA 复制焦点数据集。 应用与基于图像的模型中DNA 复制焦点数据集相同的预处理。随后，分两阶段从原始强度图像中采样4D 点云。在第一阶段，使用指数函数$e^{\lambda(\text{skewness}\times\text{intensity})}$ 缩放强度，将原始强度值转换为概率。此处，偏度(skewness)是表示分布与正态分布偏离程度的统计量，系数λ 是一个细胞内特异性比例因子，根据每个细胞内结构的随机图像采样点可视化结果凭经验确定。对于核结构，使用$\lambda = 100$，对于细胞质结构，使用$\lambda = 500$。此函数的目的是指数级增加从较高强度值采样点的概率，以防止从背景中采样。缩放后的强度随后通过除以总和进行归一化，转换为概率。从强度图像采样点云的完整概率定义为

$$P = \frac{e^{\lambda(\text{skewness}\times\text{intensity})}}{\sum e^{\lambda(\text{skewness}\times\text{intensity})}}.$$

使用这些概率采样一个包含20,480 个点，强度为编码器的输入。

点状结构的扩展数据集。 首先应用与上述基于图像的模型中点状结构扩展数据集相同的预处理，除最后一步线性缩放外。再次使用指数函数$e^{\lambda(\text{skewness}\times\text{intensity})}$，其中核结构使用$\lambda = 100$，细胞质点状结构使用$\lambda = 500$。缩放后的图像随后被归一化以获得概率密度。遵循DNA 复制焦点的相同过程，为每个点状结构采样点云。强度坐标随后使用结构特异性对比度范围进行归一化。

模型

为实现三维旋转不变的图像自编码器，采用基于R^3可操纵卷积核的图像编码器，

对SO(3)旋转群具有等变性。与普通卷积不同，R^3可操纵卷积核能在旋转下保持等变性质。利用标量场学习旋转不变的标量特征，并通过矢量场提取等变的矢量特征，进而重建三维旋转矩阵。网络包含七层可操纵卷积，最后一层输出N个标量场和两个矢量场。每卷积块后接入批归一化与ReLU激活，并在后五层引入平均池化，（经补充图1验证不影响等变性）。通过对最后一层标量嵌入进行空间池化，最终获得N 维旋转不变潜在嵌入：多态结构与点状结构分别使用瓶颈维度512 与256。

解码器部分采用六层卷积CNN结构，卷积层间插入尺度因子为2的上采样模块，参数配置为：(512,1,3), (256,1,3), (128,1,3), (64,1,3), (32,1,3) 及(16,1,3)。借助矢量表示推算旋转矩阵，以旋转规范重建结果，并使用圆柱体掩膜减少插值伪影。

引入了基于视觉Transformer（ViT）的掩膜自编码器作为替代方案，训练分为两阶段：首先以ZYX 块大小(2,2,2)、掩膜比率0.75及可学习位置嵌入进行预训练；编码器含八个Transformer 块（4头，256 维），解码器为两层（8头，192维）。第二阶段掩膜比率降为0，冻结编码器并训练新解码器进行图像重建。所有模型均以均方误差作为损失函数。

对于点云自编码器，采用基于矢量神经元（VNs）的等变编码结构，将标量神经元扩展为三维矢量以实现旋转等变性。该编码器嵌入动态图卷积神经网络（DGCNN）主干，通过EdgeConv模块在k近邻图（k=20）上提取局部特征，并逐步聚合为全局表征。除坐标信息外，部分输入还包含强度值构成4D点云，其变换与XYZ 坐标保持协同等变。网络包含六个卷积块，每块含VN线性层与VN Leaky ReLU层，最终对矢量嵌入取范数获得旋转不变表示。对照实验中，经典点云自编码器使用DGCNN编码器，其中VN层被替换为边缘卷积与ReLU激活。

3.6 小结：机器学习核心与生物医学图像处理应用

基于机器学习的生物医学图像处理中，从数据到模型的过程中，需多环节决策：

目标函数设计：构造经验风险项，结合任务特性，如分割连续性、形变合理性加入正则化项；

模型评估：通过交叉验证在有限数据下稳健评价性能，避免过拟合误判；

参数优化：通过格点搜索或经验准则选择超参数，平衡模型复杂度与泛化能力。在生物医学图像处理中，需注意数据稀缺性与领域知识的结合。考虑利用交叉验证充分利用少量标注样本，通过任务特定正则化，如对称性、平滑性融入生理先验，最终实现模型从“拟合数据”到“解决临床问题”的跨越。

第四章　深度学习

4.1　人工智能、机器学习与深度学习的关系

从技术层级来看，人工智能（AI）、机器学习（ML）与深度学习（DL）是包含与被包含的关系，机器学习是实现人工智能的核心手段，而深度学习是当前机器学习领域的主流技术分支。但这种层级关系并非“先后诞生的递进关系”，三者的起源时间相近，只是在不同阶段展现出不同的发展势能，最终形成今天的技术格局。

4.1.1　层级定位：从目标到手段

人工智能：是最顶层的目标，是让机器具备类人的智能行为，包括感知、推理、决策等。它并非特指某类技术，而是一个宽泛的领域，涵盖了实现“智能”的所有可能路径，如早期的符号逻辑、专家系统，以及现在的机器学习。

机器学习：是实现人工智能的关键路径之一。它的核心思路是“从数据中学习规律”，通过设计算法让机器从海量数据中自动拟合模型，再用模型进行预测或决策，无需人工手动编写规则。传统机器学习包含多种方法，如支持向量机、随机森林、逻辑回归等，而深度学习是其中以“深度神经网络”为核心的分支。

深度学习：是机器学习的子集，特指“使用深层神经网络，通常含3层以上隐藏层进行学习的技术”。它通过模拟人脑神经元的层级连接结构，用多层非线性变换从数据中提取复杂特征，如图像的边缘-纹理-语义，尤其擅长处理高维数据，如图像、文本、语音。

4.1.2　发展历程：同源而异流的技术演进

三者的起源时间相近，但发展节奏差异显著，这也间接塑造了它们的层级关系：**人工智能（1950s 起）**：1956 年达特茅斯会议正式提出“人工智能”概念，早期聚焦于“符号主义”，如用逻辑规则模拟推理，但因难以处理复杂现实问题陷入低谷。直到机器学习兴起，才为AI 提供了可靠的技术支撑，本质上，机器学习让AI 从“手动编码规则”转向“自动学习规则”，实现了从“空想”到“落地”的跨越。

机器学习（1950s-2010s）：几乎与AI 同步起源，随着计算机算力提升和数据积累，逐渐形成体系。20 世纪80-90 年代，支持向量机、决策树等传统机器学习方法

已能解决部分结构化数据问题，如简单分类、回归。但传统方法依赖人工设计特征，在高维非结构化数据，如医学图像上性能受限，这为深度学习的崛起埋下伏笔。

深度学习（2006 年-）：其核心“神经网络”的雏形虽在20 世纪50 年代已出现，如感知机，但因“梯度消失”“算力不足”等问题停滞多年。2006 年，Hinton在NIPS 发表论文提出“深度置信网络”，突破了浅层网络的限制；2012 年，AlexNet（卷积神经网络）在ImageNet 竞赛中准确率远超传统方法，其错误率降低10% 以上，让深度学习彻底走进大众视野。此后，GAN （2014 年）、Transformer （2017 年）等模型接连出现，如今大语言模型、医学图像分割模型的核心仍基于深度学习框架。

4.1.3 技术关联：从“工具”到“主流”

深度学习与机器学习的关系，本质是“技术迭代”与“领域拓展”的关系：传统机器学习是深度学习的基础：深度学习遵循机器学习的核心“经验风险最小化”，即通过优化损失函数从数据中学习模型参数。例如深度学习中常用的交叉熵损失，本质是最大似然估计的负对数形式；正则化、交叉验证等策略，在传统机器学习与深度学习中均通用，如UNet 训练中用L2 正则化约束卷积核参数，避免过拟合。

深度学习突破了传统机器学习的瓶颈：传统机器学习需人工设计特征，如用边缘检测算子提取医学图像的器官轮廓），而深度学习能“端到端”学习特征。以CNN为例，浅层卷积层自动提取图像边缘，深层卷积层整合出纹理或器官语义，无需人工干预。这种“自动特征学习”能力，让它在生物医学图像处理中表现突出：比如用Cellpose基于CNN分割细胞时，无需手动设计细胞形态特征，模型可直接从标注数据中学习分割规则。

深度学习是当前机器学习的“主流”：因其在高维数据上的优势，如今机器学习的前沿研究多围绕深度学习展开。例如生物医学领域，从CT 图像的病灶检测，用Faster R-CNN、三维器官分割到多模态数据融合，深度学习已成为解决复杂任务的“默认选择”，但这并不否定传统机器学习的价值，在小样本场景，如罕见病数据中，SVM 等传统方法仍因“对数据量需求低”而被优先使用。

4.1.4 生物医学图像中的关系具象

在生物医学图像处理任务中，三者的关系可通过具体场景直观体现：

人工智能的目标：让机器自动完成医学图像分析，如诊断病灶、预测预后，减少人工依赖。

机器学习的路径：通过“从标注数据中学习模型”。早期用传统机器学习，用SVM 对人工提取的病灶特征分类，如今用深度学习用CNN 直接从图像学习分类规则。深度学习的手段：作为当前最有效的机器学习方法，它通过特殊架构解决具体问题：比如用卷积层的“平移不变性”，无论病灶在图像哪个位置，都能被检测，提升检测鲁棒性；用U-Net 的“编码器- 解码器”结构保留图像分辨率，实现精准的

肿瘤分割。

三者的关系不是“替代”，而是“延伸”。人工智能是目标，机器学习是实现目标的“方法论”，深度学习是机器学习中当前最有效的“工具”。它们并非“新技术替代旧技术”，深度学习的崛起，本质是机器学习在“自动特征学习”上的突破，而机器学习的发展，又让人工智能从抽象目标落地为具体应用，如医学图像诊断。理解这种关系，能更清晰地选择技术路径：若要解决简单小样本任务，传统机器学习可能更高效；若面对复杂的医学图像等高维数据，深度学习则是更优选择，但最终，它们都服务于“用技术模拟智能”这一目标。

4.2 卷积神经网络

2012 年AlexNet 的成功让CNN 广为人知，它“能学习特征的层次结构”，尤其擅长处理网格数据，二维图像、三维图像乃至含时间序列的四维图像是典型的网格数据，因此CNN 被广泛用于图像分类、目标检测、图像分割等任务。

关于卷积神经网络的形态，我们可以做个有趣的尝试：若向大模型描述CNN 的基本结构，“CNN 是分层结构，包含输入层、输出层、激活层、卷积层，有卷积核与特定连接方式”，让它生成一张CNN 的图像，会得到一张既虚幻又颇具美感的图。这种“美感”体现在它与生物医学数据的真实分层结构高度相似，许多生物医学图像或解剖结构本就呈层级分布，一层叠一层；而“虚幻”则是因为图中虽包含输入输出、激活层、卷积层、池化层等元素，但它们的排布与真实深度神经网络的结构完全不同。

CNN 结构明确，直到今天，我们处理不同任务时用到的CNN，与它最初的形态并无本质差异。比如医学图像分割中常用的UNet、生物显微数据处理中Cellpose 的基准网络，本质都是CNN；即便与2012 年的AlexNet 相比，核心架构逻辑也一脉相承。

4.2.1 CNN 的基本结构

CNN 的基本结构包含：输入，若处理图像，输入即为图像数据；随后是多层叠加的卷积层。**卷积层**是CNN 的核心，其作用需结合“网格数据”特性理解：图像无论二维、三维还是含时间序列的四维图像本质是规则排布的网格数据（如长m× 宽n的图像含$m \times n$ 个像素），而卷积层的核心操作就是“卷积”：通过预设的卷积核在图像上滑窗，对每个局部区域做加权运算，最终生成特征图。若处理三维图像，生成的则是“特征体”。

卷积操作的本质是“特征提取”：特征图本质是输入图像像素或上一层特征图像素的加权组合，而提取的“特征”可包括图像的边缘、角点、纹理、形状等。这与传统视觉计算中“人工设计特征”的方式截然不同，比如用OpenCV 或ITK 提取边缘、角点时，需依赖预设算子；但CNN 能通过卷积层自动完成特征提取：浅层卷积

层捕捉边缘等基础特征，深层则整合出纹理甚至器官语义等高级特征。这种“自动特征学习”能力，正是CNN 在视觉计算中性能优于传统方法的核心原因之一。针对特定任务训练的CNN，提取的特征往往比人工设计的算子更优。

池化层的核心作用是“降低特征图分辨率”。以最大池化为例：若对上层特征图做2x2 池化，会将每个2x2 局部区域压缩为单个像素，使特征图的长、宽维度均缩减为原来的1/2。这既能简化计算，减少后续层的参数规模，又能保留关键特征。判断特征是否“关键”的标准可从自编码角度理解：若经卷积、池化后得到的低分辨率特征嵌入，仍能重建出原始图像，就说明未丢失核心信息。值得注意的是，特征图分辨率降低时，通道数通常会增加，因此特征的整体维度未必减小，如UNet 中，随网络深度增加，特征图分辨率降低但通道数提升。

全连接层则多用于分类任务：它将高维特征图flatten 为一维向量，经非线性激活函数处理后，通过Softmax 输出分类概率。因此常见的CNN 结构多是“级联卷积层-全连接层-Softmax”的组合，从12 年的AlexNet 到如今的各类CNN，这一核心架构始终未变。

4.2.2 CNN 的特征学习与关键特性

CNN 的“特征学习”能力可通过可视化工具直观呈现：观察网络不同卷积层的特征可知，浅层特征多是纹理等基础信息，深层特征虽分辨率降低，却包含更丰富的语义信息，如“器官”“病灶”等抽象概念。

这种特征学习能力依托于CNN 的两大特性：

平移不变性：若对输入图像x 做平移，如向右移动1 个像素，用同一2x2 卷积核卷积后，输出特征图也会做对应平移。这意味着只要学到合适的卷积核，无论目标在图像中哪个位置，都能被稳定检测。

参数共享：卷积核在图像平面上滑窗时，始终使用同一组参数，无需为每个位置单独设计参数。这既减少了模型参数量，又让模型能在不同位置学习到一致的特征，如“边缘”的特征在图像各处通用。

预训练模型的特征应用预训练大模型的特征提取能力更具代表性。以Stable Diffusion（SD）为例：将SD 预训练模型中第258 层的特征经PCA 降维可视化后会发现：两张不同摩托车图像的对应部位，特征伪彩色高度相似；对特征做聚类后，对应结构的聚类标签也基本一致。这说明无需额外训练，预训练模型就能提取图像中具有语义意义的结构特征。

这类特征可直接用于“语义对应”任务。在生物医学图像处理中，若已对一张图像的某结构做标记，通过特征匹配可建立它与未标记图像的语义对应，实现标签迁移。除SD外，基于Vision Transformer 的DINOv2 等模型提取的特征也能完成类似任务。这已成为视觉计算的新范式：不再局限于“从数据中学模型”，而是借助预训练模型/ 基础模型快速完成下游任务。

CNN 特征的应用场景还有很多：比如对显微图像做自监督学习，得到特征图

后，通过K-means 等聚类方法，可直接生成前景结构，如细胞核的分割掩膜，整个过程无需人工标注，模型仅通过浅层网络的梯度信息聚合就能实现。这也印证了深度神经网络的“表征学习”优势：它能精准捕捉图像中感兴趣的语义信息，如结构形状、纹理，且生成的表征比传统手工设计的特征描述子具有更强的表达能力。

4.2.3 神经网络中的归一化操作

归一化是神经网络的基础操作，其核心是对数据做标准化处理。在图像处理中，归一化的具体方式需结合数据维度设计，常见类型包括以下几种，其核心都是“减均值、除方差”，但计算范围不同。

批量归一化（Batch Normalization）

针对每个通道的所有数据计算均值与方差。若输入数据为批量大小N、通道数C、高H、宽W 的矩阵，维度为$C \times N \times H \times W$，批量归一化会对单个通道的所有$N \times H \times W$ 个数据计算均值μ 与方差σ^2，

$$\frac{lx - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

做标准化，ε 为小常数，通常取1×10^{-5}，用于避免方差过小时的数值不稳定。此外，还会通过可学习参数γ（缩放）与β（平移）调整归一化后的数据分布，增强模型灵活性。

批量归一化的优势明显：能加速训练收敛、降低对超参数的敏感度，甚至可作为正则化手段提升泛化能力，减少过拟合；但缺点是需额外计算均值与方差，且依赖批量大小。

层归一化（Layer Normalization）

针对单个样本的所有通道数据计算均值与方差。即对每个样本，将其所有C 个通道的$H \times W$ 个数据合并，计算整体均值与方差后标准化。它不依赖批量大小，但无法捕捉样本间的空间相关性，若批量数据存在空间关联，层归一化会忽略该信息。

实例归一化（Instance Normalization）

针对单个样本的单个通道计算均值与方差。即对每个样本的每个通道，单独计算该通道内$H \times W$ 个数据的均值与方差。这种方式常用于风格迁移任务，因它更关注图像内容而非风格，能有效分离内容与风格特征。

组归一化（Group Normalization）

将通道分组后，对每组通道数据计算均值与方差。例如将C 个通道分为若干组（如每组3 个通道），对每组内的所有数据计算均值与方差后标准化。它兼顾了批量归一化与实例归一化的优势，在小批量场景中表现更稳定。

四种归一化的核心差异在于“计算均值方差的数据范围”：批量归一化聚焦“单通道全批量”，层归一化聚焦“单样本全通道”，实例归一化聚焦“单样本单通道”，组归一化聚焦“单样本通道组”。实际应用中需根据任务场景，如批量大小、是否需保留样本关联，选择合适的方式。

Dropout

Dropout（随机失活）也是深度神经网络中常用的正则化方法，核心作用是避免模型过拟合。可以通过一个简单案例理解其原理：假设有一个含输入层、输出层和1个隐藏层的三层网络，若启用Dropout，会在训练时随机丢弃部分隐藏层神经元。

从实现逻辑来看，每个神经元被丢弃的概率为p，保留概率为$1-p$。被丢弃的神经元在训练过程中完全不参与前向传播与反向传播的计算，具体操作时，会为神经元生成一个符合伯努利分布的“掩码”，通过“按元素相乘”将需丢弃的神经元置零；随后对保留的神经元做缩放补偿，最终基于剩余神经元计算输出。

需注意的是，Dropout 仅在训练阶段随机丢弃神经元；测试阶段会保留所有神经元，无需缩放，这种“随机失活”机制能有效避免模型对训练数据特定特征的“记忆”：通过打破神经元间的协同依赖，减少“某些神经元仅依赖特定通道特征”的情况，迫使模型学习更通用的特征，从而提升泛化能力。因此，Dropout 是深度神经网络训练中极常见的正则化策略。

深度学习的三大支柱：算法、数据与算力

从2006 年自编码器的突破，到如今的大模型、基础模型与大语言模型，这一过程的推进离不开三大核心要素：**算法、数据与算力**。我们关注算法与模型，而“数据”正是深度学习落地的关键前提，尤其在视觉计算领域，深度神经网络的崛起便与大规模数据集的出现密切相关。

以卷积神经网络为例：2012 年AlexNet 在ImageNet 竞赛中的突破性表现，被视为计算机视觉“深度学习革命”的起点。ImageNet 数据集于2010 年前后问世，包含1400 万张带标签图像，标签由外包标注完成，涵盖1000 个类别，其中训练集120万张、验证集5 万张，其规模远超此前所有图像数据集。正是基于这样的大数据基础，AlexNet 的错误率（16%）远低于第二名（25%），这种性能差距在如今的研究中已极为罕见，也直接印证了“数据规模”对深度学习的重要性。

不过，大规模数据集也存在固有问题：一是**类别不均衡**，部分类别的样本量远高于其他类别，导致模型偏向性）；二是**数据质量问题**，考虑标签错误、图像噪声

等；更关键的是伦理与隐私风险，ImageNet 的图像多来自互联网爬取，而互联网数据实则受版权保护，不可随意使用。这一问题在生物医学图像处理中更为突出，需严格遵循科研伦理规范。

4.2.4 生物医学数据的伦理规范

生物医学领域的伦理要求可追溯至一系列国际准则：1947 年《纽伦堡法典》、1948 年《日内瓦宣言》（针对二战中纳粹医生的反人道罪行），1964 年世界医学协会发布的《赫尔辛基宣言》则进一步规范了所有临床操作，核心是“尊重涉及的人、动物及生物样本的权益”。生物医学图像处理相关论文需明确数据伦理声明，具体分为三类情况：

- 若仅为数值仿真，无需额外批准；
- 若使用公开数据集，如公开临床图像、生物样本数据，需声明“数据已通过原发布方伦理审查”；
- 若使用临床病人数据或动物生物样本，必须提供伦理委员会批准文件，如学校伦理委员会的审批函。

因此，无论是处理临床数据还是从互联网爬取数据，都需谨慎对待伦理与隐私问题。这是生物医学研究的基本准则。

4.2.5 CNN 的架构演进：从AlexNet 到ResNet

2012 年的 **AlexNet** 虽已是10 多年前的模型，但其核心架构与如今的CNN 并无本质差异：使用ReLU 激活函数、最大池化，借助GPU 实现并行计算，同时采用数据增强策略；网络含5 个卷积层与3 个全连接层，虽深度较浅，却奠定了CNN 的基础框架。

2014 年的 **VGG**进一步加深了网络深度，VGG16 含16 层、VGG19 含19 层），其核心改进是“用小卷积核（如3x3。堆叠替代大卷积核”，通过多层小卷积的叠加，既能减少参数规模，又能增强特征提取能力/逐层学习层级特征，这一设计至今仍被广泛采用。

2015 年**ResNet（残差网络）**的出现则彻底解决了“深层网络难训练”的问题。此前，随着网络深度增加，易出现梯度消失/ 爆炸，反向传播时梯度经多层传递后衰减至0 或激增，导致模型无法优化。ResNet 通过引入残差连接（短路连接）解决了这一问题：在网络层中增加“输入直接跳转至输出”的通路，使模型学习的是“输出与输入的残差”，而非直接学习输入到输出的映射，既降低了学习复杂度，又促进了信息流动，让网络深度得以大幅提升，如ResNet34/50/101/152，甚至更深。

此外，批量归一化，计算均值与方差并标准化，全局平均池化，减少参数以避免过拟合等策略，也进一步辅助了深层网络的训练。

4.2.6 残差连接在生物医学图像中的应用

残差连接的设计理念在生物医学图像处理中极为常见。例如常用的UNet 模型，其“编码器- 解码器”之间的长跳转连接本质就是残差连接的变体：将编码器某层的特征直接级联至解码器对应层，既保留了浅层细节特征，如边缘，又融合了深层语义特征，如器官结构，大幅提升了分割精度。

UNet 的应用场景广泛：例如预测图像形变场，二维图像需预测2 通道形变场，三维图像需3 通道）、实现多种医学图像的分割。而nnU-Net，自适应UNet则进一步实现了“模型自配置”：通过分析“数据指纹”，如灰度分布、图像模态，自动完成预处理，重采样、灰度归一化、网络拓扑选择，如2D/3D UNet、级联结构、超参数设置，优化器、学习率、数据增强策略、损失函数选择等。无需人工干预即可针对特定任务调整配置，大幅降低了实际应用门槛。

从AlexNet 到ResNet，再到UNet 与nnU-Net，CNN 的核心架构逻辑始终一脉相承，而“深度提升”“残差连接”“自适应配置”等改进，让深度学习在生物医学图像处理中愈发高效易用。

UNet在生物显微图像领域还有诸多其他应用，例如Cellpose。Cellpose 目前已更新至第三代，第二代基础架构正是UNet。如前文所述，UNet 包含较长的跳转连接，本质属残差连接的一种，即便在当下，它仍是处理生物显微图像的主流架构。Cellpose 的迭代始终围绕功能优化展开，对于第二代而言，最关键的改进是引入了人机闭环交互机制：若用户希望获得性能更优的模型，需提供更多标注数据。在固定细胞与活细胞场景，黄色区域为模型预测的细胞分割结果，紫色部分则是用户通过交互补充的标签。补充标签的原因在于模型对部分区域的预测效果欠佳。在Cellpose 2 中，若要精调模型，需根据其当前性能反馈，持续提供交互标注；只要用户不断补充标注，模型便可通过迭代调优逐步提升性能，这一过程可循环进行，直至人工补充的标签区域趋近于零，此时即认为获得了较理想的分割模型。

残差网络的特性与ODE 数值求解的关联

残差连接的引入能在训练过程中增强信息流动，从而促进模型训练与收敛。通常认为，模型深度的增加是提升性能的重要路径，通过堆叠更多残差块、引入更多参数，可优化从输入到输出的映射关系。但需注意的是，堆叠残差块会同步增加模型复杂度，这意味着训练时不仅需要更多存储空间，还需更长时间来优化大规模参数。回顾残差块的计算公式：第$n+1$个残差块的输出等于第n个残差块的输出与第n个残差块所预测的残差项之和。这一形式与常微分方程数值求解的单步迭代，如前向欧拉法完全一致：在ODE 数值求解中，从第n 步到第$n+1$步的迭代形式，与残差块的计算逻辑高度吻合。

基于此，我们可进行这样的对比：ResNet 中的每个残差块，相当于ODE 数值求解中的一次步进，从n 时刻到$n+1$时刻。这意味着，对于深层残差网络，无需依

赖大量残差块的堆叠，仅需借助ODE 数值求解器，通过在时间轴上进行多次划分与步进，即可近似实现多残差块堆叠的效果，从而大幅降低神经网络的复杂度。

在实际应用中，许多主流网络，DenseNet或用于计算光流场、位移场的网络，均通过堆叠网络块提升性能，这类网络往往存在大量残差连接。而借助ODE 神经网络，可将残差块堆叠转化为ODE 数值求解中的时间步划分，在模拟相同状态变化的同时，避免了残差块堆叠导致的复杂度提升。这一过程即“隐式神经网络”的核心逻辑：无需显式堆叠网络层，通过数值求解即可降低时间与空间复杂度。

ODE 的积分计算在医学图像处理中也有具体应用场景。例如在计算两幅图像，尤其是医学图像的形变时，需量化因生理、病理或治疗因素导致的器官形状变化。若仅计算像素级位移场，难以保障位移的可逆性，即便添加正则化项也难以完全解决；而通过“微分同胚”思想，可转而预测速度场，再通过速度场在时间上的积分得到像素级位移变化。这一积分过程恰好可通过ODE 数值求解实现，因此ODE 神经网络可用于形变场的求解。

Vision Transformer（ViT）

深度神经网络的核心目标之一是学习数据表征。无论是医学图像还是生物显微图像，均可通过网络提取有效特征。DINOv2 采用的是ViT 的特征提取。在深度神经网络的发展历程中，**Transformer** 是极具里程碑意义的模型。它是当前各类大模型的主干架构，无论大语言模型还是多模态大模型，均以其为核心。ViT作为Transformer 在视觉领域的应用。

对于图像特征提取，ViT 首先将图像分割为若干小块，转化为序列，即“token序列”，再通过Transformer 计算序列中各小块（token）间的关系。原本二维的图像经“扁平化”后，得以通过Transformer 提取特征，这种特征提取能力之所以强大，核心原因在于：图像扁平化后，Transformer 可建模所有局部小块间的关系即“空间相关性”或“长程相关性”，包括任意两个小块间的关联。

需特别关注ViT 中的“位置编码”：与CNN 不同，CNN 通过可学习卷积核对整幅图像操作，天然保留空间信息；而ViT 将图像扁平化的过程中，会丢失“小块在原始图像中位置”的信息。位置编码的作用即明确小块的位置、量化不同小块间的距离，且需保证编码的唯一性。由此引出一个关键问题：对于不同的图像处理任务，位置编码是否必需？答案需结合具体任务判断：

若任务涉及“**显著前景结构**”，如医学图像分割中对骨骼等特定器官的分割，图像扁平化后，结构的不同部分会分布在序列的不同位置，此时小块的原始位置信息具有重要意义。它可反映结构的部位特征，直接影响下游任务，如分类、分割的结果，因此位置编码必需。

若为“**非显著前景结构**”，例如显微图像任务，图像中每一小块可能对应独立的细胞、细胞核或细胞器，各小块相对独立，位置信息对任务无显著增益，此时位置编码并非必需。因此在使用ViT 时，需根据下游任务判断是否保留位置编码。这一

选择需结合任务对“空间位置信息”的依赖程度。

ViT的应用与生物医学图像处理的技术趋势

ViT 目前已成为视觉任务的主流架构，在大模型、多模态大模型及临床医学图像处理中均有广泛应用。例如在全身CT 分割任务中，ViT 架构可通过标注数据优化参数，实现精准分割；在超声数据处理中，它也能高效完成皮下多层解剖结构的分割。当前处理生物医学图像时，深度学习/深度神经网络是无可争议的主流技术，且这一领域正随大模型的发展发生显著变化。以往“拿到数据后从头训练模型”的模式已逐渐改变，基于大规模外部数据预训练的大模型，可通过“调优”直接应用于下游任务。

以Cellpose 2/3 为例，对预训练大模型调优时，仅需提供私有数据及补充标注，即可获得适配特定任务的模型。这也引发了实践中的关键思考：处理生物医学图像时，需明确数据准备方式、标注量需求，以及是否需从头训练模型。通常而言，利用基于外部数据预训练的视觉大模型，可大幅降低下游任务的训练成本。

在使用机器学习或深度神经网络进行图像配准时，通常采用端到端的计算范式：给定一对图像，训练好的模型会直接输出对应的变换参数，无论是刚性变换参数还是非刚性形变参数，均能一次性生成。这种计算方式常被称为“One - shot”，即单次计算即可得到目标参数，这也是当前机器学习与深度学习在该领域的主流模式。实际应用中也会出现“深度学习模型与传统数值优化结合”的情况：若深度学习模型的输出精度不足，对局部细节的刻画能力较弱，通常会在模型后处理环节引入迭代计算，通过传统数值优化进一步提升结果精度。从技术路径的象限分布来看，纵轴下方为传统机器学习方法，上方为深度学习方法，二者的核心差异对应着“小数据vs 大数据”“小模型vs 大模型”的选择。提到深度学习时，我们通常指向右上角象限：意味着依赖大规模数据，构建相对复杂的大模型，层数多、深度深的神经网络。

隐式神经网络通过迭代数值计算降低模型复杂度。以残差网络与ODE 神经网络的对比为例，后者可避免残差块的大量级联，显著降低网络复杂度。值得注意的是，四个象限的技术路径并未因深度学习的兴起而出现“淘汰”。在生物医学图像处理中，传统数值优化、预训练大模型的下游应用、隐式神经网络的复杂度优化、传统机器学习对小数据场景的适配等，均有其实际价值。只是当我们聚焦“深度学习”时，更多指向右上角的“大数据-大模型”路径。

4.3 时序数据处理：从循环神经网络到Transformer

卷积神经网络针对的是欧式数据，如生物医学图像，通常不涉及时序顺序；但实际中也存在有序数据，如含时间序列的信号、语音、文本等，这类数据早期多通过循环神经网络（RNN）处理。RNN 的核心是“循环”，其循环的对象是“隐式表

示”。因数据存在顺序，如时间序，t 时刻的隐式表示h_t 需依赖上一时刻的隐式表示h_{t-1}、当前输入x_t，以及模型参数，且时间步的循环可多次进行。

尽管随时间步推移，RNN 的计算过程看似复杂，但优势显著：它每次仅处理序列中的一个元素，却通过维护上一时刻的状态h_{t-1} 捕获了序列的顺序信息，最终经权重矩阵与偏移参数的映射输出结果。训练时需优化的参数包括“隐式状态与输出的权重矩阵”“输入与隐式状态的权重矩阵”“隐式状态更新的权重矩阵”等，通常通过时间反向传播（BPTT）算法实现参数更新。

RNN 的损失函数为各时间步损失的叠加，每个时间步均有输出，参数求导需依赖链式法则。而隐式状态的连续性会导致求导过程中出现“参数连乘”现象。这一特性使得RNN 在处理长序列时易出现梯度消失或梯度爆炸：当时间步过多，梯度可能衰减至零或激增，导致参数无法有效更新。

为解决这一问题，长短期记忆网络（LSTM）通过门控机制（输入门、遗忘门、输出门）避免长序列带来的梯度问题。其中，输入门控制需存入单元的信息，遗忘门决定从先前隐藏状态中丢弃的信息，输出门则筛选需传递至下一步的信息。相较于基础RNN重复模块为简单神经元，LSTM 的重复模块是含多门控的控制单元，更适配长序列场景。

4.3.1 Transformer：时序与视觉任务的主流架构

RNN 虽能处理时序数据，但当前语音、文本等任务更倾向于使用Transformer，其优势在于“可并行处理整个序列”，效率远超RNN 的串行模式。

Transformer 于2017年随论文《Attention Is All You Need》出现，如今已成为大模型的主干架构：无论是大语言模型、多模态大模型，还是各类问答系统，均以其为底层框架。其强大性能源于自注意机制，该机制能建模“长距离依赖关系”。以序列数据文本、语音为例，它可捕获序列中所有元素间的关联，不仅是相邻元素，这对语义理解至关重要：一句话中某个词的含义，可能同时受近邻词与远距离词的共同影响。

在生物医学图像处理中，长距离依赖同样关键：例如分割任务中，前景区域与图像其他部分可能存在“共生关系”，如人体结构的对称性、器官的附着关系，建模这类关联能显著提升下游任务精度。

自注意机制的核心逻辑可通过“查询（Query，Q）、键（Key，K）、值（Value，V）”的交互理解：模型生成输出时，Q 与K 通过内积计算输入序列中各元素的相关性，再结合V 得到最终的注意力输出，使得每个输出标记都与输入序列的所有部分相关联。

此外，Transformer 需通过位置编码补充序列的位置信息，自注意机制本身不包含位置信息：它为序列中每个元素添加唯一编码，与输入数据共同送入网络。位置编码的必要性需结合任务判断：若任务依赖空间位置，如医学图像中大型器官的分割，小块的位置信息能反映结构部位特征，此时位置编码必需；若处理显微图像如

细胞、细胞核等独立重复结构，位置不影响标签判断，此时位置编码并非必需。当前实践中，位置编码通常作为默认设置存在。

Transformer 的架构与优势

Transformer 的架构分为编码器与解码器两部分：编码器：包含多头自注意机制计算每个标记与序列所有标记的关联、前馈神经网络，以及残差连接与层归一化。残差连接可促进信息流动，加速模型训练；解码器：除前馈神经网络、残差连接与层归一化外，还包含“掩码多头自注意”，避免关注未来标签与“编码器- 解码器注意力”，关联编码器输出与解码器输入。如今Transformer 成为基准模型的核心原因在于：**并行化能力**：可同时处理整个序列，效率远超RNN 的串行计算；**长距离依赖建模**：通过自注意机制捕获序列中所有元素的关联，而非仅相邻元素；**可扩展性**：模型性能随规模增大而提升，适配大模型的复杂需求。

4.3.2 视觉Transformer（ViT）：图像的序列化处理

图像本身虽为静态数据并无时间序，但在视觉大模型中，可通过“分块- 序列化”转化为序列数据，从而套用Transformer 架构。

ViT 的预处理过程如下：给定大小为hxwxc 的图像,c 为通道数，灰度图c=1，彩色图c=3，高通量显微图像可能为5、8 等，将其分成pxp 的小块，总块数为$\Phi hxw\Psi/p^2$；随后将每个小块展平为向量，得到与时序数据形式一致的序列。若用于分类任务，ViT 的输出经分类层（权重矩阵+ 偏置）映射后即可得到最终结果。在生物医学图像领域，ViT 的应用已十分广泛：例如全身CT 的多类分割任务，输出118 个通道对应117 个结构，需同时分割人体各类关键结构；又如超声图像中皮下分层结构包括脂肪、肌肉、骨骼等的分割，均能通过ViT 实现高精度处理。

需注意的是，自注意机制的计算复杂度与序列长度呈二次相关，序列过长时计算成本极高。当前已有多种优化方案，例如“稀疏Transformer”通过分块限制自注意的计算范围，仅关注块内元素，或通过“加法替代内积”降低关联计算复杂度。

4.4 生成模型

4.4.1 生成对抗网络（GAN）

在生成对抗网络（GAN）的衍生模型中，WGAN（Wasserstein GAN）是解决传统GAN 训练不稳定问题的关键突破，其改进在于距离度量方式的优化。在WGAN 中，度量生成数据与真实数据分布差异时，采用了推土机距离（Wasserstein Distance，也称为Earth-Mover Distance），替代了原始GAN 中使用的JS 散度（Jensen-Shannon Divergence）。这种替换带来了显著优势：即便面对极端情况，即生成数据与真实数据的分布完全无关时，模型也不会出现传统GAN 中常见的梯度消失问题。

尽管“分布完全无关”属于非常极端的场景，但凸显了推土机距离的意义：它能为GAN 的训练稳定性提供明确且可靠的保障，避免模型因梯度信息缺失而陷入训练停滞。

案例：WGAN 在颌面破损图像修复中的应用

以“颌面破损CT 图像自动修复”为例：假设存在一张反映人体上颌破损的CT图像，模型的目标是将其修复为该个体在无生理异常情况下的正常颌面形态。WGAN 的“对抗机制”发挥了关键作用。这里的“对抗”，本质是让模型生成的“特定患者健康颌面图像”，与该患者真实的健康颌面图像或同类健康个体的标准颌面图像尽可能一致，最终达到让判别器无法区分两者的效果：既无法判断某张健康颌面CT 是真实采集的，还是由算法生成的。

利用算法合成大量不同形态的颌面破损图像，构建训练数据集；用这些合成数据训练WGAN 的生成器，使其具备修复破损图像的能力；通过判别器的“对抗监督”，确保生成器修复出的图像与真实健康颌面图像在分布上完全对齐，即判别器无法分辨；最终，通过这一目标驱动，实现颌面破损图像的自动化、高保真修复。

GAN 的增强模型：StyleGAN 及其核心特性

除WGAN 外，GAN 家族还有诸多针对特定需求优化的增强模型，StyleGAN 是“风格生成”领域代表性的模型。GAN 的生成任务场景十分广泛，除了“图像修复”，还包括“图像域翻译”，如将素描图转化为彩色照片、将白天场景转化为夜景等。而StyleGAN 的独特价值，体现在其对“生成图像属性控制”的精准性上。例如，它可以基于同一张患者照片，生成该患者“更健康”或“病情更严重”的外观图像。这类生成任务的应用场景也很明确：既可以用于患者护理指导、疾病相关的社会认知科普，也能辅助医生向患者直观展示病情发展或愈后效果。

StyleGAN 与传统GAN 的核心差异，在于其生成方法的优化：传统GAN 的生成器通常从隐空间中随机采样一个向量z，直接基于z 生成图像；而StyleGAN 会在z的基础上引入额外的向量w，其中w的不同通道对应图像的不同属性，如面部的肤色、纹理、五官比例等“风格特征”。通过对w的精准调控，StyleGAN 能够生成风格更丰富、属性更可控的图像，极大提升了生成结果的实用性和灵活性。

4.4.2 扩散模型

除了GAN 系列模型，当前生成式AI 领域另一种核心技术是扩散模型。相较于GAN、VAE等其他生成模型，扩散模型的最大优势在于其坚实的理论基础，它的核心逻辑直接对应物理学中的“扩散过程”。通过在训练阶段模拟“向真实数据中逐步添加噪声，使其最终变成随机噪声”的过程，再在生成阶段反向学习“从随机噪声中逐步去除噪声，还原出真实数据分布”的路径。

这种与物理过程紧密结合的设计，不仅让扩散模型的生成过程可解释性更强，也使其在图像、文本、音频等多个领域的生成质量上达到了当前业界领先水平。在模型设计与调参中，我们常会频繁提到“凭经验设置参数”“凭经验设计模块”，但“经验”本身是模糊且难以量化的。而扩散模型的核心优势恰恰在于其设计有明确的理论基础：无论是“为何能从高斯噪声生成数据”，还是“为何需对原始数据逐步加噪至高斯噪声”，都能找到理论依据，无需依赖主观经验判断。

扩散模型：“破坏- 重建”逻辑与核心特性

扩散模型的生成逻辑常被概括为“破坏- 重建”，这一过程可拆解为两个核心阶段：**破坏（正向扩散加噪）**：针对目标数据图像，通过多步迭代逐步添加噪声，最终将其“破坏”为完全随机的高斯噪声。以图像为例，直观上就是从一张清晰的正常图像，逐步变成一张无意义的噪声图。

重建（反向扩散去噪）：与正向过程相反，从高斯噪声或含噪声的“问题数据”出发，通过迭代去噪逐步还原数据的真实分布，实现“重建”。“去噪”并非仅指从纯噪声中恢复图像，其场景更广泛：若输入是破损图像，如颌面缺损CT，去噪/ 重建就是还原为完整图像；若输入是病理异常图像，如含病灶的医学影像，去噪/ 重建则是生成对应的健康状态图像，即消除病理异常结构。

扩散模型的性能与应用

在当前主流生成模型中，扩散模型的综合能力被广泛认为是最强的，其生成质量通常优于GAN 和VAE，核心原因在于其基于物理扩散过程的理论严谨性，避免了GAN 的训练不稳定，如梯度消失和VAE 的生成模糊等问题。扩散模型的应用覆盖多类生成任务，尤其适用于医学图像处理：

低代价数据生成：医学图像采集需患者授权、专业设备支持，成本极高。通过训练好的扩散模型，可生成符合真实分布的医学图像，如超声图像，且能融入领域特性，如模拟超声探头“近清远糊”的物理特性；

文本驱动的定向生成：属于多模态生成任务，通过输入文本描述，如“右肺正常、左肺存在结节的胸片”，可生成符合描述的定向医学图像，用于模型训练或教学演示；

无监督异常检测：以脑CT 异常检测为例，扩散模型可从“异常脑CT”出发，生成该患者对应的“正常脑CT”；通过计算“异常输入”与“正常生成结果”的残差图，即可定位异常区域，无需标注异常样本；

图像超分辨率：针对低分辨率医学图像，如设备限制导致的低分数据，以“低分图像”为条件，生成高分辨率图像并补充细节。此时扩散模型的“去噪”可分为两种路径：直接对图像像素进行迭代去噪，如DDPM，去噪扩散概率模型；先将低分图像编码为隐式表示，对隐式表示进行去噪后，再通过解码器生成高分图像，后者更适用于高分辨率图像生成，可降低计算成本。

扩散模型的局限

计算效率低：反向去噪需数百至数千步迭代，每一步均需模型计算，生成过程耗时较长；

内存需求高：训练去噪模型时，处理高分辨率图像，如高清医学影像易出现内存不足问题，需依赖隐式表示，如LDM或模型压缩技术缓解。

4.4.3 VAE 及其变体：从隐空间分布到量化表征

VAE（变分自编码器）是另一类经典生成模型，核心是“从隐空间采样生成数据”，与GAN 的“随机采样隐向量z 生成”有相似之处，但设计思路更侧重“分布学习”。VAE 的架构由“编码器- 隐空间- 解码器”组成，形成闭环：

编码器：将输入数据x 映射到隐空间，输出隐分布的均值（μ）和方差（σ^2）（通常假设隐分布为高斯分布）；隐空间（潜在空间）：从编码器输出的高斯分布中采样得到隐向量z，每个z 对应数据分布中的一个样本；

解码器：将隐向量z 解码，生成与输入x 尽可能一致的重构结果$\hat{x}$。

与传统自编码器的差异

2006 年提出的传统自编码器仅通过“编码- 解码”让输出与输入一致，本质是将数据投影到低维空间；而VAE 的“变分”体现在：它不直接输出隐向量，而是输出隐分布的均值、方差参数，通过学习“符合高斯先验的隐分布”，让隐空间更具泛化性，可通过采样生成新数据，而非仅重构输入。

损失函数：重建损失+ KL 散度正则

VAE 的训练目标包含两部分，确保“重构准确”与“隐分布符合先验”：

重建损失：衡量生成结果$\hat{x}$与输入x 的差异，确保模型具备重构能力；

KL 散度： 衡量编码器输出的“隐分布”与均值为0、方差为1 的“标准高斯先验”的差异，起到正则化作用，避免隐空间坍塌。

掩码自编码器

MAE 是VAE 的重要变体，核心改进是引入掩码机制：训练时，对输入图像随机掩盖部分区域，要求模型仅利用未掩盖区域，重构出完整的原始图像；通过“不完整数据重构完整数据”，强迫模型学习更鲁棒的图像表征，需捕捉全局结构而非局部像素；可用于医学图像修复，例如如颌面缺损、器官影像破损修复；强表征学习，如心脏影像、电镜图像的特征嵌入，甚至结合ViT提升表征能力。

VQ-VAE（向量量化变分自编码器）

VQ-VAE简化表征的离散化方案VQ-VAE 针对VAE“隐空间为连续分布、计算成本高”的问题，引入向量量化机制：

量化逻辑：预设一个“码表”，包含多个离散的“码向量”；编码器输出连续的隐表示后，将每个隐向量与码表中的码向量对比，分配最相似的码向量的标签，例如“与第3 个码向量最接近，则标签为3”；量化后的隐表示从“多通道连续张量”变为“单通道离散标签图”，极大简化表征形式。

量化的利弊：

弊端：量化过程是“有损”的连续分布→离散标签；

优势：离散表征可大幅降低后续任务的计算复杂度，尤其适用于跨维度映射任务。

应用实例：2D-X 光重建3D-CT：分别为2D-X 光和3D-CT 构建VQ-VAE 自编码器，得到两者的“单通道离散表征”；由于表征已简化，只需建立“2D 离散表征→3D 离散表征”的低复杂度映射，即可实现从2D-X 光到3D-CT 的重建；还可构建“3D-CT 离散表征→2D-X 光离散表征”的映射，实现3D 影像的体绘制，如调整透明度观察骨骼、软组织的二维投影。

4.5 图神经网络：从欧式数据到图表示的扩展

此前讨论的图像，2D 矩阵、3D 张量均属于欧式数据结构规则，但实际场景中还存在大量非欧式数据，其结构不规则，需用“图”表示，图神经网络（GNN）正是处理这类数据的核心工具。

图由节点和边组成：节点：代表数据的基本单元，如社交网络中的人、分子中的原子、医学影像中的器官区域；边：代表节点间的关联，如社交关系、原子间的化学键，可带权重表示关联强度或无权重仅表示“有无关联”。

图的结构可通过矩阵量化，常见形式包括：**邻接矩阵**：对称方阵，行/ 列均对应节点，元素值为“1”表示两节点有边连接，“0”表示无连接；**相似矩阵**：元素值表示节点间的相似性，如特征相似度，是邻接矩阵的扩展；**图拉普拉斯矩阵**：由邻接矩阵和度矩阵推导而来，用于捕捉图的全局结构信息，是图卷积的核心矩阵。

4.5.1 图卷积网络的退化问题

CNN通过“滑动卷积核”聚合固定邻域的像素特征，GNN则通过“图卷积”聚合节点的**关联邻居特征**，两者逻辑一致但实现不同：

- **聚合对象**：以节点为中心，聚合其“一阶近邻”直接相连的节点或“多阶近邻”；收集每个节点的邻居节点特征，通过加权求和、平均等方式聚合；

- **聚合方式**：通过可学习的权重矩阵控制邻居特征的贡献度，常见方式包括：求和聚合：直接累加邻居特征；平均聚合：对邻居特征取平均，降低异常值影响；最大聚合：保留邻居特征中的最大值，突出关键信息。

图卷积网络核心操作包括邻域聚合与消息传递GNN的核心是“利用节点邻居信息更新节点特征”，分为两步：**邻域聚合**与**消息传递**。将聚合后的邻居信息与自身特征融合，更新节点的最终特征，通常结合非线性激活函数增强表达能力。

4.5.2 图神经网络：数据表征新范式及医学应用

在数据表征中，“张量”是我们熟悉的规则形式，例如图像的2D矩阵、CT的3D张量，图是更灵活的表征。两者的核心差异在于“近邻定义”：张量数据的近邻是固定的，二维图像有上下左右4 近邻、三维图像有6 近邻，若考虑$3 \times 3/3^2$邻域块则数量固定；而图数据的近邻由“边”决定，需通过图矩阵邻接矩阵、相似矩阵或图拉普拉斯矩阵记录节点间的连接关系，结构更灵活。在**蛋白质结构生成**中，蛋白质的三维结构可表示为图，原子为节点，化学键为边，通过图神经网络，如AlphaFold中的GNN模块可预测蛋白质的功能结构。

图表示：为什么要在图像处理中引入图？

图表示的优势在于**降低数据规模**，传统张量表征的痛点是**规模过大**：一张手机照片的像素可达上百万，三维CT的体素甚至上亿，若维度进一步升高，如4D动态影像，数据量会呈指数级增长。而图能通过“聚合相似单元”大幅降维：

- **聚合规则**：采用“超像素分解”或“超体素分解”，将相邻且表观一致的像素/体素聚合为一个**节点**，再用“边”连接空间相邻的节点块；
- **规模可控**：聚合后图的节点数可灵活设定，如将百万像素的图像聚合成5000 个节点，相较于原始张量规模显著降低；
- **不影响下游任务**：即使规模缩小，图仍能保留数据的全局结构与局部关联，可正常支持图像检测、分割、场景理解等下游任务。

图的应用场景扩展

在医学图像处理中，图的典型形式是“超像素/超体素图”，但图的价值不止于此：若需保留局部细节，可缩小聚合粒度，增加节点数；若需降低计算成本，可放大粒度,减少节点数，兼顾“细节保留”与“规模控制”；在其他领域，如社交网络、分子结构，图的规模可能更大，如百万级节点的社交图，因此“大规模图的高效处理”是图神经网络的核心研究方向之一。图与隐式神经网络突破了传统张量表征的局限：图通过“灵活聚合”降低规模，适配不规则数据。

（1）退化问题：节点特征过平滑 当GNN层数加深时，会出现**特征退化**现象，本质是“图矩阵连乘导致的特征坍塌”若对GNN 做简化假设，无非线性激活函数、权重矩阵为单位阵、不优化权重参数，得到的“线性图卷积网络”会出现退化现象：

- **退化原理**：若简化GNN，不考虑非线性激活、权重矩阵设为单位阵，多层图卷积等价于“图矩阵，如相似矩阵的多阶连乘”。根据线性代数原理，矩阵连乘后会逐步趋近于**秩1矩阵**，所有节点的特征向量都会退化为图矩阵的“占优特征向量”对应最大特征值的特征向量；

- **直观表现**：节点特征“过于平滑”，每个节点的特征与邻居高度一致，丢失了个体差异，无法区分不同节点的属性，类似图像平滑抹除高频细节。

针对该问题，借鉴数值计算中的“幂迭代收缩法”，在GNN层递进时逐步“消去占优特征向量的影响”，从而提取图矩阵的其他特征向量，保留节点的差异化信息。

（2）动态图与注意力机制 传统GNN的图结构是固定的，难以处理动态变化的图，如分子动态运动、动态社交关系，且默认所有邻居的贡献度相同，忽略了“邻居重要性差异”。引入**注意力机制**可解决这一问题：

计算节点与邻居的“相关性系数α”，通过归一化得到，α 越大表示该邻居对当前节点的特征更新越重要；动态调整邻居的贡献权重，让GNN更关注关键邻居，提升特征表征的针对性。

（3）大规模图的处理：采样+聚合 当图的节点数极大，如10万级以上时，直接聚合所有邻居会导致计算量爆炸。无论原图结构多复杂，都能通过采样控制每一层的计算规模，实现大规模图的高效处理。

GraphSAGE分为两步：

- **采样阶段**：对每个节点，随机采样固定数量的一阶近邻，如从10 个邻居中采5个，避免邻居过多导致的计算过载；

- **聚合阶段**：从外层采样的邻居开始，逐步向内层节点聚合特征，更新当前节点的表征。

4.6 隐式神经网络：神经辐射场

隐式神经网络突破传统张量表征的范式，其核心是“用网络计算数据属性，而非存储固定格点值”。神经辐射场是典型代表，尤其适用于三维重建任务。

4.6.1 隐式vs. 显式表征

- **显式表征（张量）**：直接存储数据在固定格点上的取值，如3D CT的每个体素值，优点是直观，缺点是维度升高时数据量剧增；

- **隐式表征（NeRF）**：用一个轻量级多层感知机替代固定格点存储，输入“三维坐标(x, y, z)”和“观测方向”，MLP输出该位置的属性，如医学图像的组织密度、自然图像的RGB 颜色+ 透明度；

为什么叫“隐式”？因为数据的属性不直接存储，而是通过网络“实时计算”得到，无需预先定义所有格点的取值。

4.6.2 NeRF在医学三维重建中的应用

传统NeRF针对“特定场景”，如特定病人的X光片重建，若要构建“泛化模型”，适用于任意病人，需结合生成模型如扩散模型，以“病人的二维X光片”为条件，通过扩散模型指导NeRF的MLP 学习“不同病人的三维CT 分布”，实现“输入任意二维X光，生成对应病人的三维CT”。NeRF的核心能力是“从少量二维图像重建三维场景/物体”，可用于三维医学图像重建。以“从二维X光片重建三维CT”为例，NeRF的训练与推理流程如下：

- **输入数据**：一组正交的二维胸片，无需真实三维CT作为监督；

- **训练目标**：通过优化MLP，让其生成的“三维隐式表征”投影到二维平面后，与输入的X 光片完全一致，进行自监督训练；

- **体绘制过程**：为了验证投影一致性，需模拟医学影像的“体绘制”，给不同组织，如骨骼、肌肉，设置不同透明度，感兴趣结构设为低透明度以显影，无关结构设为高透明度以隐藏，将三维隐式表征转化为二维投影图，与输入X光片对比优化；

- **推理阶段**：训练完成后，输入任意三维坐标，MLP即可输出该位置的组织属性，生成完整的三维CT图像。

传统CT扫描会产生电离辐射，若患者治疗过程中的疗效监测需多次采集，累积辐射剂量可能超出临床安全范围；NeRF仅需少量二维X光片，辐射剂量远低于CT，即可重建出高质量三维CT，在保证诊断需求的同时，大幅降低辐射风险。NeRF可扩展至4D动态重建，如心脏跳动的动态CT，通过输入不同时间点的二维影像，生成动态的三维隐式表征，捕捉组织的运动规律。NeRF 通过“实时计算”实现高效三维重建，兼顾精度与临床安全性。

4.7 3D高斯溅射

在三维表征领域，3D高斯溅射（3D Gaussian Splatting, 3DGS） 与神经辐射场的差异在于“表征形式”：

- NeRF通过隐式的MLP 计算三维坐标的属性，如密度、颜色，不直接存储数据；
- 3DGS则采用显式的离散高斯粒子表征三维场景，在目标空间中放置一系列“三维高斯函数”，每个高斯函数由均值/三维坐标、方差/空间分布范围、权重/贡献度 定义，通过这些参数可直接计算任意位置的属性值,如医学图像的组织密度、自然场景的颜色。

3DGS的优势

- **直观的场景划分与属性控制**：3DGS将三维场景拆分为独立的“高斯粒子块”，每个块可单独赋予材质属性，如反光率、透明度 和运动属性，如速度、旋转参数。这种拆分让场景的精细化控制更易实现，尤其适用于**物理仿真任务**：
  例如，模拟人体器官运动时，可给心脏、肺部的高斯粒子分别设置不同的运动规律，如心脏的收缩频率、肺部的呼吸扩张幅度，无需像传统张量表征那样处理整体动态，操作性更强。
- **高效的二维投影绘制**：与NeRF类似，3DGS需将三维高斯粒子投影为二维图像，核心是“累积求和”：
  对观测方向上所有经过的三维高斯粒子，按其权重加权求和，得到二维投影面上每个像素的属性值。相较于NeRF依赖MLP实时计算，3DGS的离散高斯粒子聚合更易并行优化，绘制效率更高。

4.8 视觉大模型：重塑图像处理范式

近年来“大模型终结计算机视觉”的讨论，本质是**视觉大模型** 突破了传统图像处理的局限，形成了“预训练-迁移”的新范式。其强大能力源于两大核心：大规模参数规模大：通过亿级甚至千亿级参数刻画复杂数据分布，可建模图像的细粒度语义关联；海量数据训练：训练数据覆盖自然场景、医学影像、工业图像等，总量远超人类个体一生的视觉经验，能学习到跨领域的通用特征。

4.8.1 大模型的核心能力

超越传统的特征表征

传统图像处理依赖“手工设计特征”，如边缘检测、角点提取或“任务特定CNN特征”，这类特征缺乏语义信息；而视觉大模型的预训练特征具备**强语义性**：

- 无需额外标注，模型能自动识别不同图像中的相同结构，如如不同胸片中的“肺叶”“肋骨”，在特征空间中呈现相同的聚类模式；
- 直接从预训练模型的中间层提取多通道特征，即可用于分割、检测等下游任务，甚至无需训练专属解码器。

零/少样本的任务适配

在生物医学图像处理中，大模型的**迁移学习能力**尤为关键，多数临床任务数据稀缺，如罕见病影像，但通过“预训练大模型+少量任务数据微调”，即可快速构建高性能模型：

- CLIP（Contrastive Language-Image Pre-training）通过“图文对齐预训练”，可实现**无监督分割**：将医学图像输入CLIP得到视觉表征，与“正常肝脏”“肿瘤”等文本表征比对，即可定位对应区域，无需标注分割掩码；
- DINOv2（Self-Supervised Visual Transformer）通过“自监督对比学习+ 自蒸馏”（教师模型引导学生模型学习一致表征），仅用无标注数据就能学到通用视觉特征，微调后可用于病理切片的细胞分类。

端到端的复杂任务处理

传统视觉任务需分步骤优化，如“先检测目标再分割”，而大模型可直接通过**提示（Prompt）驱动**完成端到端处理：

- 输入自然图像，大模型能同时输出“个体分割，区分不同人/物体”“深度估计判断物体远近”“场景分类识别室内/室外”；
- 在医学场景中，输入“右肺结节的CT影像”提示，大模型可直接生成“结节大小测量”“良恶性概率预测”“三维重建结果”，无需手动调用多个工具。

4.8.2 大模型的局限性与适用场景

大模型并非“万能”：若任务数据与预训练数据差异极大，如特殊设备采集的微观生物影像，其性能可能不如专用小模型；但对于多数常规医学影像，如CT、MRI、超声，大模型的预训练特征已能覆盖核心结构，是当前效率最高的解决方案。

4.9 实例：基于生成模型的虚拟染色

荧光显微镜在生物和医学图像分析中对于监测亚细胞结构的形态和动力学至关重要 [5, 36, 40, 44]。然而，荧光染色昂贵、耗时，并且存在光毒性和光漂白的风险，特别是在活细胞中 [19, 39]。虚拟染色先河的计算机辅助着色（in silico painting）技术，利用像素到像素的转换，将无标记的透射光显微镜图像转换为细胞器特异性的荧光图像 [10, 34]，提供了非侵入性、非破坏性的成像方式，使得分析亚细胞结构的形状、功能和生理特性成为可能 [2, 7, 23, 37]。与传统染色相比，虚拟染色提高了采集速度和多路复用能力，支持细胞器检测和配准等下游任务，这对于生成亚细胞结构的统计模型至关重要 [4, 47]。深度学习技术如CNNs [43]、U-Nets [20]、条件GANs [12, 38]以及transformers [48]，已被应用于无标记虚拟染色。任务感知先验 [57]和稀疏视图方案已被探索用于3D亚细胞结构预测 [23]。Wieslander 等人 [49]采用密集U-Nets和GANs进行细胞绘画，将并行虚拟染色与细胞核分割相结合。鉴于亚细胞结构形态和动力学的多样性，需要强大的图像生成器来在输出的荧光图像中突出显示亚细胞结构。

扩散模型因其通过迭代去噪和灵活的分布建模性能，在图像生成领域日益突出。与可能遭遇模式崩溃和训练不稳定的GANs不同，扩散模型更易于训练，并且能够从高斯噪声开始，以概率方式生成高质量、多样化的图像。然而，在扩散模型中编辑风格码以实现细粒度结构重建仍然具有挑战性。条件扩散模型已被应用于荧光图像生成 [11, 32]，其中模型最小化与输入的距离 [8] 或从带噪输入图像中去噪 [30]。类引导的去噪扩散概率模型已被用于荧光图像重建，通过精心准备的类先验指导去噪过程 [11]。然而逐像素图像生成通常会修改整个图像以与目标域的分布对齐，这可能导致局部结构细节的丢失。

DiffStain从明场图像生成亚细胞结构特异性荧光图像的新颖框架。DiffStain采用条件扩散方法，其中亚细胞结构掩膜指导迭代去噪过程。不依赖于预先选择的图像滤波器进行亚细胞结构分割，而是提出了一个深度神经谱聚类模块来从荧光图像中提取掩膜。通过利用预训练的DINOViT特征 [33]并在谱嵌入空间中进行k-means 聚类，无监督NSC模型能够有效地从带噪或模糊的荧光图像中识别亚细胞结构。为了增强荧光图像生成过程，在在线推理过程中加入了掩膜引导。NSC生成的掩膜被反馈给去噪器，确保迭代去噪过程能够突出感兴趣的亚细胞结构。我们通过在公共显微镜数据集上的大量实验评估了所提出的DiffStain的有效性。

虚拟染色条件扩散模型

扩散模型通过逐步对随机高斯噪声去噪以生成匹配目标分布的图像，从而实现多样化图像生成。采用条件扩散模型进行从明场图像x到多通道荧光图像y的图像到图像翻译，通过以$p(y|x)$ 的形式对输入图像去噪。扩散模型包含两个主要过程：前向扩散过程和反向去噪过程。在前向过程中，高斯噪声以马尔可夫方

式迭代添加到图像中。前向过程定义为：$q(y_t|y_{t-1}) := \mathcal{N}(y_t; \sqrt{1-\beta_t}y_{t-1}, \beta_t\mathbf{I})$，其中$\beta_t$表示第$t$ 步的噪声方差。时间步t的图像y_t 以原始图像y_0为条件，其转移定义为：$q(y_t|y_0) = \mathcal{N}(y_t; \sqrt{\alpha_t}y_0, (1-\alpha_t)\mathbf{I})$，其中$\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$表示噪声随时间的累积效应。

在反向去噪过程中，模型旨在通过训练神经网络ϵ_θ预测每一步的噪声来重建原始图像。反向过程由条件概率分布描述：$p_\theta(y_{t-1}|y_t) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, x, t), \beta_t\mathbf{I})$，其中$\mu_\theta(y_t, x, t)$是分布的可学习均值。使用朗之万动力学估计数据对数似然的梯度，迭代反向去噪过程可写为：

$$y_{t-1} = \frac{1}{\sqrt{\beta_t}}\left(y_t - \frac{1-\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x, y_t, \alpha_t)\right) + \sqrt{1-\beta_t}\epsilon_t. \tag{4.1}$$

神经网络ϵ_θ通过最小化每个时间步实际噪声与预测噪声之间的距离进行训练，损失函数$\mathcal{L}_{dm} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,(x,y),\alpha}\left\|\epsilon_\theta^{(t)}(x, y_t, \alpha) - \epsilon\right\|_1$。该损失函数对应于关于加权变分下界的最大化似然。通过最小化损失函数，$\mu_\theta(y_t, x, t)$和$\epsilon_\theta^{(t)}(x, y_t, \alpha)$ 均被优化以执行图像翻译并生成高保真荧光图像。

神经谱聚类

NSC在谱嵌入空间中对荧光图像进行亚细胞结构分割的无监督方法。与传统的交互式图像滤波器和操作符不同，NSC利用预训练的DINOViT特征[15]和谱聚类来对抗识别细粒度结构中的高频扰动。给定荧光图像$y \in \mathbb{R}^{m\times n\times l}$，估计亚细胞结构掩膜$z \in \{0,1\}^{m\times n\times l}$，其中$l$ 代表对应不同亚细胞结构的通道数。

首先，使用预训练的DINOViT特征从荧光图像构建图像块图，这些特征通过自注意力捕获重复性细粒度结构间的长程关系。考虑到亚细胞结构的细粒度，图像被细分为小视场（FOV）图像$y_s \in \mathbb{R}^{q\times q}$，其中图像块大小与亚细胞结构相当。

接下来对图像块图的拉普拉斯矩阵进行特征分解。采用联合谱嵌入以避免谱失真和小FOV图像间不一致的聚类分配。这是通过使用锚点图像y_a的信息增广亲和矩阵，确保跨图像一致的亚细胞结构识别。增广亲和矩阵定义为：$\tilde{A} = \begin{bmatrix} A_a & A_{a-s} \\ A_{a-s}^T & A_s \end{bmatrix}$，同时考虑了图像间和图像内的块间关系。$A_a$和$A_s$ 分别表示锚点图像和小FOV图像的亲和矩阵。A_{a-s}表示y_a和y_s 间的块间亲和度。增广亲和矩阵$\tilde{A} \in \mathbb{R}^{2n_p\times 2n_p}$使用余弦相似度计算，$\tilde{A}_{ij} = \frac{F_iF_j^T}{\|F_i\|_2\|F_j\|_2} \odot (F_iF_j^T > 0)$，其中$F_i$ 和F_j 表示图像块i 和j的DINOViT特征。n_p表示小FOV 图像的图像块数量。归一化拉普拉斯矩阵$L = I - D^{-1/2}\tilde{A}D^{-1/2}$（其中$D$是度矩阵，$D_{ii} = \sum_j \tilde{A}_{ij}$）用于谱聚类。对应前$r$个非零特征值的特征向量$\Psi \in \mathbb{R}^{2n_p\times r}$提供了荧光图像的谱嵌入。

最后对谱嵌入应用k-means聚类并生成聚类分配矩阵$Q \in \{0,1\}^{2n_p\times k}$。使用共享锚点图像，通过跨图像标签映射函数$\theta_{i,j} : \{1,\ldots,k\} \to \{1,\ldots,k\}$ 从图像i 到图像j同步小FOV图像间的聚类分配，当$v = \arg\max_{v^*}\lfloor Q_i\rfloor_u \cdot \lfloor Q_j\rfloor_{v^*}$ 时，$\theta_{i,j}(u) = v$。操作符$\lfloor\cdot\rfloor_u$返回第u列的前n_p 维向量。在Ψ上进行的k-means 聚类对荧光图像中的高频噪

声具有鲁棒性，确保了可靠的分割。此外，通过锚点图像的联合谱嵌入确保了一致的聚类标签分配，避免了额外的跨图像聚类同步。

掩膜引导去噪

在多通道荧光图像虚拟染色的逐像素图像翻译中，在图像生成过程中保留细粒度亚细胞结构至关重要。我们引入了一种针对不同亚细胞结构的掩膜引导方案，通过聚焦于基于NSC的掩膜中嵌入的亚细胞形状来增强迭代去噪过程。我们通过扩散模型的前向过程对NSC获得的掩膜添加噪声，并将加噪后的掩膜作为y_t 用于反向去噪。

NSC获得的掩膜z可以过滤早期生成的亚细胞结构图像中的高频噪声，而一般轮廓和语义特征等低频信息可以在前向过程中保留。掩膜引导增强了感兴趣区域的像素生成，实现了语义感知去噪并进一步合理生成亚细胞结构细节。这种语义感知去噪方法通过同时对输入明场图像的分布和目标荧光图像的期望特征进行建模，确保最终输出突出显示亚细胞结构。

4.10 小结

在生物医学图像处理中，深度学习的发展，特别是“突破传统表征局限，提升模型的理解与适配能力”，既解决了“数据采集难、标注成本高”的行业痛点，也为“精准诊断、低风险治疗”提供了技术支撑，是未来医学影像智能化的核心方向。

第五章　计算机视觉

5.1　计算机视觉的演进

5.1.1　计算机视觉的历史：从基础到关键突破

计算机视觉的发展与人工智能、机器学习几乎同步，可追溯至计算机诞生初期（20世纪40 年代），但真正形成独立领域并爆发，经历了“生理基础→算法基础→ 设备与数据基础→任务演进”关键阶段。计算机视觉近60年的发展，离不开“神经生理、物质、算法”三大基础的支撑，这些基础共同定义了“机器如何看见”的技术框架。

5.1.2　视觉大模型：重塑任务边界的“新范式”

“大模型是否会终结传统计算机视觉？”，这类看似耸人听闻的标题，实则指向了当前技术的核心变革：若视觉大模型能接触到足够多的广泛数据，包括生物医学影像、自然场景、工业图像等，并通过少量任务数据完成“调制适配”，那么传统视觉任务，如分割、检测、三维重建或许都能得到高效解决。

海量数据训练的大规模预训练模型，已能在多数基础视觉任务中展现令人满意的性能”。无论是GPT-4o 的图像分割与深度估计、CLIP的无监督医学影像匹配，还是DINOv2 的通用特征提取，其核心前提都是“模型在预训练阶段见过足够多样的数据”，这种数据优势让大模型能跨越领域泛化，甚至无需专门训练就能处理部分生物医学图像处理任务。

生理基础：揭开视觉感知的底层逻辑

计算机视觉的第一个重要突破，源于神经生理学对“人类视觉机制”的探索，1950年代，神经生理学家David Hubel 与Torsten Wiesel（1981年诺贝尔生理学或医学奖得主）的“猫视觉实验”，为后续视觉模型提供了核心灵感。实验发现猫的视觉初级皮层单元对“移动边缘”刺激高度敏感，且大脑存在分层视觉机制，低级皮层处理边缘、角点等基础特征，高级皮层则负责物体识别、场景理解等复杂任务。

20世纪50-60年代的“猫视觉实验”（Hubel与Wiesel）揭开了生物视觉的底层逻辑：大脑皮层通过分层处理解析视觉信号，低级皮层对边缘、角点等基础特征敏感，

高级皮层负责物体识别与场景理解。这一发现直接催生了卷积神经网络的核心设计。局部感受野模拟皮层单元对局部刺激的敏感，分层特征提取，低级卷积层提边缘，高级层提语义。如今，人类神经生理学研究进一步证实“大脑不同区域对视觉、听觉等信号的特异性响应”，为视觉模型的“语义对齐”提供了更精细的生物参考，本质都是对生物视觉机制的工程化模仿。

算法基础：从感知机到早期视觉模型

算法是计算机视觉的“核心引擎”，其发展与机器学习同步，可分为两个阶段：

早期基础（1950-2010年）：以感知机（1958年）、新认知机（1980年，CNN雏形）为代表，模型浅、参数少，依赖手工设计特征，如Sobel边缘检测、HOG形状描述子；20世纪50-80 年代，是计算机视觉“算法框架奠基”的关键期，核心突破集中在“神经网络雏形”与“早期视觉任务定义。

深度学习革命（2010年至今）：以AlexNet（2012年）为起点，深度神经网络凭借“大规模参数+ 海量数据”，在分类、检测、分割等任务中全面超越传统方法，成为当前视觉算法的绝对主流。

- 感知机：神经网络的起点（1958年）心理学家Frank Rosenblatt在《智能自动机的设计》中提出“感知机”，这是首个具有“输入- 权重- 激活”结构的单层神经网络：其结构包含输入层、可调整权重的连接、阈值激活函数，与今天的神经网络结构高度相似。感知机仅能处理线性分类任务，无法解决“异或（XOR）”等非线性问题，这一局限直接引发了20世纪60 年代的“第一次人工智能寒冬”，学界认为神经网络难以应对真实世界的非线性任务，但它为后续深度神经网络提供了“分层计算”的基本思路。

- 新认知机：CNN的雏形（1980年）日本学者福岛邦彦提出“新认知机”，首次将“分层特征提取”与“池化”引入神经网络，被视为CNN 的直接前身：核心设计包含“输入层→卷积特征提取层→池化层”的分层结构，支持通过调整内部参数自适应学习输入特征；新认知机已具备CNN的所有关键模块，包括局部感受野、卷积、池化、分层提取，只是规模较今天的CNN 更小，但其架构逻辑至今仍是视觉模型的核心。

- 马尔的视觉理论：定义视觉计算流程（1970年代末）计算机科学家David Marr在《视觉》一书中提出“视觉计算三阶段理论”，为“机器如何理解图像”提供了首个系统性框架：输入图像→提取边缘/骨架等基础特征→生成三维模型→ 完成场景理解。这一流程至今未变，即便今天的深度神经网络，处理三维重建任务时仍遵循“输入图像→ 特征提取→三维表征生成”的逻辑。《视觉》一书近年仍在再版，足见其理论影响力。

设备与数据基础：从胶片到数字化图像

计算机视觉的落地，离不开“图像采集设备”与“数字化介质”的突破：在医学成像设备方面，1970年，Godfrey Hounsfield在英国EMI实验室发明CT扫描仪，核心算法为“拉东变换”，为医学图像处理提供了首个标准化三维数据来源，今天医院的CT设备仍是这一技术的迭代；自然图像依赖相机，医学图像依赖CT、MRI等专业设备，1970年CT扫描仪发明，为医学图像处理提供标准化三维数据。

在数字化图像方面，1959年，Russell Kirsch团队首次将其儿子的照片转化为“二进制灰度图像”，这是人类历史上第一张数字化图像。从此，计算机有了可处理的“视觉介质”。1950-60年代，数字图像技术诞生，使图像能以“矩阵/张量”形式在计算机中存储，今天我们处理的所有图像，二维照片、三维CT、图结构数据均基于数字化形式，这是后续算法处理的前提。

关键任务演进：从“无法解决”到“常态化应用”

计算机视觉的核心任务，分割、三维重建、分类等并非近年才提出，而是历经60余年迭代，从“理论设想”逐步走向“工程落地”：

（1）三维重建：60年未变的核心任务（1963年至今）

1963年，Lawrence Roberts在博士论文《三维固体的机器感知》中首次提出“从二维图像推导三维结构”，这是三维重建任务的起点。

早期三维重建依赖“线框图匹配”，能力极弱；20世纪90年代后，随着“立体视觉”（双目三角测量）、“运动恢复结构（SfM）”等算法出现，重建精度大幅提升。今天，三维重建已实现规模化应用，如谷歌地图的三维地形、城市三维建模、自动驾驶的实时环境重建。

GPT-4V生成深度图像、NeRF从二维X光重建三维CT，本质都是这一任务的“深度化延伸”。

（2）图像分割：从MIT暑期项目到今天的AI分割（1960年代至今）

1960年代，MIT启动“夏季视觉项目”，首次提出“自动分割前景与背景、提取非重叠物体”的目标，但因当时算法无机器学习支撑、数据无大规模标注，项目失败。

20世纪70-80年代，形态学算子，例如Sobel算子检测边缘、Canny算子提取轮廓成为分割核心工具；今天，大模型可通过“提示驱动”直接完成像素级分割，例如区分自然图像中的个体、医学影像中的肿瘤边界，实现了60 年前的目标。

（3）图像分类：从统计学习到深度学习革命（1980年代-2012年）

20世纪80-90年代，统计机器学习，如SVM、隐马尔可夫模型HMM、主成分分析PCA，是分类核心，通过手工设计特征（如HOG、SIFT）+ 分类器，实现简单图像分类。1998年，Yann LeCun提出“LeNet-5”，首次用CNN处理手写数字分类，即MNIST 数据集，架构与今天的图像分类CNN完全一致，即卷积→池化→ 全连

接→Softmax。2012年，AlexNet在ImageNet竞赛中以远超传统算法的精度夺冠，标志着“深度学习主导计算机视觉”的时代到来，从此，深度神经网络在分类、检测、分割等任务中全面超越传统模型。

历史发展：任务未变，能力跃迁

回顾计算机视觉60余年的发展，一个核心规律愈发清晰：“核心任务从未改变，改变的是解决任务的技术能力”。无论是1960年代的三维重建、MIT项目的分割目标，还是1970 年代马尔提出的“视觉理解流程”，这些任务至今仍是生物医学图像处理的核心。但今天的大模型，通过“海量数据学习的先验知识”，让这些曾经“无法解决”的任务变得触手可及。

这种“先验”的价值，正如人类单眼仍能感知深度，其源于一生的视觉经验，大模型见过的医学影像、自然场景远超个体专家，其学到的通用特征、语义关联，为生物医学图像处理提供了“低数据依赖”的解决方案。未来，随着大模型对领域数据的进一步适配，或许我们真的会进入“无需为每个任务设计专属模型”的新阶段，这不是“终结计算机视觉”，而是让视觉技术更高效地服务于临床与科研。

5.2 人类视觉、机器视觉与大模型范式

5.2.1 人类视觉与机器视觉的核心目标对齐

人类的视觉能力本质是“传感器（眼睛）→ 感知→ 任务指导”的闭环：通过眼睛捕捉二维图像，从中检测物体、感知形状与深度，最终指导日常行动，如抓取物品、判断距离。而机器视觉的核心目标，正是复刻这一能力，让计算机像人一样理解视觉信息，甚至突破人类视觉的局限。当前最先进的视觉大模型已展现出与人类视觉高度对齐的能力：

- 单目物体感知：无需训练即可检测图像中的物体及细节，如区分人的肢体部位、前景与背景场景，精度可媲美人类肉眼；

- 双目深度估计：模拟人类双眼的立体视觉，从单张或多张二维图像中推断深度信息，即物体与观测者的距离。这一能力的关键在于“先验知识”，正如人类单眼能感知深度，依赖对三维场景的经验认知，GPT-4o 通过海量数据学习的三维场景先验，可补全遮挡区域的部分三维信息，实现精准深度估计。

5.2.2 视觉方法的演进：传统与深度学习的分野与关联

在生物医学图像处理的论文中，视觉方法常被分为“传统方法”与“深度学习方法”，两者并非割裂，而是存在传承与突破：

传统方法（深度学习前）

传统方法包括手工特征算法如形态学滤波器、边缘检测算子、传统机器学习模型，SVM、贝叶斯隐马尔可夫模型。依赖专家设计特征，需“小数据+迭代优化”，计算复杂度低但泛化能力弱。可进行简单图像分割，例如阈值分割、基础目标检测例如模板匹配。

深度学习方法（2012年至今）

深度学习方法可进行端到端学习，无需手工设计特征，通过大规模数据学习通用语义，泛化能力强；与传统方法的关联。在流程上遵循马尔视觉理论的“输入图像→特征提取→ 任务处理”逻辑，例如CNN 的编码器提取特征、解码器完成分割。在模型方面新认知机的“卷积+池化”结构，是今天CNN的直接雏形；贝叶斯隐马尔可夫模型的“概率分布建模”，影响了变分自编码的隐空间设计。

深度视觉计算学习的优势与局限

在多数生物医学任务，例如CT肿瘤分割、眼底图像分类中，深度学习的精度远超传统方法，并未解决所有问题，例如早期自动驾驶视觉模型无法识别“伪装路障”，而GPT-4o虽能突破这一局限，但仍无法应对所有极端场景，例如对抗攻击图像。计算机视觉的核心任务，包括检测、分割、深度估计始终围绕“理解图像语义”展开，方法从“有监督”向“无监督”演进。

Mask R-CNN是图像分割的重要模型，基于“区域卷积神经网络（R-CNN）”改进，实现“检测+ 分割”一体化，至今仍是生物医学图像分割的常用基准模型，如器官分割、细胞分割等。需要**多阶段计算**：

区域检测：生成物体包围框，进行定位目标；

分类：判断包围框内物体的类别；

像素级分割：生成分割掩码，明确每个像素所属类别。

无监督的谱方法与图表示无需标注数据的无监督分割，依赖“图拉普拉斯矩阵的谱分解”，无需监督信号，仅通过“空域→ 谱域”变换即可完成分割，适用于数据稀缺场景。需要**多阶段计算**：

图像转图：将像素/超像素视为节点，用相似度构建图拉普拉斯矩阵；

谱分解：对矩阵做特征值分解，提取“费德勒向量”，除0特征值外最小特征值对应的向量；

分割：通过阈值或K均值聚类，基于费德勒向量实现前景-背景二分类或多类别分割。

5.3 大模型驱动范式变革：“训练模型”到“提示任务”

视觉大模型的出现，彻底改变了传统“数据→训练→模型”的流程，形成“提示（Prompt）→任务解决”的新范式。传统视觉任务需“任务专属数据+标注标签”，而大模型通过“预训练+ 提示”，可实现零样本任务处理。

GPT-4o的零样本分割：输入自然图像或医学影像，无需训练即可输出物体检测框、分割掩码与深度图。

CLIP的提示驱动分割：通过“点提示（点击目标）、包围盒提示（框选目标）、文本提示（如‘黑色椭圆区域’）”，CLIP可精准分割目标区域，无需标注任务专属数据。

视觉大模型的适配：从通用到特定任务。预训练视觉大模型对自然图像具备出色的处理能力，但其泛化性依赖“训练数据覆盖度”。若面临模型未见过的场景，如罕见病医学影像、特殊生物显微图像，需通过模型精调实现自适应。

基于预训练大模型特征，用少量特定任务数据，如目标领域的图像与标注微调部分参数，如输出层、中间特征层，让模型适配任务特性。无需从头训练模型，即可让精调后的模型输出符合需求的预测结果，如特殊病理切片的病灶分割。

范式本质：预训练先验的复用

大模型的能力源于“海量数据学习的通用先验”，其预训练过程已覆盖自然场景、医学影像等多领域，存储了“物体形状、三维结构、语义关联”等知识。当用户输入图像与提示时，大模型无需重新训练，仅需调用预存先验即可完成任务，大幅降低了对任务数据的依赖。

未来趋势：视觉计算的“低门槛化”

在生物医学图像处理中，这一范式的价值尤为显著：临床医生无需掌握深度学习知识，仅需提供影像与简单提示，如“分割左肺”，预训练大模型即可输出结果。这种“低门槛、高效率”的模式，正成为视觉计算与医学影像结合的主流方向。

从人类视觉的生物机制，到计算机视觉的三大基础，再到大模型驱动的范式变革，视觉技术的核心始终是“让机器更好地理解视觉世界”。当前，大模型已实现从“模仿人类视觉”到“超越部分人类视觉”的突破，而其最大的价值在于将视觉计算从“需要专业训练的技术”，转变为“人人可用的工具”，这不仅重塑了计算机视觉的发展路径，也为生物医学图像处理的临床落地提供了新的可能。

计算机视觉是人工智能的核心分支，目标是让机器具备“人类视觉般的理解能力”，通过算法模型处理图像/视频数据，实现对现实世界的**感知、分析与决策**。它与机器学习（提供学习框架）、深度学习（提供表征能力）、图像处理（提供底层操作）紧密衔接，是AI落地最广泛的领域之一。

从人类智能视角看，计算机视觉对应加德纳“多元智能理论”中的**视觉-空间智能**，即“准确感知空间信息、加工三维结构、再现视觉场景”的能力。在生物医学图像处理中，这种能力进一步延伸：不仅要“看到”器官结构，还要“理解”病理意义，例如区分正常组织与病灶。

计算机视觉的核心任务

计算机视觉的任务体系覆盖“从底层处理到高层理解”，且全部可迁移至生物医学场景：

任务类型	目标	生物医学应用案例
图像分类	判断图像所属类别	区分正常胸片/肺炎胸片、良性/恶性肿瘤切片
目标检测	定位图像中目标的位置与类别	检测CT中的肺结节、眼底图像中的黄斑病变
图像分割	划分图像中不同语义区域（像素级）	分割脑CT中的灰质/白质/脑脊液、肝脏肿瘤边界
图像生成	生成符合真实分布的图像	生成低剂量CT的“虚拟高剂量影像”、破损颌面的健康图像
图像增强	提升图像质量（去噪、去模糊、超分）	增强超声图像的清晰度、修复病理切片的污染区域
三维重建	从二维图像恢复三维结构	从X光片重建3D CT、从显微镜图像重建细胞三维模型
目标跟踪	追踪动态目标的运动轨迹	跟踪心脏超声中瓣膜的运动、内窥镜下病灶的位置变化

5.3.1 大模型对计算机视觉的颠覆性突破

近年大模型在计算机视觉领域的突破，打破了传统“数据-训练-部署”的闭环：

- **无需训练的即时任务处理**：以往处理分割任务需标注数据、训练模型，现在输入“分割CT 中的肝脏”提示，GPT-4V等多模态大模型可直接输出结果；
- **超越传统模型的泛化能力**：传统自动驾驶视觉模型易被“伪装路障”，如绘制场景图案的屏障欺骗，而大模型能通过深度估计与语义理解识别伪装，避障能力更接近人类；

- **跨模态的协同理解**：不仅能处理图像，还能结合文本、语音完成复杂任务，如输入“生成吉卜力风格的心脏解剖图”，大模型可同时满足“医学准确性”与“动画风格”需求；

- **辅助科研与生产效率**：科研中需绘制“半监督学习图像处理架构图”时，仅需文字提示，大模型即可生成专业、简洁的示意图，无需手动设计排版。

5.3.2 经典视觉模型的案例

除大模型外，传统视觉模型也通过迭代优化实现长期应用，典型代表包括：

YOLO：单阶段检测的高效范式

YOLO（You Only Look Once）自2016年提出以来，历经多代迭代，至今仍是主流目标检测模型，核心优势在于单阶段回归设计.速度快、部署成本低，适用于实时检测场景，如医学影像中的快速病灶定位。与两阶段模型，如R-CNN、Mask R-CNN需先生成候选区域再分类不同，YOLO直接将图像划分为网格，通过回归预测每个网格内目标的类别概率与边界框坐标，无需分阶段处理。

GAN：从图像生成到分割的“风格迁移”逻辑 GAN的核心能力是“图像翻译”，例如StyleGAN生成特定风格图像，可直接迁移至图像分割任务。将“原始图像”与“分割掩码”视为两种“风格”，通过GAN的生成器学习“图像→掩码”的映射，或反向“掩码→图像”的生成。生成模型可从患者面部图像生成不同疾病阶段的外观变化图，也可从胸部CT图像生成对应的肺叶分割掩码，实现分割与生成的联动。

ViT与CLIP：大模型的特征对齐与零样本能力 ViT（Vision Transformer）：作为当前视觉大模型的核心架构，通过“图像分块→注意力建模”提取全局特征，广泛用于特征嵌入。

CLIP：通过“视觉-文本嵌入对齐”实现零样本分割，输入图像与文本提示，如“肝脏肿瘤”，CLIP可通过比对视觉特征与文本特征的相似度，定位目标区域，无需任务专属标注。CLIP的无监督分割能力，本质是预训练阶段学到的“图文语义关联”，可直接迁移至未见过的类别或场景。

5.4 图像特征

在深度神经网络前，图像特征依赖手工设计，通过边缘、角点、纹理等低级特征描述图像语义，是传统视觉任务，诸如匹配、重建的基础。

边缘检测：Canny算子

边缘是图像中灰度突变的区域，是物体轮廓的基础。Canny算子是最经典的边缘检测算法，步骤清晰且鲁棒性强：通过梯度捕捉灰度变化，通过阈值与NMS筛选真

实边缘，是后续轮廓分析、目标定位的基础。

图像平滑：用高斯滤波器去除噪声，避免噪声导致的虚假边缘；

梯度计算：计算二维图像在x、y方向的梯度，如用Sobel算子，得到梯度幅值反映边缘强度与梯度方向反映边缘走向；

非极大抑制（NMS）：在梯度方向上，仅保留局部极值点，消除边缘的“宽化”，让边缘更细；

双阈值筛选：设置高阈值与低阈值：

幅值高于高阈值：强边缘，直接保留；

幅值介于两阈值之间：弱边缘，仅保留与强边缘连通的部分；

幅值低于低阈值：非边缘，剔除；

边缘连接：将离散的强边缘与连通弱边缘连接，形成完整物体轮廓。

角点检测

角点是图像中“多方向梯度突变”的点，如矩形的顶点，是图像匹配、三维重建的关键特征，1988 年提出的Harris 算子通过矩阵分析定位角点。步骤如下：

窗口变化分析：在图像中滑动窗口，计算窗口内灰度值随窗口偏移（Δx, Δy）的变化量；

泰勒展开近似：对灰度变化量做一阶泰勒展开，简化为“梯度向量×偏移向量”的形式，引入Harris 矩阵M：

$$M = \sum_{(x,y)\in\text{窗口}} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

其中I_x、I_y分别是x、y方向的梯度；

3. 特征值判断：对M做特征值分解，得到两个特征值λ_1、λ_2，通过特征值分布判断区域类型：

λ_1、λ_2均很小：平坦区域（无灰度变化）；

λ_1很大、λ_2很小或反之：边缘区域，仅单方向梯度变化；

λ_1、λ_2均很大：角点，多方向梯度变化；

4. 响应函数简化：为避免复杂的特征值分解，定义Harris响应函数：

$$R = \det(M) - k \cdot \text{trace}(M)^2$$

其中$\det(M) = \lambda_1\lambda_2$（行列式），$trace(M) = \lambda_1 + \lambda_2$（迹），k 为经验常数（通常取0.04∼0.06）；

$R >$ 阈值：角点；

$R < 0$：边缘；

$|R|$很小：平坦区域。

Harris算子无需预设模板，可自适应检测任意形状的角点，是后续特征匹配，如SIFT的基础。

纹理特征：LBP与灰度共生矩阵

纹理是图像中“重复出现的灰度模式”，如皮肤纹理、织物花纹，反映物体表面的微观结构。LBP（局部二值模式）是简单高效的纹理编码 LBP通过比较“中心像素与邻域像素的灰度差异”生成二进制编码，描述局部纹理模式，计算简单、旋转不变，通过旋转邻域至主方向实现，适用于纹理分类，如病理切片的细胞类型区分。

邻域选择：取中心像素的3×3邻域（可扩展至更大邻域）。

灰度对比：将每个邻域像素的灰度值与中心像素比较：邻域灰度≥ 中心灰度：编码为1，邻域灰度< 中心灰度：编码为0。

二进制转十进制：将8个邻域的二进制编码按顺时针/逆时针顺序拼接，转成十进制数，作为该中心像素的LBP 值。

纹理统计：统计图像中LBP值的直方图，作为图像的纹理特征。

灰度共生矩阵（GLCM）是基于灰度关联的纹理描述，通过统计“图像中灰度值为i与j的像素在特定方向、距离上同时出现的概率”，反映灰度分布的空间关联：GLCM 能捕捉纹理的全局统计特性，适用于复杂纹理分析，如遥感图像的地物分类、医学影像的组织病变判断。

矩阵构建：设图像灰度级为L（如8级灰度），构建L×L的GLCM矩阵，矩阵元素$P(i,j,d,)$ 表示“沿方向 θ 、距离d，灰度i与灰度j相邻出现的概率”。

纹理特征计算：基于GLCM的概率分布，计算4类核心纹理特征：

对比度：

$$\sum_{i,j}(i-j)^2 \cdot P(i,j)$$

反映灰度差异的明显程度，对比度高则纹理更清晰；

能量：

$$\sum_{i,j}P(i,j)^2$$

反映纹理的均匀程度，能量高则纹理更规则；

相关性：

$$\sum_{i,j}\frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i\sigma_j}$$

反映灰度的线性关联，相关性高则纹理具有方向性；

同质性：

$$\sum_{i,j}\frac{P(i,j)}{1+|i-j|}$$

反映纹理的局部均匀性，同质性高则纹理更平滑。

特征匹配与三维重建：SIFT的旋转不变性

SIFT（尺度不变特征变换）是经典的“关键点检测+特征描述”算法，解决了“尺度、旋转变化下的图像匹配”问题，是三维重建、全景拼接的核心工具：手机拍摄街景生成三维地图、双目相机的三角测量，通过匹配点计算三维坐标、全景图拼接等，至今仍是传统三维重建的重要工具。步骤：

尺度空间极值检测：构建高斯金字塔与差分高斯（DoG）金字塔，在26个邻域，同尺度8 个+ 上下尺度各9 个中检测极值点，实现尺度不变性；

关键点定位：剔除低对比度、边缘响应的极值点，精确定位关键点坐标与尺度；

主方向计算：统计关键点邻域内的梯度方向直方图，取峰值方向作为主方向，实现旋转不变性；

特征描述子生成：将关键点邻域划分为4×4子区域，每个子区域计算8方向梯度直方图，生成4×4×8=128维特征向量；

特征匹配：通过“最近邻搜索”匹配不同图像中的SIFT特征点。

5.4.1 传统特征与深度特征的对比：范式的转变

表 5.1: 传统手工特征与深度神经网络特征的比较

	传统手工特征（如Canny、Harris、SIFT）	**深度神经网络特征（如ViT、DINOv2）**
特征来源	手工设计，依赖领域知识	数据驱动，自动学习语义关联
语义性	低级特征（边缘、角点），无语义信息	高级特征（物体部件、类别关联），含强语义
泛化性	仅适用于特定场景，如SIFT适用于纹理丰富图像	跨领域泛化，预训特征可用于医学、自然图像）
任务依赖	需为不同任务设计不同特征	通用特征可迁移至多个任务

深度特征并未完全取代传统特征，在数据稀缺、实时性要求高的场景，传统特征，如LBP、Harris仍因“计算简单、无需训练”具备优势；而在复杂语义任务，如医学影像分割、三维重建中，深度特征的强语义性与泛化性更具价值。

从视觉大模型的精调适配，到传统图像特征的手工提取，两者共同构成了视觉计算的技术脉络：大模型通过“预训练+ 精调”解决了通用到特定任务的适配问题，而传统特征则为理解“视觉如何被描述”提供了基础逻辑。在生物医学图像处理中，

这种技术融合尤为重要，既可以通过CLIP、ViT 等大模型快速实现零样本分割，也可以在数据有限时结合传统特征，如GLCM的纹理分析提升模型鲁棒性，最终实现“高效、精准”的视觉任务处理。

关键点检测的加速：ORB算法的优化逻辑

在计算图像梯度的幅值与方向时，为降低复杂度，可通过盒式滤波器（Box Filter）替代传统的差分高斯（DoG）计算，盒式滤波器无需复杂的高斯卷积，仅通过邻域像素求和即可近似梯度，大幅提升运算速度。在此基础上，结合海森矩阵的极值检测，可快速定位关键点。

SIFT需构建DoG金字塔，在26个邻域，同尺度8个+上下尺度各9个中搜索极值点，计算成本高。ORB则通过盒式滤波器简化梯度计算，再利用哈尔小波响应生成方向直方图，无需统计梯度的幅值与方向，仅通过邻域像素的数值求和与绝对值求和即可形成特征描述子，最终将计算复杂度降低一个数量级，同时保持了旋转不变性与尺度鲁棒性。ORB的优势在于“速度优先”，适用于实时场景，如移动端的特征匹配、视频流的目标跟踪，但其特征维度（32 维或64维）远低于SIFT（128维），在纹理稀疏图像中的匹配精度略逊于SIFT。

5.4.2 三维形状的表征：Shape DNA与“等谱不等距”问题

人类可轻易区分不同姿态的三维模型，如不同姿势的猫的网格模型，但计算机需通过形状描述子实现这一能力。Shape DNA（形状DNA）是经典的全局三维形状描述子，通过图拉普拉斯矩阵的特征值表征三维表面模型：

图拉普拉斯矩阵构建：将三维模型的表面网格视为图，计算该图的拉普拉斯矩阵，其元素反映顶点间的连接权重与局部结构；

特征值提取：对图拉普拉斯矩阵做特征值分解，取前k个最小特征值作为Shape DNA，这些特征值蕴含了三维模型的全局拓扑与几何信息。

“等谱不等距”的局限性

“听到鼓的声音，能否确定鼓的形状？”鼓的“声音”对应图拉普拉斯矩阵的特征值（谱），鼓的“形状”对应顶点间的距离（等距）；答案是“等谱不等距”：即便两个矩阵的特征值完全相同（等谱），也无法保证对应的三维模型形状一致（等距），即存在“不同形状的鼓，能发出相同的声音”。尽管存在理论局限，但实验表明：相似的三维形状，如不同姿态的同一只猫，其Shape DNA 的特征值会聚集在相邻区间，因此Shape DNA仍能用于三维模型的相似度检索，如工业零件分类、生物医学中的骨骼形状比对。

5.4.3 深度模型的特征：从Feature Map到跨模态适配

深度特征的可视化：语义对齐的价值

当前视觉大模型的特征不再是传统的“边缘、角点”，而是通过深度神经网络生成的特征图或特征体。以自监督预训练模型DINOv2为例：从DINOv2的第2/5/8层提取的Feature Map中，不同图像的同类物体，如两张汽车图像、两张狗图像在语义对应位置，如汽车的车轮、狗的耳朵呈现相同颜色，这意味着Feature Map的数值在语义相似区域高度一致，无需额外标注即可实现“图像间的语义对应”。

DINO特征的强语义性，彻底替代了传统手工特征，如SIFT、Harris角点：传统特征需通过“检测关键点→ 匹配描述子”实现对应，而深度特征可直接通过Feature Map的数值相似性定位语义对应区域，大幅简化了图像匹配、三维重建等任务的流程。

三维模型的深度特征适配

当前视觉大模型多基于二维图像训练，若需为三维网格模型提取特征，可通过“投影- 特征提取- 反向投影”实现：

多视点投影：将三维模型按不同视角投影到二维平面，生成一系列二维投影图。二维特征提取：用预训练二维大模型提取每张投影图的像素级Feature Map。反向投影：将二维Feature Map的像素特征，按投影矩阵映射回三维模型的对应顶点，最终得到三维顶点的特征向量。

这种方法的优势在于“复用二维大模型的泛化能力”，无需训练专属三维模型，即可为三维网格赋予强语义特征，可用于三维模型的稠密对应，如不同姿态猫的顶点匹配；生物医学中的三维形状变化度量，如肿瘤治疗前后的体积变化，需通过对应顶点的位移计算，直接对顶点序列相减无意义。

5.4.4 动态特征与深度估计

图像的动态性，如视频帧间的物体运动可通过光流描述，其本质是“相邻帧间像素的位移场”，反映每个像素在时序上的运动轨迹。光流的计算逻辑与生物医学图像的图像配准完全一致。光流侧重“时序上的像素位移”，如视频中行人的行走轨迹。图像配准侧重“空间上的语义结构位移”，如CT影像中器官的位置偏差。他们均需从一对图像中推测语义对应像素的位移，深度学习端到端模型，例如FlowNet、RAFT直接输出位移场，无需手工设计匹配规则。

人类通过单眼即可感知深度，依赖场景先验，视觉大模型也可通过预训练实现这一能力。输入单张自然图像，可通过Transformer架构提取多尺度特征，结合注意力机制建模像素间的全局关联，最终输出像素级的深度图，数值越大表示距离越远；自动驾驶中的障碍物距离判断，如识别伪装路障的深度异常、医学影像中的器官层

次划分，如CT 中骨骼与软组织的深度区分。

5.4.5 复杂物体的表征：部件模型的应用

对于由多个子部分组成的物体，如人脸、人体需通过部件模型（Part-Based Model）建模整体与局部的关联。这类模型最早在2008年左右兴起，将物体拆分为可复用的部件，通过部件的组合与形变描述整体形状。

传统部件模型，如deformable part model（DPM），将人体拆分为头部、躯干、四肢等部件，通过统计部件的位置偏移与尺度变化，实现目标检测，如行人检测。

深度学习部件模型：在预训大模型中嵌入部件注意力机制，如人体姿态估计模型HRNet，通过多分辨率分支建模关节与肢体的关联，可直接输出部件的语义分割与位置坐标。

部件模型的价值在于“处理形变鲁棒性”，即便物体存在姿态变化，通过部件的相对位置约束，仍能准确识别整体类别与局部结构。

5.4.6 从词袋模型到深度学习：图像表征与分类的范式演进

词袋模型最初用于文本处理，后被迁移至图像领域，其核心逻辑是“将数据拆解为无序的基础单元，通过单元的出现频率表征整体”。

词袋模型的局限

无法表征顺序与位置关系 文本域中无法体现句子中词汇的语序，如“猫追狗”与“狗追猫”在BoW中表示相同，因词汇集合一致。

图像域中无法刻画特征点的空间位置关系，例如，两张包含“边缘、角点”特征的图像，即便特征点分布完全不同一张是人脸，一张是汽车，BoW也会因“特征词集合一致”将其判定为相似，导致分类任务失效。

缺乏语义表达能力 BoW中的“词”是孤立的基础单元，无法捕捉“词与词的组合语义”。文本中，“红色”与“苹果”的组合语义，例如“红色苹果”无法通过BoW体现，仅能记录两个词的单独出现频率。

图像中，“边缘+角点”的组合可能对应“矩形”，但BoW仅能分别统计“边缘”“角点”的数量，无法关联两者的组合语义，导致无法区分复杂物体。

依赖手工设计特征 BoW中的“词”需通过手工设计的特征描述子生成，图像领域的典型特征包括：**SIFT**：通过梯度计算、方向直方图生成128维特征向量，刻画局部表观信息。**Haar特征**：通过矩形区域的像素差值，提取边缘、线条、中心环绕等低级特征。其他手工特征：如ORB、SURF等，均需人工定义计算规则，如梯度方向、邻域划分方式。

这些特征的生成流程相似，先检测图像中的特征点，再计算每个特征点的描述子，最后通过聚类生成“词表”，将特征描述子映射为词表中的索引，形成图像的BoW表示。深度学习的出现彻底改变了图像表征的逻辑，其与传统BoW模型的差异体现在“特征生成、语义层次、上下文处理”三个方面。

表 5.2: 词袋模型与深度学习模型的比较

	词袋模型	深度学习模型
特征生成方式	手工设计特征，SIFT、Haar等，规则固定	自动学习特征，通过卷积核、注意力机制，数据驱动
语义层次	仅捕捉低层局部特征，边缘、角点，无语义	层次化特征：低层（边缘）→中层（部件）→高层（语义）
上下文处理能力	无顺序/位置信息，无法捕捉全局关联	可建模局部-全局上下文，CNN 靠感受野扩大，ViT 靠自注意力
特征量化与聚合	人工聚类生成词表，量化为索引后简单计数	自动向量量化，如NeRV的特征聚类，多尺度聚合，如金字塔池化

尽管差异显著，深度学习仍继承了BoW“特征量化-聚合”的思路，并进行了优化：

- **特征量化**：BoW通过K-Means聚类生成词表，深度学习则通过神经网络自动学习量化规则，有深度神经网络将多通道特征嵌入量化为单索引，简化表示；
- **特征聚合**：BoW仅对“词频”进行统计，深度学习则支持多尺度聚合，深度神经网络也使用空间金字塔池化，从不同尺寸特征图中聚合信息，保留更多空间细节；

在二维X光到三维CT的重建任务中，深度学习通过量化二维图像的隐式嵌入，建立与三维CT 嵌入的映射，这一过程与BoW的“词表映射”一致，但因特征是自动学习的语义特征，重建精度远高于BoW。

图像分类：从传统模型到深度学习范式

图像分类是视觉任务的基础，其发展历程直接反映了从BoW到深度学习的范式转变。传统分类依赖“BoW特征+浅层分类器”，例如SVM、逻辑回归。

- **决策边界依赖人工阈值**：分类时需手动设置概率阈值，高于阈值为类别A，低于为类别B，阈值选择依赖经验，鲁棒性差。

- **特征表达能力弱**：仅靠低层特征无法区分语义相似但外观不同的物体，如“猫”与“狗”，BoW可能因均包含“毛发纹理”特征导致混淆。

当前主流的图像分类模型基于**卷积神经网络**或**ViT**，典型架构包括三部分：

特征提取模块：

- **CNN**：通过卷积层学习卷积核提取局部特征，池化层降维并增强鲁棒性，随着网络加深，感受野扩大，逐步捕捉高层语义。
- **ViT**：将图像分块为Patch，通过自注意力机制建模全局Patch关联，直接学习全局语义特征。

分类头模块：

- 通常由全连接层构成，将高层特征映射为类别概率。
- 最终通过Softmax函数归一化概率，概率最大的类别即为预测结果。

训练优化：

- **损失函数**：二分类用“二分类交叉熵”，多分类用“多分类交叉熵”，通过预测概率与真实标签的差异计算损失。
- **优化器**：常用随机梯度下降、Adam等，最小化损失函数以更新模型参数。

深度学习分类模型的发展

- LeNet-5（1998年）：首个用于图像分类的CNN，解决手写数字（MNIST）分类问题，架构包含卷积、池化、全连接层，奠定CNN基础；
- AlexNet（2012年）：深度学习爆发的标志，在ImageNet竞赛中以远超传统模型的精度夺冠，首次证明“大规模数据+深度CNN”的强大能力，推动深度学习在视觉领域的普及；
- VGGNet加深网络层数、ResNet引入残差连接解决梯度消失、ViT 首次将Transformer用于图像分类，建模全局关联，以及当前基于预训大模型的分类器如用DINOv2提取特征，仅微调分类头。

5.4.7 度量学习

度量学习（Metric Learning）是深度学习中优化“特征相似度”的关键技术，核心目标是“让相似样本的特征距离更近，不相似样本的特征距离更远”，解决传统分类模型“仅关注类别标签，忽略样本间相似性”的问题。

度量学习的核心问题包括：相似性约束：通过对比损失（Contrastive Loss）、三元组损失（Triplet Loss）等，定义“相似样本对”，例如同一类别的不同图像，以及“不相似样本对”，如不同类别的图像；特征优化：训练特征提取器，使相似样本的特征向量在嵌入空间中聚类，不相似样本的特征向量远离，最终得到“具有判别性的特征表示”。

典型应用

- 细分类任务：如区分不同品种的猫，外观相似但类别不同，传统分类模型易混淆，度量学习可通过优化特征相似度实现精准区分；
- 检索任务：如医学影像检索，给定一张肿瘤CT，检索数据库中相似的病例，度量学习可保证检索结果的语义一致性；
- 半监督/无监督学习：在标签稀缺场景下，通过度量学习挖掘样本间的相似性，实现无监督特征聚类，如DINOv2的自监督对比学习。

表示学习：从线性到非线性的特征映射

表示学习的目标是**将原始特征空间映射到更优的新空间**，使样本在新空间中满足”相似样本更近、不相似样本更远”的判别性标准，为后续分类、检索等任务奠定基础。其方法按映射类型可分为”线性”与”非线性”两类：

线性表示学习：马氏距离与降维

在传统机器学习中，线性表示学习通过学习**线性变换矩阵**实现特征空间优化。可以利用马氏距离（Mahalanobis Distance）学习一个变换矩阵，消除原始特征的冗余与相关性，使变换后的特征在新空间中更易区分，例如，通过矩阵映射，让同一类别的正样本在新空间中聚类，不同类别的负样本远离；通过降维优化，当原始特征维数过高，如高分辨率图像的像素特征时，通过PCA （主成分分析）、LDA （线性判别分析）等算法，在保留关键信息的同时降低维度，减少计算成本，同时提升后续任务的鲁棒性。

非线性表示学习：从核技巧到深度学习

当原始数据呈非线性分布，如”异或”问题时，线性映射无法满足判别性要求，需通过**非线性映射**将数据投射到高维空间，使其变得线性可分：

- 传统核技巧（Kernel Trick）：以SVM为例，通过”核函数，如RBF 核、多项式核”将低维非线性数据隐式映射到高维空间，无需显式计算高维特征，仅通过核函数计算样本间的相似度即可实现线性分类；

- 深度学习非线性映射：现代表示学习以深度神经网络为核心，通过卷积层、注意力层等非线性模块，自动学习从原始数据到判别性特征的映射。例如CNN通过堆叠卷积核以及非线性激活函数，逐步将像素级原始特征映射为包含语义信息的高层特征，无需人工设计映射规则。

相似度的判别标准

度量学习是表示学习的核心分支，专注于**定义与优化"样本相似度测度"**，确保新空间中的距离度量能准确反映样本的语义关联。其核心围绕"锚点-正样本-负样本"的三元组展开，通过损失函数驱动特征优化。

度量学习的损失函数

三元组损失（Triplet Loss） 三元组损失是最经典的度量学习损失函数，需输入"锚点样本（Anchor）、正样本（与锚点同类）、负样本（与锚点异类）"三组样本，优化目标明确：通过迭代优化，使同类样本在特征空间中紧密聚类，异类样本远离，最终降低后续分类任务的决策边界复杂度，例如，原本线性不可分的"猫/狗"图像，经三元组损失优化后，可通过简单的线性边界区分。通过最小化"锚点-正样本"的距离（$d(A,P)$），最大化"锚点- 负样本"的距离（$d(A,N)$），且要求$d(A,P)+\alpha<d(A,N)$ ，α 为间隔阈值，确保正负样本距离差足够大；

对比损失（Contrastive Loss） l直接对"样本对"（正样本对/负样本对）进行优化，逻辑更简洁：三元组损失依赖"锚点- 正- 负"三元组，对比损失仅需"样本对"，数据准备更灵活，适用于标签稀缺场景。

- 正样本对：最小化两者的欧氏距离（d^2），确保相似样本接近；
- 负样本对：若两者距离小于预设阈值（m），则最小化$\max(m-d,0)^2$，强制不相似样本距离大于m。

典型度量学习网络：孪生网络（Siamese Network）

孪生网络是度量学习的经典架构，通过"共享权重的双分支子网络"实现特征一致性。网络结构包含两个完全相同的子网络，如CNN、ViT，权重共享，确保输入的两张图像（I_1, I_2）经过相同的特征提取逻辑。相似度计算中，将两张图像的特征嵌入（$f(I_1), f(I_2)$）输入相似度计算模块，常用"欧氏距离"或"余弦相似度"度量关联。进行任务适配，输出"相似/不相似"的判别结果，或通过损失函数，如对比损失优化特征提取模块，最终实现"同类别特征距离小、异类别距离大"的目标。

自监督度量学习：无标签的数据增强策略

当缺乏人工标签时，可通过**自监督学习**生成"伪标签样本对"，实现度量学习。

DINOv2 的自监督对比学习，通过对图像生成多个增强视图，优化视图间的特征相似度，使模型学到具有泛化性的判别特征，无需人工标注即可用于下游分类任务。

正样本对生成：对同一张图像进行数据增强，如剪切、旋转、高斯噪声、医学图像的仿射变换，增强后的图像与原图构成正样本对，即同源数据，语义一致。负样本对生成：不同图像的增强结果互为负样本对。

5.5 迁移学习：预训模型的知识复用与适配

迁移学习的核心是**将"原任务（已学知识）"的经验迁移到"目标任务（新任务）"**，解决目标任务数据稀缺、训l练成本高的问题，是当前深度学习应用的主流范式。

迁移学习的定义与场景

无需从头训练模型，仅通过"微调预训模型"即可快速适配目标任务，大幅降低算力与数据成本。其中原任务的数据量充足、泛化性强的任务，如ImageNet自然图像分类，对应预训练大模型。目标任务一般是数据稀缺的特定任务。

迁移学习的三种方式

样本迁移：加权适配目标任务 当目标任务样本极少时，对原任务中"与目标任务相似的样本"赋予更高权重，辅助目标任务训练，例如在"罕见病CT分类"任务中，对原任务中"肺部相关自然图像"加权，提升模型对肺部结构的关注度。

特征迁移：复用预训模型的特征提取能力 是最常用的迁移方式，冻结预训模型的特征提取模块，仅微调下游任务层：

- 特征提取模块复用：预训模型的编码器已学到通用特征，如边缘、部件、语义，直接用于目标任务的特征提取，无需重新训练。

- 微调策略：冻结编码器权重，仅训练"分类头"或"解码器"，如分割任务的掩码生成层，例如，用CLIP的视觉编码器提取病理切片特征，仅训练一个小规模全连接层实现"肿瘤/正常组织"分类。

模型迁移：预训权重初始化目标模型 将原任务预训模型的权重作为目标任务模型的初始权重，替代随机初始化。模型从"已有知识"开始训练，收敛更快、泛化性更强。在"生物医学图像分割"中，用ImageNet预训的ResNet 权重初始化U-Net的编码器，再用少量医学图像微调解码器，分割精度远高于随机初始化的U-Net。

视觉语言大模型的迁移适配

视觉语言大模型通过”图文对齐预训”具备跨模态泛化能力，迁移到生物医学领域需针对性调制：**预训基础**：CLIP在海量自然图像-文本对上训练，实现”视觉特征与文本特征的对齐”。**生物医学调制**：用生物医学图像-文本对，如”肺结节CT-右肺上叶磨玻璃结节”微调CLIP 的视觉编码器与文本编码器，使其适配医学语义。**下游任务适配**：调制后的模型可直接用于”无监督分割”，文本提示”肺结节”，匹配CT 图像的视觉特征。

域内与域外测试：迁移学习的泛化性边界

迁移学习的泛化性能受”原任务与目标任务的域一致性”影响：

- **域内测试**：测试数据与预训/微调数据的分布一致，如用”胸部CT”微调的模型，测试其他胸部CT，性能通常较好。
- **域外测试**：测试数据分布与训练数据差异大，如用”胸部CT”微调的模型，测试”腹部CT”，性能会下降。

若目标任务与原任务域差异过大，如自然图像预训模型直接测试生物显微图像，需增加”域适应”步骤，缩小数据分布差异，提升泛化性。

实例：图像配准中的迁移学习应用

图像配准的传统深度学习方法需”训练特征提取器+解码器”，而基于预训大模型的方案可大幅简化流程：

- **传统方法**：训练CNN编码器提取两张图像（参考图像/移动图像）的特征，再训练解码器预测形变场，使移动图像经形变后与参考图像重合，需大量标注的配准数据。
- **大模型迁移方法**：对于特征提取，直接用预训视觉大模型提取两张图像的高层语义特征，无需训练编码器。考虑形变场优化，不训练解码器，通过”线性迭代优化”，如基于特征相似度的迭代最近点算法，直接求解形变场，使两张图像的特征分布对齐。

大模型迁移方法减少对标注数据的依赖，降低训练成本，同时利用预训模型的强语义特征，提升配准精度。

端到端模型的演变：从“无迭代”到“预训+迭代”的融合

在深度学习语境中，“端到端模型”的定义是“输入直接对应任务原始数据，输出直接为任务目标，如配准的形变场、分割的掩码”，在线预测阶段无需额外优化步

骤。例如传统图像配准的端到端模型，会直接通过神经网络输出从“移动图像”到“参考图像”的形变场，无需手动设计迭代优化流程。但随着预训大模型的普及，这一范式发生了调整。

预训模型的特征提取能力极强，能生成具有强语义性的图像表征，大幅降低了“迭代优化”的计算压力。因此，传统的“数值迭代优化算法”，如基于梯度下降的形变场求解，重新被整合进深度学习流程，例如在医学图像配准中，先用预训模型提取两张图像的高层特征，再通过轻量级迭代优化，而非训练复杂解码器求解形变场，既保留了端到端的便捷性，又提升了精度。

这种“预训特征+迭代优化”的融合模式，打破了“端到端=无迭代”的刻板认知，也体现了技术发展的包容性：传统数值优化、机器学习小模型、深度学习大模型并非相互替代，而是在不同任务场景中协同存在。

迁移学习的典型应用：从自然图像到医学图像的适配

迁移学习在生物医学图像处理中将自然图像预训模型的“通用特征提取能力”迁移到医学场景，解决医学数据稀缺、标注成本高的问题。

跨模态特征迁移的局限性 自然图像预训模型，如ResNet、ViT常被认为具有“模态无关”的特征提取能力，但实际应用中需注意：

自然图像的特征与医学图像的特征，如CT的组织密度、MRI的信号强度存在本质差异；预训模型对自然图像的“边缘检测”能力，无法直接适配医学图像中“模糊病灶边界”的识别，需通过少量医学数据微调特征提取模块，才能让模型理解医学场景的特殊语义，如肿瘤与正常组织的密度差异。

分割模型的医学化迁移：自然图像分割模型迁移到医学场景时，需针对性优化。SAM（Segment Anything Model）在自然图像分割中表现出色，但对医学图像的“微小病变”“多器官重叠区域”识别精度低，因其预训数据中缺乏医学领域知识，如器官解剖结构、病变形态。

通过“医学图像+专家标注”微调模型参数，构建专用模型。针对CT/MRI的多模态特性，调整模型的输入通道。针对“小病灶分割”，优化模型的局部感受野，如增加高分辨率分支。微调后的模型可适配多器官、多组织的分割任务，甚至无需额外训练即可处理常见医学影像，性能远超原始自然图像模型。

5.5.1 数据增强：生物医学图像的“样本扩充与泛化提升”

在医学图像样本量较小时，数据增强是缓解过拟合、提升模型泛化能力的关键手段，其模拟真实场景中的数据变异，既扩大数据集规模，又让模型学习到结构的鲁棒特征。医学图像的增强需兼顾“解剖结构合理性”与“任务适配性”，常见方法分为两类：

几何变换

- 刚性变换：通过旋转、缩放、平移、错切等线性变换，模拟图像采集时的姿态差异。
- 非刚性变换：通过仿射变换、弹性形变，模拟人体组织的自然形变，如呼吸导致的肺部形态变化、不同个体的骨骼结构差异。
- 裁剪与拼接：对高分辨率医学图像（如全身CT）进行局部裁剪，或拼接多幅图像的局部特征，增强模型对局部结构的关注度。

表观与语义增强

- 颜色/强度调整：对MRI图像调整信号强度、对病理切片调整染色浓度，模拟不同设备的成像差异。
- 噪声与伪影添加：加入高斯噪声、椒盐噪声，模拟低剂量CT的噪声特性；加入金属伪影（如牙科CT 的假牙伪影），提升模型对伪影的鲁棒性。
- 遮挡与擦除：随机遮挡图像局部区域，或使用“掩码自监督学习”，让模型通过残缺信息还原完整结构，增强特征提取能力。

医学图像的数据增强并非“无意义的随机变换”，而是对真实场景的还原，例如通过非刚性形变模拟“不同患者的肝脏形态差异”，通过噪声添加模拟“低剂量CT的成像质量”，让模型在训练阶段就接触到临床中可能遇到的各类数据变异，避免在实际应用中因“未见过的结构形态”导致预测失败。

5.5.2 生物医学图像处理面临的挑战

尽管深度学习与迁移学习带来了技术突破，但生物医学图像的特殊性仍带来诸多未解决的问题：

图像多变性：结构与表观的复杂差异

医学图像的“多变性”源于解剖结构的个体差异与成像条件的不稳定，具体表现为：

结构形变：同一器官在不同生理状态下形态差异大，如呼吸导致的肺部膨胀/收缩、心脏的舒张/收缩；

表观差异：即使是同一结构，因患者性别、年龄、体重差异，在图像中的信号强度、纹理特征也不同；

成像偏差：不同设备、不同参数导致图像质量差异，甚至同一设备的不同扫描批次也会产生偏差。这些差异使得“同一结构、不同图像”的表观相似度极低，大幅增加了模型的识别难度。

噪声与伪影：关键信息的丢失与干扰

医学图像的噪声与伪影是影响模型性能的障碍：

低剂量成像的噪声：为降低患者辐射剂量，低剂量CT的图像噪声显著增加，导致微小结构的边界模糊甚至消失。

金属伪影的干扰：患者体内的金属植入物，如假牙、心脏支架，会导致“射线硬化伪影”，使伪影区域的图像完全退化，丢失底层结构信息，如金属附近的骨骼形态无法识别。

模态专属伪影：MRI的“运动伪影”，患者扫描时的轻微移动、超声的“斑点噪声”，均会干扰模型对病变的判断。

数据不均衡与模型偏见

医学数据的“类别不均衡”是普遍现象，直接导致模型存在偏见，泛化能力差：2018年CellSeg细胞核分割数据集以欧美人群样本为主，训练出的模型对非洲人群的细胞核解析精度极低，因不同人种的细胞核大小、染色特性存在显著差异；2024年补充的“African-CellSeg”数据集，才缓解了这一偏见；若训练数据仅覆盖某一地域、某一疾病阶段的样本，模型会优先学习“多数类样本”的特征，对“少数类样本”，如罕见病、小众人群的预测精度骤降，严重影响临床应用的公平性。

异常检测的泛化难题

生物医学图像的“异常检测”，如肿瘤、病变识别面临“异常多样性”的挑战。异常类型繁多且不断新增，如罕见病的新型病灶形态，模型无法通过有限的训练数据覆盖所有异常模式。大脑CT的“异常”可能包括脑出血、脑梗死、肿瘤等，每种异常的形态、位置差异极大，且新的病变类型无法被现有模型识别，如何让模型具备“未知异常的检测能力”，仍是当前研究的难点。

生物医学图像处理的技术发展，既受益于深度学习的端到端范式与迁移学习的知识复用，也需直面图像多变性、噪声干扰、数据不均衡等固有挑战。未来的突破方向，不仅需要更先进的模型架构，如多模态大模型、自监督异常检测模型，更需要“技术与临床的深度融合”，结合医生的解剖知识优化模型设计，基于多中心、多样化的医学数据构建无偏见数据集，最终让AI模型真正适配临床场景的复杂需求，实现“精准、公平、鲁棒”的医学影像分析。

5.6 视觉计算的应用

人脸三维重建

人类可从单张二维人脸照片感知三维形状，状如雕塑家通过照片创作三维雕像，计算机可通过“自编码器+ 形状补全”实现。输入部分遮挡的人脸图像，如口罩遮

挡，自编码器的编码器提取人脸特征，解码器先补全二维完整人脸，再通过三维形状预测分支生成对应的三维人脸网格，包含深度信息。在生物医学中可用于颌面修复，根据患者二维照片生成三维颌面模型，用于定制修复体。

动作与手势识别

通过二维视频帧重建三维人体动作，是当前视觉计算的成熟应用。用二维大模型检测人体关键点，再通过三角测量或运动恢复结构（SfM），将二维关键点序列重建为三维动作轨迹，最终驱动三维人体模型运动。游戏中的动作捕捉，如体感游戏、电影中的角色动画生成、康复医学中的运动姿态评估。

场景理解与三维重建

仅通过RGB相机，视觉计算可实现“场景重建+物体识别+关系推理”的端到端处理。先通过特征匹配（如ORB）找到不同帧间的对应点，再通过SfM生成场景的稀疏点云，最后通过泊松重建生成稠密三维模型。同时，用预训大模型识别场景中的物体，并推理物体间的空间关系，如“杯子在桌子上”。

当前三维资产生成算法可生成大量真实三维数据，为三维大模型的训练提供支撑，未来视觉计算对三维数据的理解与生成能力将进一步提升。

5.7 视觉计算范式变革

从传统到深度学习的范式变革10-20年前，生物医学图像处理依赖传统方法，如手工特征+SVM分类，需为每个任务设计专属特征，如用GLCM 提取纹理特征区分正常与病变组织。如今，深度学习范式彻底改变了这一流程。无需手工设计特征，通过预训大模型直接提取语义特征。少量任务数据即可实现模型微调，如用100张肿瘤切片微调CLIP，即可实现肿瘤分割。端到端任务处理，如用GPT-4o直接输入医学影像与提示“分割肝脏肿瘤”，无需中间步骤。

这种范式变革的核心是“大模型的泛化能力+数据驱动的特征学习”，大幅降低了视觉任务的技术门槛，同时推动生物医学图像处理向“高精度、低数据依赖”方向发展。

5.8 实例：虚拟染色图像生成幻觉

生成式模型的一个已知缺陷是易产生“幻觉”，即生成并非基于真实数据或现实的信息/ 图像。这类幻觉的表现形式多样，既可能是细微的结构或颜色不一致，也可能是完全虚构的内容。在虚拟组织病理学与虚拟染色领域，这些幻觉可大致分为两类：（1）“技术性非真实幻觉”；（2）“拟真幻觉”。由技术问题引发的非真实幻觉包

括模糊区域、组织折叠区域或异常染色区域，这类幻觉与传统玻璃切片病理学中常见的伪影类似，通常会被专业的组织技术人员或病理医师在图像检查中发现。然而，第二类“拟真幻觉”的外观极具“真实感”，例如，部分组织成分被替换为虚构的成分。这类拟真幻觉可能误导病理医师，使其对实际组织标本中不存在的特征做出诊断，尽管从组织染色质量来看，这些幻觉特征显得真实可信。此类拟真幻觉可能导致多种问题：如曲解组织特征、干扰诊断流程、改变肿瘤分级、影响治疗反应预测等。

目前，已有多种基于深度学习的工具被开发用于检测标准组织化学染色病理切片中的技术性或非真实伪影，但这些工具的准确率参差不齐，且质量保障仍需依赖“人在环路”的持续检查。更重要的是，这些工具均未针对“检测虚拟染色模型引发的幻觉”进行专门训练或评估。目前尚无可用工具能评估并检测虚拟染色组织图像中那些“真实可信的拟真幻觉”。

AQuA识别生成式虚拟染色模型所产生的形态学伪影与幻觉 [18]，既包括技术性/ 非真实幻觉，更能检测至关重要的拟真幻觉。使用人类肾脏与肺部活检样本的虚拟染色苏木精- 伊红（H&E）组织切片集，对AQuA 网络（AQuA-Net）进行了训练、验证与测试。通过特定的数据构建方法，在肾脏与肺部组织数据集中引入了虚拟染色图像中可能出现的各类形态学幻觉与误差模式。借助创新的架构设计，AQuA-Net 无需任何真实参考图像（金标准）信息，就能自动检测这些形态学幻觉与低质量虚拟染色图像。

构建了基于无标记自体荧光（AF）的虚拟组织染色模型（即VS 模型：AF→H&E），并开发了其对应的反向图像转换模型，虚拟自体荧光模型（VAF 模型：H&E→AF）。通过在H&E 与AF 图像域间进行虚拟迭代，并将迭代结果作为AQuA-Net 的输入，AQuA 可在不获取组织化学染色H&E 真实图像（金标准）的情况下，自主检测虚拟染色H&E 图像中的伪影与幻觉。

VS网络架构与训练方案

针对肾脏和肺部组织样本的虚拟染色（VS）神经网络，采用生成对抗网络（Generative Adversarial Network, GAN）架构 [?]。该架构包含两个深度神经网络：VS生成器网络和判别器网络。

生成器网络：采用5层U-net结构 [?]，学习从输入自体荧光（AF）图像到对应明场苏木精- 伊红（H&E）染色图像的统计转换关系；

判别器网络：基于标准卷积神经网络分类器架构，旨在区分生成器生成的VS H&E 图像与真实组织化学染色（HS）H&E图像。

通过竞争性训练过程，生成器与判别器交替优化，逐步提升各自性能。

VAF网络架构与训练方案

肾脏和肺部样本的虚拟自体荧光（VAF）网络与VS生成器网络采用相同的5层U-

net 结构，仅输入与目标图像相反：VAF网络以H&E染色图像为输入，将其转换为对应的AF图像，与VS 生成器模型功能镜像。

VS模型的“优质/劣质染色”定义

在VS模型训练过程中，保存每个轮次（e）后的模型检查点及其对应的验证损失（l），验证损失的计算形式与训练损失一致，但基于验证集数据。

收敛判定：当验证损失在100个及以上轮次内无显著变化时，判定模型完全收敛；

阈值确定：基于验证损失曲线的波动情况，确定收敛起始点对应的经验阈值(e_0, l_0)（e_0 为轮次阈值，l_0 为验证损失阈值）；

灰色区域划分：为避免该区域内检查点的质量歧义，划分“灰色区域”：- 肾脏组织VS模型检查点：灰色区域为$e \in [e_0 - 50, e_0 + 50]$ 且$l \in [0.97l_0, 1.03l_0]$；- 肺部组织VS模型检查点：灰色区域为$e \in [e_0 - 100, e_0 + 100]$ 且$l \in [0.97l_0, 1.03l_0]$；

模型分类：

- 灰色区域外，满足$e < e_0$ 且$l > l_0$ 的检查点：标注为“劣质染色VS模型”；- 灰色区域外，满足$e > e_0$ 且$l < l_0$ 的检查点：标注为“优质染色VS模型”。

选择方法基于验证损失指标（包括平均绝对误差、SSIM、总变差），是图像生成与转换领域的常用方法。由具备执业资格的病理医师进一步验证了模型定义的正确性。补充图12 展示了肾脏和肺部组织VS模型的阈值设定与检查点选择过程。

AQuA网络架构与循环序列定义

AQuA网络的输入起始于组织染色域与自体荧光（AF）域之间的T次VS-AF循环。设初始AF测量图像为x_0、给定虚拟染色（VS）图像为y_0，输入序列$(x_0, y_0, \ldots, x_{T-1}, y_{T-1})$通过将$x_t$ 与y_t 分别迭代输入VS网络G_{VS}和VAF网络G_{VAF}生成，具体公式如下：

$$\begin{cases} x_{t+1} = G_{VAF}(y_t), & t = 0, \ldots, T-2 \\ y_t = G_{VS}(x_t), & t = 1, \ldots, T-1 \end{cases} \tag{5.1}$$

其中，G_{VS}与G_{VAF}始终采用预先确定的、固定的“最优检查点模型”（即验证损失最低的模型），该模型与输入序列(x_0, y_0)相互独立。

需特别说明：对于人类肺部组织切片的组织化学染色（HS）图像数据集，无法获取真实AF测量图像。因此，我们将VAF循环序列向后偏移一个时间步长，并以$G_{VAF}(y_0)$替代真实AF 测量图像。此时，用于HS图像的AQuA网络输入序列为$(y_0, x_0, \ldots, y_{T-1}, x_{T-1})$，定义如下：

$$\begin{cases} x_t = G_{VAF}(y_t), & t = 0, \ldots, T-1 \\ y_{t+1} = G_{VS}(x_t), & t = 0, \ldots, T-2 \end{cases} \tag{5.2}$$

AQuA网络特征提取与分类模块

AQuA网络的预训练骨干网络基于在ImageNet-1K数据集上训练的ResNet-50模型 [?]，包含该模型的所有残差块与二维平均池化层。特征提取与分类流程如下：

- **空间特征提取**：将T幅VS图像序列与T幅AF图像序列（含虚拟AF 图像与真实AF 测量图像）分别输入预训练骨干网络，提取空间特征$f_s \in \mathbb{R}^{B\times 2T\times C}$，其中$B$为批大小（batch size），$C$为二维平均池化层输出特征的通道数；

- **时空特征融合**：通过两个含1×1卷积核与ReLU激活函数的时间卷积层，对循环过程中的时间信息进行整合，生成沿时间轴压缩后的时空特征$f_{st} \in \mathbb{R}^{B\times C}$；

- **分类预测**：由ReLU激活函数连接的两个全连接层将时空特征f_{st} 映射为“优质染色”（正类）与“劣质染色”（负类）的预测对数概率。

5.9 小结

视觉计算的发展，本质是“特征表征能力”的不断进阶：从传统手工特征的“边缘、角点”，到深度特征的“语义对齐”，再到三维特征的“跨模态适配”，每一步都围绕“更高效、更鲁棒地描述视觉信息”展开。

从词袋模型到深度学习，图像表征与分类的核心进步在于“从人工设计到数据驱动，从低层特征到语义表达，从孤立单元到上下文关联”。词袋模型作为早期范式，奠定了“特征量化- 聚合”的基础，但因缺乏语义与上下文能力，难以应对复杂视觉任务；而深度学习通过自动学习层次化语义特征、建模全局上下文，彻底突破了这一局限，尤其预训大模型的出现，进一步降低了分类任务的训练成本，仅需微调少量参数，即可利用大模型的泛化能力实现高精度分类。未来，随着度量学习、多模态学习的发展，图像分类将向“更细粒度、更低数据依赖、更强语义关联”的方向持续演进。

表示学习、度量学习与迁移学习共同构成了深度学习的“特征优化-知识复用”体系：表示学习解决“如何将原始数据映射到判别性空间”；度量学习解决“如何定义相似度测度，优化样本在空间中的分布”；迁移学习解决“如何复用预训知识，降低新任务的训练成本”。

扩散模型的理论驱动、VAE 的分布学习，到图神经网络的非欧式数据处理，不同生成模型与表征学习工具各有侧重。在实际任务中，例如医学图像处理，需根据数据特性，欧式/ 非欧式、任务需求，生成/ 修复/ 表征选择适配的模型，甚至结合多模型优势，例如扩散模型+ VAE 的隐空间提升性能。

在生物医学图像处理中，通过预训大模型的特征迁移，如DINOv2提取医学图像特征、度量学习的样本优化，如三元组损失区分病灶/正常组织，可在数据稀缺的临床场景中实现高精度任务，如分割、配准、诊断，是当前医学AI 落地的核心技术路

径。同时，伦理问题的凸显也提醒我们：技术进步需与公平性、责任感同行，只有解决数据集偏差、模型透明性等问题，视觉计算才能真正赋能生物医学、自动驾驶等关键领域，实现“让机器更好理解世界”的终极目标。

第六章　图像分割

生物医学图像分割是临床诊断与科研的基础任务，核心目标是**从多模态医学影像中精准提取感兴趣结构**，为后续的疾病诊断、治疗规划、疗效评估提供量化依据。当前大模型主导的分割范式已实现“交互提示驱动”的高效流程：无需大规模标注数据，仅需提供简单提示。文本提示如“分割肝脏肿瘤”、点提示、包围框提示，大模型（如SAM、MedSAM）即可输出高精度分割结果。例如在胸部CT分割中，输入“分割右肺上叶磨玻璃结节”的文本提示+结节区域的点提示，模型可自动生成结节的像素级掩码，精度接近专家标注。

6.1　生物医学图像分割：从传统方法到多模态大模型

生物医学图像分割的发展与人工智能、计算机视觉的技术迭代同步，可清晰划分为“传统方法→ 机器学习方法→深度学习方法→大模型方法”四个阶段，且当前各方法仍在不同场景中并存。

传统分割方法：基于形态学与能量优化

传统方法依赖手工设计规则，核心是“通过图像的低层特征划分区域”，适用于简单场景或作为辅助工具：依赖人工参数调整，对模糊边界、复杂结构的分割精度低，泛化能力差。

基础形态学方法通过腐蚀、膨胀、开运算、闭运算等算子，去除噪声、填充孔洞，提取目标区域，例如，对CT图像进行阈值分割，设定灰度阈值区分骨骼与软组织后，用膨胀算子填补骨骼边缘的微小间隙。

活动轮廓模型（Snakes模型）通过定义能量函数，如轮廓内部能量、外部能量，让初始轮廓在能量最小化过程中逐步贴合目标边界。

机器学习分割方法：从监督到无监督

深度学习普及前，机器学习方法通过“手工特征+分类器”实现分割，将分割转化为像素/区域的分类问题：

- **监督学习方法**：用手工设计的特征描述子，如LBP纹理特征、灰度共生矩阵提取图像特征，再通过决策树、随机森林、SVM等模型，将像素/超像素映射到

“目标/背景”标签，例如用SVM 分类病理切片的像素，区分肿瘤细胞与正常细胞。

- **无监督学习方法**：无需标注数据，通过聚类算法将“表观相似的像素”归为一类，例如基于CT图像的灰度分布聚类，自动划分肺、心脏、骨骼等区域。

手工特征的语义表达能力弱，无法捕捉复杂医学结构的形态差异，如不同患者的肝脏形状变异，对噪声与伪影敏感。

深度学习分割方法：端到端的语义学习

深度学习彻底改变了分割范式，通过“端到端模型自动学习语义特征”，实现复杂结构的高精度分割：架构以U-Net及其变体为代表，采用“编码器- 解码器”对称结构。编码器通过卷积层提取图像的高层语义特征，如肿瘤的形态特征。解码器通过转置卷积恢复图像分辨率，结合编码器的低层特征，如边缘信息，生成像素级分割掩码。

训练：需要“图像-分割掩码”成对的标注数据，通过交叉熵损失、Dice 损失等优化模型参数，使预测掩码与真实掩码尽可能一致；可处理多模态、多器官分割任务，如全身CT 的多器官同时分割，对小病灶，如肺部微小结节的识别精度远超传统方法；依赖大规模高质量标注数据，但医学标注成本极高，泛化能力受限，如CT 上训练的模型无法直接用于MRI。

大模型分割方法：零样本/小样本的多模态适配

近年来，多模态预训练大模型的出现，解决了深度学习“数据依赖”的痛点，实现“零样本分割”：复用大模型在海量自然图像+文本数据上学习的“图文对齐”与“通用视觉特征”，通过少量医学数据微调或直接提示，适配医学场景；无需训练专属模型，可直接处理多模态医学影像，甚至无需医学标注即可完成常见结构分割。以“SAM+CLIP”联合分割方法中用SAM对医学图像进行初步分割，生成多个候选区域掩码。将候选掩码与文本提示输入CLIP，通过图文特征对齐，筛选出与文本匹配的目标掩码。以筛选出的掩码为提示，再次输入SAM，优化边界精度，得到最终分割结果。

6.1.1 分割工具的“方法并存”

无论是开源软件，如3D Slicer、ITK-SNAP，还是商业软件，均同时提供传统方法与深度学习工具，满足不同场景需求：**传统工具的应用**：在快速预览、手动修正时使用，例如在ITK-SNAP 中，先用阈值分割初步划分CT的硬组织区域，再用“区域增长”工具扩展边界，最后用“画笔”手动修正分错的像素。**深度学习工具的应用**：在批量处理、高精度分割时使用。传统方法操作灵活、无数据依赖，适合个性化调整；深度学习/大模型方法精度高、效率高，适合标准化任务。

6.2 生物医学图像的多模态特性与成像机制

医学图像的多模态差异源于“成像机制不同”，这也是分割模型泛化能力的核心挑战，不同模态的表观差异极大，需针对性适配。

表 6.1: 医学影像模态特点与分割任务对比

	成像机制	图像特点	典型分割任务
CT	X射线透射成像，通过组织密度差异成像	骨骼/高密度组织显影清晰，软组织对比度低	肺部、肝脏、骨骼分割
MRI	核磁共振，通过氢原子在磁场中的共振信号成像，含T1、T2、DWI等序列	软组织对比度高，可区分肿瘤/正常组织	脑肿瘤、脊髓、关节软骨分割
超声	超声波纵波传播，通过组织反射/散射差异成像	实时动态成像，受伪影（如气体）影响大	胎儿、心脏、甲状腺分割
病理切片	组织染色后显微镜成像，反映细胞级结构	分辨率极高，视野小，需拼接	肿瘤细胞、免疫组化区域分割
PET	注射放射性药物，通过正电子湮灭产生的光子成像，反映组织代谢活性	功能成像，可定位高代谢病灶（如肿瘤）	肿瘤分期、转移灶分割
SPECT	单光子发射成像，与PET类似但分辨率较低，常用于心血管、骨骼成像	功能成像，辐射剂量低	心肌灌注区域、骨骼病变分割

电子显微镜（EM）：高分辨率的电子束成像

电子显微镜通过**电子束替代可见光**，突破光学衍射极限，实现纳米级分辨率成像，可观察亚细胞结构（如线粒体、核糖体），是细胞生物学、病理学研究的关键工具；样品需真空环境，无法观察活细胞，制备过程复杂，如脱水、包埋，成像成本高。电子束生成依赖电子枪加热至高温，电子溢出形成电子云，施加高压（通常数

十万伏）加速电子，生成高速电子束；在交互与成像中，电子束与生物组织切片样品薄片交互，通过透射、散射等作用形成信号，再经电磁透镜放大，最终在探测器上生成图像。

光场显微镜（Light Field Microscopy）：三维体成像

以共聚焦光场显微镜为代表，结合共聚焦层析与光场多视角信息，实现快速三维体积成像：无需机械切片即可获取三维结构，如活体细胞的动态变化，兼顾分辨率与成像速度；传统共聚焦需逐点扫描，成像慢；光场显微镜通过微透镜阵列一次性记录多视角信息，成像效率提升1-2个数量级。激光光源提供单色相干光，扫描振镜控制光束扫描样品，针孔过滤非聚焦光，提升分辨率，微透镜阵列记录光线角度信息。在成像中激光激发样品荧光，荧光信号经针孔过滤后，由微透镜阵列将“不同角度的光线”映射到传感器，通过算法重建三维体图像。

多模态表观差异的影响 无论是EM、光场显微镜，还是CT、MRI，不同模态的**成像机制差异**导致表观完全不同。**EM图像**以“电子散射强度”为信号，呈现黑白对比的亚细胞结构。**光场显微镜**以“荧光信号”为核心，彩色标记特定分子，如荧光蛋白标记的细胞膜。这种差异使得“一种模态训练的分割模型无法直接迁移到另一种模态”，例如EM上训练的细胞分割模型，无法识别光场显微镜下的荧光标记细胞，这也是传统分割方法泛化能力弱的原因。

在大模型出现前，模型的多模态泛化能力极差：CT上训练的肝脏分割模型，直接用于MRI时，因MRI的软组织对比度与CT的密度对比度完全不同，分割精度会骤降。即便是同模态的不同参数，如常规CT与低剂量CT，模型泛化能力也受限，低剂量CT 的噪声更大，常规CT 训练的模型会将噪声误判为病灶。大模型通过海量多模态数据预训，大模型可学习跨模态的通用特征，例如“肝脏的形态特征”在CT与MRI 中一致，从而实现多模态泛化。

6.2.1 生物医学图像分割的未来方向

轻量级适配器（Adapter）优化：在大模型基础上，设计小规模适配器，用少量医学数据微调，进一步提升模型对特定模态、特定疾病的适配能力。**多模态融合分割**:结合不同模态的优势，提升复杂病灶的分割精度。**交互式分割的智能化：**结合GPT等大语言模型，自动生成医学提示，如从影像报告中提取“右肺上叶磨玻璃结节”作为文本提示，无需人工输入提示，实现“全自动分割”。

6.3 传统生物医学图像分割方法

传统分割方法依赖手工设计规则，基于图像的低层特征，例如灰度、边缘、纹理，实现分割，适用于简单场景或辅助工具，但其性能受限于“人工参数与特征的

表达能力”。

阈值分割：基于灰度分布的二分类

阈值分割是最基础的分割方法，通过灰度阈值区分目标与背景，适用于目标与背景灰度差异显著的场景：计算速度快，无需复杂模型；但仅适用于“单峰-单峰”灰度分布，无法处理目标内部灰度不均的情况，对噪声敏感。**阈值选择方法**包括：

- 直方图谷底法：统计图像灰度直方图，目标区域对应一个波峰，背景对应另一个波峰，阈值取两波峰之间的谷底，实现二分类。
- 最大熵法：选择“分割后图像信息熵最大”的阈值，熵越大，目标与背景的灰度区分越显著，适用于灰度分布不均匀的图像，如病理切片。
- 迭代阈值法：初始设定阈值，计算阈值两侧的平均灰度，以新平均灰度的均值更新阈值，重复至阈值稳定，适用于动态调整场景。

区域增长：基于像素相似性的扩展

区域增长通过“种子点+相似性准则”实现分割，从种子点出发，将表观相似的像素纳入目标区域：可用于CT 肺部分割；但依赖种子点选择，目标区域需具备均匀的表观属性，如肿瘤内部的坏死区域会导致分割中断，对噪声与伪影敏感。首先人工或自动选择种子点，如大脑MRI中手动标记的脑组织中心点。定义相似性准则，如灰度差小于阈值、纹理特征一致。进行迭代扩展，将种子点的邻域像素中“满足相似性准则”的像素加入目标区域，直至边界不满足准则的像素停止。

边缘检测：基于梯度突变的边界提取

边缘是目标与背景的“灰度突变区域”，边缘检测通过算子提取边界，再连接边界形成目标轮廓：可用于超声图像的边缘清晰，灰度差异显著胎儿轮廓提取；仅能获取边界，无法直接得到闭合区域，需额外的边缘连接算法，对模糊边缘提取精度低。图像平滑：用高斯滤波器去除噪声，避免噪声导致的虚假边缘。其次进行梯度计算，用Sobel、Prewitt等算子计算x/y方向梯度，得到梯度幅值/边缘强度与方向/边缘走向。最后进行阈值筛选，设定高/低阈值，保留强边缘、连接弱边缘。

主动轮廓模型（Snakes模型）：能量驱动的轮廓演化

主动轮廓模型通过“能量函数驱动初始轮廓贴合目标边界”，适用于复杂形状的分割：可处理不规则形状，对初始轮廓位置不敏感；依赖能量参数调整，内部/外部能量权重，计算复杂度高，对噪声敏感。

能量函数构成：

- 内部能量：控制轮廓的平滑性与连续性
  - 弹性能量：限制轮廓的拉伸程度，避免过度变形。
  - 弯曲能量：限制轮廓的弯曲程度，确保轮廓光滑。
- 外部能量：驱动轮廓向目标边界运动
  - 梯度能量：图像梯度越大（边界区域），外部能量越小，轮廓被“拉向”边界。
  - 用户交互能量：人工指定约束点，引导轮廓演化。

扩展模型包括：

- 梯度向量流（GVF）：增强对弱边界、凹陷区域的捕捉能力。
- 水平集方法：将轮廓表示为“水平集函数的零水平集”，支持轮廓的拓扑变化，如分割多根牙齿时，轮廓可从单一根变为多根。

条件随机场（CRF）：像素关系的后处理优化

CRF常作为深度学习分割的“后处理模块”，通过建模“像素间的关系”优化分割结果：深度学习分割后的边界优化，如FCN分割病理切片后，用CRF消除“孤立的错误像素”。弥补深度学习分割的“局部不一致”问题，提升边界精度；计算复杂度随图像尺寸增加而上升，不适用于超大规模图像，如全切片病理图像。通过最小化总势能函数，得到最终的像素标签；将图像的像素/超像素视为图的节点，相邻节点间建立边。势能函数定义：

- 一元势能：基于深度学习模型的初始预测，如FCN的像素类别概率，确保优化后标签与初始预测一致；
- 二元势能：基于相邻节点的表观相似性，如灰度、纹理，相似像素赋予相同标签，实现分割结果的平滑化；

6.4 深度学习分割方法：从监督到无监督的语义学习

深度学习通过“端到端模型自动学习语义特征”，彻底突破传统方法的局限，成为生物医学图像分割的主流技术，核心包括“监督分割”“无监督分割”与“专用模型适配”。

监督分割：U-Net及其变体的编解码架构

U-Net是生物医学图像分割的“基准架构”，采用对称的“编码器-解码器”结构，适配多模态、多器官分割任务：可处理多模态、多器官分割，对小病灶的识别精度远超传统方法；依赖大规模高质量标注数据，医学标注成本极高，泛化能力受限。其中编码器（Encoder） 由卷积层、池化层构成，逐步降低图像分辨率，提取高层语义特征。解码器（Decoder） 由转置卷积层构成，逐步恢复图像分辨率，结合编码器的“低层特征”，生成像素级分割掩码。长连接（Skip Connection）将编码器的某一层特征直接连接到解码器的对应层，补充低层细节，解决“分辨率恢复时的特征丢失”问题。训练需要“图像-分割掩码”成对的标注数据，通过Dice损失、交叉熵损失优化模型参数。变体模型包括：3D U-Net：处理三维图像，建模三维空间关系。Attention U-Net：在编码器与解码器间加入注意力机制，聚焦目标区域，提升分割精度。

无监督分割：谱聚类与深度谱方法

当缺乏标注数据时，无监督分割通过“数据自身的表观相似性”实现聚类，深度谱方法是当前的主流方向：无标注的病理切片分割；无需标注数据，降低成本；聚类精度低于监督方法，对特征嵌入的质量依赖高，特征提取模型需具备强判别性。**深度谱聚类**计算方法如下：

- 特征嵌入：用深度模型提取图像的高层特征，替代传统手工特征。
- 图构建与谱分解：将特征嵌入构建为图，计算图拉普拉斯矩阵，对矩阵做特征值分解，得到“谱嵌入”。
- 聚类：基于谱嵌入做K-Means聚类，相似特征的像素被归为一类。

谱嵌入的“费德勒向量”，最小非零特征值对应的向量，可直接区分前景与背景，仅需单通道即可实现二分类；多类聚类可通过增加特征值数量实现，适用于多结构分割。

Cellpose的流场

Cellpose是针对“细胞分割”的专用模型，突破传统“概率掩码”输出模式，通过“流场”实现高精度细胞边界分割：无需手动设计细胞形态先验，可分割不规则形状的细胞。Cellpose具有流场输出：

- 模型输出“像素相对于细胞中心的位移场”/流场，而非传统的类别概率。
- 流场的“收敛点”对应细胞中心，流场的方向指向中心，通过“流场积分”可确定每个像素所属的细胞，自然形成闭合边界。

初始自动分割后，人工校正错误区域，如合并的细胞、漏分的小细胞，校正数据可用于模型微调，提升泛化能力，适配不同类型的细胞图像。

生成模型在分割中的应用

生成模型如GAN、扩散模型通过“对抗学习”或“图像翻译”逻辑，为分割任务提供新的解决方案，将分割视为‘图像到标签’的生成问题。

GAN的对抗损失：提升分割精度

GAN在分割中的核心作用是“通过对抗损失迫使生成的分割掩码与真实掩码一致”，典型架构如下：可改善分割掩码的细节（如消除锯齿状边界），提升小病灶的分割完整性；训练不稳定（易模式崩溃），需平衡生成器与判别器的能力。

- 生成器（Generator）：即分割模型，输入图像，输出分割掩码；
- 判别器（Discriminator）：输入“真实掩码”或“生成掩码”，判断其真实性；
- 损失函数：
  - 分割损失：确保生成掩码与真实掩码的像素级一致性。
  - 对抗损失：迫使判别器无法区分生成掩码与真实掩码，提升生成掩码的“真实感”，如边界光滑、区域完整。

图像翻译：分割与生成的双向性

GAN的“图像翻译”能力可直接用于分割，将分割视为“从图像到标签的单向翻译”。

- 单向翻译（分割任务）：生成器学习“图像→分割掩码”的映射，输入自然图像/医学图像，输出对应的分割结果。
- 反向翻译（生成任务）：生成器学习“分割掩码→图像”的映射，输入分割掩码，生成对应的自然图像/医学图像。

在图像分割任务中，无标注数据时，用GAN生成“伪分割掩码”，辅助监督学习。在图像生成任务中，模拟不同疾病阶段的医学图像，如肿瘤增大后的CT图像，用于模型训练数据扩充。实现分割与生成的联动，解决标注数据稀缺问题；生成的分割掩码/图像可能存在“伪影”，需结合真实数据验证。

6.4.1 生成模型在分割中的核心要求：结构一致性与条件生成

生成模型用于图像分割时，需满足**“结构严格一致”**，这一要求在生物医学图像中尤为关键。

GAN将分割视为“从输入图像到分割掩码的单向图像翻译”，要求**源图像与目标分割图的结构完全对齐**。

- 自然图像分割中，输入图像的物体轮廓需与分割掩码的区域边界完全匹配，不能出现“结构错位”。
- 医学图像中，该约束更为严格，例如用GAN实现“CT→MRI”跨模态翻译时，肺、心脏等解剖结构的位置、形态必须与原图一致；若用于分割，则肿瘤的空间位置、大小需与原图完全对应，否则分割结果无临床意义。

无监督GAN分割需通过设计“结构保持损失”，如感知损失、结构相似性损失，避免生成器因对抗训练过度追求“像素相似”而忽略结构一致性。

扩散模型的分割范式：条件驱动的去噪生成

扩散模型通过**“条件去噪”**实现分割，核心是将“输入自然图像”作为条件，引导噪声逐步生成目标分割图：

- 扩散模型的基础逻辑：从纯高斯噪声出发，通过多步去噪过程学习“真实分割图的分布”，训练时用大量“分割图-噪声图”对优化。
- 分割时的条件注入：输入新自然图像后，将其特征作为“条件信号”融入每一步去噪过程，迫使生成的分割图与自然图像的结构匹配。

相比GAN，扩散模型生成的分割图更稳定，不易模式崩溃，且可通过调整条件强度控制分割精度，如增强自然图像特征的权重，提升边界对齐度。

6.4.2 深度学习的算力与数据挑战：从“小模型”到“大模型适配”

SAM（Segment Anything Model）是分割大模型的代表，其核心是**“可提示交互+强泛化能力”**，适配医学场景需针对性调制：

SAM的核心架构与提示工程

SAM的架构围绕“图像编码-提示编码-解码”三层设计，实现零样本分割：

- **图像编码器**：基于ViT，提取输入图像的全局语义特征，如自然图像中的“物体轮廓”、医学图像中的“器官形态”；

- **提示编码器**：将用户交互提示（点、包围盒、文本）编码为特征向量，例如文本“肝脏肿瘤”编码为语义向量，点提示“点击肿瘤中心”编码为空间向量；

- **解码器**：融合图像特征与提示特征，生成像素级分割掩码，支持多类别同时分割，如一次分割CT中的肝脏、肿瘤、血管。

MedSAM的调制过程

自然图像预训的SAM直接用于医学分割时，因“模态差异”，如CT的密度特征与自然图像的纹理特征不同性能有限，需通过以下方式调制：

- **大规模医学数据微调**：用CT、MRI、超声等多模态医学图像及标注，微调SAM的图像编码器与提示编码器，使其学习医学语义；

- **少样本调制**：若标注数据稀缺，如仅含少量包围盒标注的病理切片，可通过“提示扰动”优化，对现有提示（如包围盒）做微小调整，如缩放、平移，生成伪标注数据，仅微调提示编码器，即可提升分割精度；

调制后的Medical SAM可支持“零样本医学分割”，如输入文本提示“分割脑转移瘤”+CT图像，无需训练即可生成转移瘤掩码。视觉语言大模型通过“图文对齐”能力增强分割性能，复用预训大模型特征，仅训练小模块。

6.4.3 主流融合方案

跨模态注意力对齐：将医学图像，输入预训视觉编码器，将医学报告，如“ 右肺上叶磨玻璃结节”，输入预训文本编码器，如CLIP 文本端，通过“跨模态注意力层”对齐图文特征，仅训练该注意力层，即可实现“ 报告引导的分割”，例如文本中的“ 磨玻璃结节”语义引导模型定位并分割对应区域；

视觉-语言图匹配：将图像分割掩码转换为图结构，像素为节点，边缘为相邻关系，将文本描述转换为语义图，如“肿瘤-血管”的关联，训练小模块实现两图匹配，优化分割掩码的语义一致性；

解剖结构-文本协调：引入解剖学知识图谱，如“肝脏-胆囊”的位置关系，训练小模块将文本描述，如“胆囊结石”与解剖结构特征对齐，避免分割时出现“结构错位”。

所有方案均遵循大模型不动，小模块可训的原则，预训视觉/文本编码器，如CLIP 的权重冻结，仅训练跨模态对齐层、图匹配层等小规模模块，参数通常仅数百万级，大幅降低训练成本。

6.5 实例：电镜亚细胞结构分割

纳米级细胞器分割在理解细胞结构的复杂形态和组织方面起着至关重要的作用 [27, 41, 51, 58]。高通量成像技术能够高效地提供细胞器结构的高分辨率成像。虽然人工分割仍是多种下游生物学任务的金标准，但其需要大量人力且高度依赖专家知识。基于Transformer的自监督学习技术 [6, 15]已展现出捕捉局部与全局空间关联的强大潜力，非常适合亚细胞结构的鲁棒性表征。基于深度神经网络的表征学习促进了显微镜图像的可扩展自动化结构分割 [14, 24, 51]。

聚类长期以来是一种基于特征相似性的强大无监督分割方法，可在空间域或谱域中实施。空间聚类方法通常通过识别局部像素相似性直接对提取的图像特征进行操作，但这类方法对显微镜图像中的高频噪声敏感，易导致次优分割结果。另一方面，谱聚类在谱嵌入空间中运作，对高频扰动具有更强鲁棒性。然而数值谱嵌入与聚类在处理大规模图时计算成本高昂 [13]，且需要额外的同步操作 [29] 以维持跨图像聚类分配的一致性。

协同空间与谱聚类学习的新型无监督双分支深度神经聚类框架CS²C借鉴自监督表征学习与深度嵌入聚类 [50]，利用从MAE模型提取的丰富语义特征进行空间聚类。针对显微镜图像特征中的高频扰动，我们采用深度谱嵌入并在低维谱嵌入空间进行聚类。通过结合聚类损失与空间-谱聚类分配的一致性正则化，端到端地训练双分支聚类模型。该方法推动了细胞器分割在空间与谱域的统一解决方案。

给定输入VEM图像，目标是通过训练聚类模型使其能够接收VEM切片I作为输入，并输出针对不同细胞器结构的聚类分配结果P。框架遵循基于特征相似性的特征提取与聚类推断主流范式，利用双分支空间-谱聚类模型实现细胞器分割。空间聚类分支直接对MAE 模型提取的丰富语义超像素特征执行聚类，而谱聚类分支则在图拉普拉斯矩阵近似谱基张成的低维空间中进行神经谱嵌入与聚类。引入端到端训练策略，将聚类损失与空间- 谱聚类分配间的一致性正则化相结合。该方法有效缓解显微镜图像特征中的高频扰动，并促进聚类的空间-谱一致性。

协同空间-谱神经聚类

框架核心包含两个并行聚类模型：空间模型与谱模型。它们基于超像素特征预测聚类分配，充分利用谱聚类对高频扰动的鲁棒性及其与原始特征空间中基于质心的概率分布的对齐特性。

谱聚类在谱分支中，细胞器分割被建模为谱嵌入空间内超像素图的切割问题。传统独立图谱聚类方法往往忽视图像间关联，导致聚类分配不一致。深度神经谱聚类模型利用其泛化能力确保跨图像标签一致性。基于MAE特征生成超像素图，其中亲和矩阵$A \in \mathbb{R}^{n_s \times n_s}$ 采用RBF 核定义：$a_{ij} = \exp\left(-\frac{\|F_i - F_j\|_2^2}{\kappa}\right)$（$1 \le i, j \le n_s$）。$\kappa$为核方差，$n_s$为超像素数量。进一步缩减带宽并消除弱关联值：$A \leftarrow \max\left(A - \frac{\max A}{\alpha}, 0\right)$，其中$\alpha$ 控制削弱弱连接的排斥力强度。

采用基于图卷积网络的谱嵌入模块近似图拉普拉斯矩阵L的谱基$\Phi \in \mathbb{R}^{n_s \times u}$，避免计算昂贵的数值特征分解中特征向量切换和符号翻转问题。u表示近似谱基数量。谱嵌入通过两个约束进行优化：(1)对称拉普拉斯矩阵L 的对角化；(2) 近似谱基Φ的正交性。谱嵌入损失函数：

$$\mathcal{L}_{\text{emb}} = \sum_{i \neq j} \left[(\phi_i^T L \phi_j)^2 + (\phi_i^T \phi_j)^2 \right], \tag{6.1}$$

其中ϕ_i和ϕ_j为Φ的第i和第j列向量。对称拉普拉斯矩阵L的特征分解满足$\Lambda = \Phi^T L \Phi$（Λ为特征值对角矩阵）。该损失确保$\Phi^T L \Phi$ 的非对角元素趋近于零，并强制Φ正交性。

引入神经聚类模块f_θ基于谱嵌入分配聚类标签，无需匈牙利匹配 [1, 29]等标签同步操作。聚类分配概率矩阵$P = \text{softmax}(f_\theta(\Phi))$利用抗扰动的谱嵌入定义超像素图的切割。

空域聚类在空间分支中，聚类直接应用于MAE特征。受深度嵌入聚类方法[50]启发，通过最小化软聚类分配与辅助目标分布间的距离来优化可学习聚类中心。软聚类分配矩阵Q采用学生t分布核函数 [45] 计算：

$$q_{i,j} = \frac{(1 + \|F_{s,i} - \mu_j\|/\tau)^{-(\tau+1)/2}}{\sum_{j'} (1 + \|F_{s,i} - \mu_{j'}\|/\tau)^{-(\tau+1)/2}}. \tag{6.2}$$

$\mu \in \mathbb{R}^{k \times 2l}$表示通过k均值初始化的可学习中心，$\tau$控制自由度。辅助目标分布$R$ 通过对Q进行频率引导归一化推导：

$$r_{i,j} = \frac{q_{i,j}^2 / \sum_k q_{k,j}}{\sum_{j'} (q_{i,j'}^2 / \sum_k q_{k,j'})}. \tag{6.3}$$

通过最小化软分配Q与目标分布R间的KL散度优化聚类中心：

$$\mathcal{L}_{\text{clu}} = \text{KL}(R \| Q) = \sum_i \sum_j r_{i,j} \log \frac{r_{i,j}}{q_{i,j}}. \tag{6.4}$$

最小化损失$\mathcal{L}_{\text{clu}}$可优化中心μ，提升聚类纯度并聚焦高置信度分配。

空-谱一致性正则化谱聚类因图拉普拉斯特征基近似的低通特性而对图像特征高频扰动具有鲁棒性。空间聚类应用于MAE特征时能有效迭代优化软聚类分配。为结合两种聚类优势，提出最小化其概率分配矩阵间距离。具体采用KL散度度量谱聚类矩阵P与空间聚类矩阵R 间的差异：

$$\mathcal{L}_{\text{reg}} = \text{KL}(P \| R) = \sum_i \sum_j p_{i,j} \log \frac{p_{i,j}}{r_{i,j}}. \tag{6.5}$$

相较于交叉熵损失，KL散度能以更温和的方式更新模型，有助于避免数据表征的突变。

总训练损失为神经谱嵌入损失$\mathcal{L}_{\text{emb}}$、聚类损失$\mathcal{L}_{\text{clu}}$ 和空间- 谱一致性正则化损失$\mathcal{L}_{\text{reg}}$ 的加权组合：

$$\mathcal{L} = \mathcal{L}_{\text{emb}} + \gamma_1 \mathcal{L}_{\text{clu}} + \gamma_2 \mathcal{L}_{\text{reg}}, \tag{6.6}$$

其中γ_1和γ_2是平衡正则化空间与谱聚类的超参数。最小化组合损失函数$\mathcal{L}$ 可同步优化神经谱聚类与空间聚类模块。

在线推理过程中，给定VEM图像I：首先通过MAE特征提取器生成超像素特征并构建超像素图；随后谱聚类模块以加权超像素图为输入，将各超像素分配至聚类，产生聚类分配矩阵P 从而实现细胞器分割。

6.6 小结

从传统方法到多模态大模型，各类分割技术并非相互替代，而是**按需并存、综合利用**，适配不同场景需求：

	优势	适用场景	典型应用
传统方法	操作灵活、无数据依赖	快速预览、手动修正、简单结构分割	ITK-SNAP 中的阈值分割、区域增长
普通深度学习	精度高、适配特定任务	标注数据充足的标准化任务	U-Net用于分割
生成模型	无监督/半监督、细节优化	标注稀缺、需边界平滑的分割	GAN优化分割边界
多模态大模型	零样本/少样本、跨模态泛化	标注稀缺、多模态医学图像分割	MedSAM 用于分割
跨模态融合	语义增强、降低标注依赖	需医学报告/解剖知识引导的分割	CLIP引导的MRI分割

表 6.2: 方法类型及其应用场景

主流实践范式

当前最常见的是“**预训大模型特征+小Adapter**”的综合范式：

- **复用预训大模型**，如SAM的视觉特征、CLIP的文本特征，获取强泛化性表征。
- 设计并**训练小规模Adapter**，如分割解码器、跨模态对齐层，适配特定任务。
- 可选**引入传统方法**，如用形态学算子修正分割掩码的微小漏洞，进一步提升精度。

图像分割经历了从“数据依赖”到“零样本适配”在图像分割任务中，技术演进可清晰划分为三个阶段，当前的核心突破是“预训大模型的表征复用”：

传统阶段：依赖形态学算子、阈值分割、传统机器学习，需手工设计特征（如LBP 纹理、灰度共生矩阵），仅适用于“目标与背景灰度差异显著”的简单场景。

深度学习阶段：以U-Net、FCN为代表的端到端模型，通过CNN自动学习语义特征，可处理复杂结构分割（如肿瘤分割），但需大规模标注数据，且泛化能力受限。

大模型阶段：基于预训大模型的零样本/小样本分割成为主流，无需训练专属模型，仅需输入简单提示即可生成高精度分割掩码。当前面对新分割任务时，优先用预训大模型测试基础性能；若精度不足，再用少量标注数据训练“轻量型Adapter”，无需重训大模型，大幅降低成本。

生物医学图像分割的技术演进，是“从人工规则到数据驱动，从低层特征到高层语义”的过程：传统方法适用于简单场景与辅助工具，深度学习方法通过端到端学习实现复杂结构的高精度分割，生成模型则为无标注场景与细节优化提供新路径。未来，随着多模态大模型与生成模型的融合，分割技术将进一步向“低标注成本、高跨模态泛化、高细节精度”方向发展，更好地适配临床与科研需求。

第七章　图像配准

医学图像配准、三维形状特征表示学习及基于深度学习的医学图像配准是医学图像分析中的重要研究方向。医学图像通常是空间变化的物理量映射，主要成像模态包括X线计算机断层摄影（CT）、核磁共振成像（MRI）、正电子发射断层影像（PET-CT）及单光子发射型计算机断层显像（SPECT）等。这些模态均具有一致的数字张量表示。

医学图像配准用于对齐不同视角、模态或时间点获取的同一场景或物体的多幅图像，在纵向研究中量化治疗或生长发育导致的形态变化，为智慧精准医疗提供关键技术支撑。其应用包括统计形状建模、多模态图像融合、属性迁移与自动标注等。

配准方法可分为刚性配准与非刚性（形变）配准两大类。刚性配准仅涉及全局的旋转（$\mathbf{R}$）和平移（$\mathbf{T}$）变换（六自由度），可用线性变换表示：$\mathbf{V}_t = \mathbf{V}_r \cdot \mathbf{M}$。形变配准则采用非线性稠密空间变换模型（形变场$\mathbf{W}$），通过对源图像施加变换$\mathbf{W}$ 实现与目标图像的对齐，其求解通常归结为优化以下能量函数：

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \left[\mathcal{D}(\mathbf{T}, \mathbf{S} \circ \mathbf{W}) + \lambda \mathcal{R}(\mathbf{W})\right]$$

其中$\mathcal{D}$ 为相似性测度，$\mathcal{R}$ 为正则化项。该问题常为病态问题，需结合适当的优化策略。

7.1　图像配准的概念：从自然图像到医学图像

7.1.1　配准的目标：建立“语义对应”

图像配准是计算机视觉与医学影像分析中的基础任务，其本质是在不同条件下的图像间建立结构的语义对应关系，并量化这种对应所需要的空间变换。这一任务在自然图像与医学图像中均有体现，但核心需求因场景差异而不同:

自然图像配准：需处理“物体姿态变化”“光照差异”等干扰，目标是找到不同图像中同一物体的对应结构，通常输出“位移场”“置换矩阵”或“对应概率矩阵”，用于目标跟踪、图像拼接等任务。

医学图像配准：是后续图像分析，如病变检测、疗效评估的前提，只有建立了

结构间的语义对应，才能有效研究图像间的解剖变化，如肿瘤体积增减、器官形态改变。若缺乏对应关系，直接对图像矩阵，做范数计算得到的“距离”无实际意义，因为其未考虑解剖结构的空间一致性。

7.1.2 医学图像配准的应用场景

医学图像配准的价值体现在多类临床与科研任务中，核心应用可归纳如下：

纵向研究与变化度量：对同一患者不同时间点的图像进行配准，量化解剖结构或病变的动态变化，如肿瘤直径缩小幅度。

多模态图像融合：不同模态的医学图像，如CT提供解剖结构、PET提供代谢信息需先配准，才能融合多源信息，为临床诊断提供更全面的依据。

图像属性迁移：建立“标注图像”与“未标注图像”的稠密对应后，可将标注图像中的结构标签迁移到未标注图像，实现“One-Shot自动标注”，降低人工标注成本。

统计模型构建：对多例患者的图像进行配准，将其映射到同一标准空间，构建解剖结构的统计模型，如正常人群的脑结构均值与方差。

手术导航与放疗规划：术前图像与术中实时图像配准，可将术前规划的靶点位置映射到术中场景，指导手术器械定位或放疗剂量投放。

7.1.3 医学图像配准的要素：变换模型与损失函数

医学图像配准求解“空间变换参数”，根据变换的灵活性，可分为刚性变换、仿射变换与非刚性变换三类，对应不同的解剖结构特性。典型的配准流程包含以下几个组成部分：变换模型定义图像之间的空间映射关系；插值方法用于在非网格位置估计图像强度；相似性测度评估变换后图像的对齐程度；优化算法寻找最优变换参数。

需注意非刚性变换的实现可通过“局部刚性变换组合”，如将图像划分为多个子块，每个子块做刚性变换或“稠密位移场”，直接预测每个像素/体素的位移，前者参数规模小，后者精度更高但计算复杂度大。相似性度量是评估配准质量的关键：

- 均方误差（MSE）：衡量像素强度的差异。

$$\mathrm{MSE}(I_1, I_2) = \frac{1}{N}\sum_{i=1}^{N}(I_1(x_i) - I_2(x_i))^2$$

- 互信息（MI）：衡量图像强度间的统计依赖性，适用于多模态配准。

$$\mathrm{MI}(I_1, I_2) = H(I_1) + H(I_2) - H(I_1, I_2)$$

其中$H(I_1)$, $H(I_2)$ 为各图像的熵，$H(I_1, I_2)$ 为联合熵。

表 7.1: 医学图像配准的变换模型分类与特性

变换类型	自由度/参数形式	特性	适用场景
刚性变换	6个自由度（3个平移+3个旋转）	保持结构的体积、形状与距离不变，仅做全局位置/角度调整	骨骼、植入物（如人工关节）的配准
仿射变换	12个自由度（刚性+3个缩放+3个剪切）	保持平行线不变，允许全局缩放与剪切，仍为线性变换	近刚性结构（如肺部大体位置配准，忽略局部呼吸形变）
非刚性变换（形变配准）	稠密参数（按像素/体素的位移场）	允许局部形变，可刻画复杂解剖变化（如呼吸导致的肺形变、肿瘤生长导致的组织挤压）	软组织的精细配准

- 归一化互相关（NCC）：对图像强度差异进行归一化，具有一定模态无关性。

医学图像配准问题本质是“带约束的优化问题”，损失函数通常由数据项与正则项两部分构成，二者共同保证配准结果的“准确性”与“合理性”：

数据项：保证配准的准确性，核心目标是使“经变换后的移动图像（Moving Image）”与“固定图像（Fixed Image）”在对应结构上一致。

正则项：保证配准的合理性。由于配准问题是“病态的”，同一数据项可能对应多个变换解，需通过正则项约束变换的物理合理性，避免出现“结构折叠”“不连续形变”等不符合解剖规律的结果，常用约束包括：**平滑性约束**：要求形如$\|\nabla T\|_2^2$的位移场的梯度较小，保证形变连续。**可逆性约束**：要求变换满足“微分同胚（Diffeomorphism）”，即变换可逆且雅可比行列式为正，避免结构拓扑变化，如脑沟回折叠。**保体积约束**：针对肝、肺等不可压缩器官，约束变换前后的体素体积不变。

7.2 医学图像配准的关键技术挑战

对于三维医学图像，配准面临巨大数据量、大尺度形变等挑战。空域测度和空间对齐受硬件限制，因此需引入特征提取与变换预测模块，以高效地建立图像对应关系。在纵向成像中，图像配准可用于监测组织或肿瘤的体积变化，尤其在肿瘤学中用于评估治疗响应。在脑影像研究中，连续MRI图像的配准有助于追踪神经退行性疾病，如阿尔茨海默病或帕金森病中的脑结构变化。在手术规划与导航中，图像

配准能够将术前影像与术中实时图像对齐，辅助外科医生定位病灶、识别关键解剖结构或规划手术路径，提升手术精准性与安全性。尽管多模态图像提供互补信息，如MRI显示解剖细节、PET反映功能代谢，但其强度分布差异显著，给诸如均方误差的相似性度量带来挑战。此外，配准过程计算复杂度高，尤其对于非刚性形变及三维图像，需依赖高效算法与强大算力。图像中存在的噪声和伪影也可能影响配准精度。医学图像配准面临多类独特挑战，这些挑战直接影响配准方法的设计与性能：

跨模态表观差异：不同模态图像的像素值无直接可比性，如CT中骨骼为高值，MRI中骨骼为低值，传统基于像素的相似性测度失效，需依赖互信息等模态无关测度。

大尺度与非刚体形变：软组织因呼吸、心跳或病理变化产生大尺度形变，需高精度非刚性变换模型刻画局部细节；现有网络对大尺度形变的刻画能力有限，需设计多尺度网络，如从粗到精的级联结构逐步优化形变场。

高维与大规模数据：医学图像常为三维体数据，直接处理会导致高计算复杂度，需优化算法效率。深度学习模型需大量标注数据，如真实形变场、Landmark等，而医学数据标注成本高；同时，模型在训练集分布外的数据上泛化性较差。

采样与插值问题：图像为离散像素/体素网格，变换后的数据可能落在非网格点上，需通过插值生成规则网格图像，插值精度直接影响配准结果。

物理合理性约束：需在网络设计中融入正则项（如微分同胚约束），避免出现不符合解剖规律的形变，如结构折叠，常用方法包括在损失函数中加入雅可比行列式正则项，或使用特殊网络结构保证变换可逆。

局部最优问题：端到端模型的前向传播输出的形变场缺乏传统迭代优化的“局部最优保障”，可能出现局部结构对齐问题。

7.3 医学图像配准的方法演进：从数值优化到深度学习

医学图像配准技术的发展历经三个关键阶段：传统数值优化方法（20-30年前）、传统机器学习方法（10 余年前后）、深度学习方法（当前主流）。三类方法虽技术路径不同，但目标一致，“建立图像间的精准语义对应”，且在“目标函数/损失函数设计”上存在共性。

7.3.1 方法演进中的共性：目标函数/损失函数的一致性

尽管三类方法的技术路径差异显著，但在“函数设计”上存在本质共性，均以“数据项保证准确性、正则项保证合理性”为原则，具体体现为：

传统数值优化方法的“目标函数”与深度学习方法的“损失函数”结构一致：均包含“数据项”与“正则项”，数据项保证配准的准确性，使变换后图像与固定图像对齐，正则项保证形变的合理性，符合解剖规律。

数据项一致：无论是传统方法的“互信息、NCC”，还是深度学习方法的“互信息损失、MSE 损失”，均是“度量变换后图像与固定图像的相似性”，相似性越高，数据项越小；

正则项一致：无论是传统方法的“平滑性约束、弹性势能约束”，还是深度学习方法的“位移场L2 正则、雅可比行列式惩罚”，均是“约束形变场符合物理/解剖规律”，避免不合理形变。

医学图像配准的核心矛盾始终是“准确性与合理性的平衡”，方法演进的本质是“用更高效的技术路径，如深度学习实现这一平衡”，而非改变配准问题的本质目标。未来方法的发展仍进一步融合“物理先验,如生物力学模型”与“数据驱动，如大模型”，实现更高精度、更高效率、更强泛化性的配准。

7.3.2 传统配准方法：基于优化的迭代策略

在深度学习方法兴起前，传统医学图像配准主要依赖“迭代优化框架”，流程为“初始化变换→计算损失→更新参数→收敛判断”。

传统数值优化方法（20-30年前）

基于物理模型或几何规则构建目标函数，通过迭代优化求解形变参数；定义“数据项（如互信息、NCC）+ 正则项（如平滑性约束）”构成目标函数，采用梯度下降、L-BFGS 等数值优化算法迭代更新形变参数，直至目标函数收敛。算法包括：demons配准（扩散模型）、弹性体配准（物理模型）、FFD配准（插值模型）；基于优化的迭代策略理论严谨、物理意义明确，但迭代次数多、依赖人工设计的相似性测度与正则项，泛化性差。

传统机器学习方法（10余年前后）

用机器学习模型替代部分人工设计模块，如特征提取、相似性测度，提升方法的适应性。通过手工设计特征，如SIFT、HOG或浅层模型，如随机森林、SVM学习图像间的相似性度量，或通过回归模型直接学习“图像对→形变参数”的映射。算法包括：基于随机森林的Landmark匹配、基于SVM的特征相似性判断、基于高斯过程的形变场回归。传统机器学习方法相比数值优化，泛化性有所提升，但仍依赖人工设计特征，难以处理复杂形变与跨模态场景。

基于物理模型的方法

将解剖结构视为具有特定物理属性的介质，通过物理方程刻画形变：

弹性体模型：基于胡克定律，将结构视为弹性体，通过求解线性/非线性弹性力学方程得到形变场，保证拓扑不变性。

流体模型：将形变视为粘滞流体运动，通过求解纳维-斯托克斯（Navier-Stokes）方程处理大尺度形变，适用于脑等软组织。

扩散模型：如demons 配准算法，将图像像素视为“魔鬼（Demons）”，通过计算“魔鬼力”迭代更新形变场，正则项采用高斯滤波保证平滑性。

基于插值的方法

通过“稀疏控制点+插值函数”降低形变场参数规模，核心是用少量控制点的位移推导全局稠密形变场：

薄板样条：以稀疏地标点为约束，构建低阶多项式样条函数，保证形变平滑且通过所有控制点。

自由形变模型：在图像外构建规则控制网格，通过调整网格节点位移，结合三线性/三立方插值得到图像内部的稠密形变场，适用于三维图像的非刚性配准。

径向基函数：如高斯RBF，以控制点为中心构建径向衰减的基函数，通过线性组合基函数生成全局形变场，灵活性高但计算成本随控制点数量增加而上升。

基于统计模型的方法

利用多例样本的统计信息约束形变，降低参数空间维度：

主成分分析（PCA）模型：对多例配准后的形变场做PCA，用前k个主成分构建形变子空间，仅需优化主成分系数即可表示形变，大幅降低计算复杂度。

基于生物力学的统计模型：结合解剖结构的生物力学特性，如骨骼的刚性、肌肉的弹性，构建统计约束，使形变符合生理规律。

7.3.3 深度学习驱动的医学图像配准方法

随着深度学习技术的发展，传统迭代优化方法的“计算效率低”“依赖人工设计特征”等问题被逐步解决。深度学习配准方法的核心是用神经网络直接学习“图像对→空间变换”的映射，实现端到端的配准，优势体现在：

效率提升： 训练完成后，测试阶段仅需一次前向传播即可输出形变场，无需迭代优化，适用于临床实时场景。

特征学习自动化： 无需人工设计相似性测度或特征描述子，神经网络可从海量数据中学习模态无关的配准特征。

多任务融合能力： 可将配准与其他任务，如图像分割、病变检测联合训练，利用多任务信息提升配准精度，如结合分割标签实现弱监督配准。

基于深度学习方法构建端到端学习“图像对→形变场”的映射，由模型自动学习特征与相似性测度。以CNN、Transformer等深度模型为核心，输入固定图像与移动图像，直接输出稠密形变场，损失函数仍保留“数据项+ 正则项，如平滑性损失、

雅可比行列式损失”的结构。基于深度学习方法实现端到端推理效率高、自动学习模态无关特征泛化性强，但需大量标注数据、模型可解释性差。

7.3.4 深度学习配准框架

目前主流的深度学习配准框架可分为“基于回归的端到端模型”与“基于特征匹配的模型”两类。近年来，基于深度学习的方法显著推动了三维医学图像配准的发展。这类方法通过卷积或变换器网络提取层次特征，构建从图像到形变场或速度场的端到端映射，支持解剖形态分析与生长建模。有监督方法利用真实标注，如结构对应或形变场优化网络参数，但受限于三维标注的高成本与个体间差异。无监督方法则借助空间变换网络实现可微配准，通过最小化图像差异进行优化，避免对标注数据的依赖，如VoxelMorph及其微分同胚变体实现了高效且拓扑保持的配准。

端到端回归模型 如VoxelMorph，以固定图像与移动图像为输入，通过编码器-解码器网络直接回归稠密位移场，损失函数结合数据项与正则项，训练过程为无监督，无需真实形变场标签。

特征匹配模型 先通过CNN提取图像对的深层特征，再通过注意力机制或图神经网络匹配特征点，最后基于匹配的特征点用插值生成形变场，适用于需显式特征对应关系的场景。

此外一些研究致力于改善计算效率与特征表示：PointNet和PointNet++通过点云层次划分学习局部与全局特征。OctNet和基于八叉树的卷积神经网络利用体素稀疏性降低计算开销。图卷积网络从三维几何中提取任务相关的形状描述符。然而，医学图像中多解剖结构共存，其非稀疏性限制了层次卷积方法的效率。因此，当前研究更侧重于从原始数据中学习鉴别性特征，并利用内在结构对应关系克服姿态和形态变异带来的干扰。DeSmooth 方法通过替换Sinkhorn层以增强置换矩阵估计，也有工作探索基于谱域映射的特征表示，以提升配准的鲁棒性与泛化能力。

7.3.5 弱监督与半监督配准：降低数据依赖

为解决医学图像分析领域中“标注数据稀缺”问题，弱监督与半监督配准方法已成为当前研究热点。通过“低成本标注信息”，诸如稀疏解剖学标志点、局部结构分割标签等辅助模型训练，在降低标注成本的同时保障配准性能：

- **弱监督配准**：以稀疏Landmark为核心约束条件，在模型损失函数中引入“变换后Landmark 与目标Landmark的距离损失项”，通过该约束引导网络学习图像间的正确对应关系。该方法无需获取稠密形变场标签，大幅降低了数据标注的难度与成本。

- **半监督配准**：采用“少量有标签数据+大量无标签数据”的混合训练模式，其中有标签数据通常为带有真实形变场的图像对，无标签数据为普通未标注图

像。模型通过迁移学习将有标签数据学到的知识迁移到无标签数据上，或通过一致性约束确保模型对同类数据的预测一致性，最终在“标注成本控制”与“配准精度提升”之间实现平衡。

基于深度学习的配准方法凭借数据驱动优势，通过卷积网络或Transformer网络构建端到端训练框架，能够从原始图像中自动学习具有判别性的最优特征表示，并直接输出用于图像变换的形变场或速度场。

- VoxelMorph：作为早期端到端卷积配准网络的经典方案，构建了“图像输入-形变场输出”的直接映射关系，简化了传统配准方法的多步骤流程，为后续深度学习配准研究奠定了基础。

- 半监督增强策略：在端到端框架中融入弱标注信息，通过“配准结果与分割标签的空间一致性约束”优化模型学习过程，进一步改善复杂解剖结构区域的配准精度。

- Transformer类模型：模型诸如TransMorph相较于传统卷积网络，Transformer的自注意力机制具备更大的有效感受野，能够更高效地建模图像全局上下文信息与跨区域语义对应关系，使配准过程中生成的形变场曲面更平滑，减少局部失真问题。

- 无监督对应评分机制：引入置信度评分网络，对配准过程中生成的特征对应关系进行自适应权重分配，对噪声干扰、组织遮挡等因素导致的误匹配区域降低权重，从而提升模型对复杂图像场景的鲁棒性。

- 级联结构模型：模型诸如LapIRN、IIRP-Net采用“从粗到细”的金字塔特征提取策略或迭代推理机制，先通过粗粒度网络学习全局形变趋势，再通过细粒度网络逐步优化局部位移场细节，在保证配准效率的同时显著提升定位精度。

近年来，以DINOv2、CLIP为代表的基础模型凭借大规模预训练获得的通用特征表示能力，为医学图像配准带来了全新技术范式，推动领域向“低标注依赖”“跨模态适配”方向发展。

- DINO-Reg方法：直接复用预训练DINOv2模型提取的通用视觉特征进行配准计算，无需针对特定配准任务进行模型微调，大幅缩短了模型训练周期，同时依托预训练特征的强泛化性，在多模态医学图像配准中表现出优异性能。

- DINO-Tracker方法：融合预训练DINO特征与轻量级卷积网络，通过卷积网络预测特征残差以优化DINO 基础特征，构建高精度点跟踪模型，可用于动态医学图像序列中解剖结构的运动轨迹追踪。

- 通用匹配模型：诸如MatchAnything通过大规模跨模态数据预训练，具备强大的跨模态图像匹配能力，能够有效处理模态差异大、结构特征不重叠的图像配准任务，在组织学切片、视网膜OCT 等细分场景中验证了其优异性能。

- 数据生成与增广技术：结合扩散模型，例如StyleBooth，以及深度估计模型，例如DepthAnything等工具，生成红外图像、深度图像、事件流数据及艺术风格化医学图像等多模态数据，用于扩充配准模型训练集，缓解标注数据稀缺问题，进一步增强模型对不同成像条件、不同模态数据的泛化能力。

7.4 基于插值的配准：Landmark采样与基函数选择

插值是医学图像非刚性配准中“从稀疏约束推导稠密形变场”的核心技术，其通过有限个特征点的位移，结合特定基函数的插值计算，生成覆盖全图像的稠密位移场。插值效果的关键在于“Landmark 采样方式”与“基函数类型”的匹配，二者需根据解剖结构特性与配准精度需求选择。

7.4.1 Landmark的采样方式

Landmark是插值计算的“约束锚点”，其采样方式分为规则采样与不规则采样，对应不同的配准场景与图像特性：

规则采样: Landmark按固定网格间距分布，如50×50×50体素网格，形成规则控制格点。其优势是“结构简单、参数化统一”，无需依赖图像内容即可生成，适用于“无明显局部结构差异”的全局形变建模。自由形变模型（FFD）通过规则控制格点的位移，结合三线性插值生成全图像形变场，常用于三维图像的整体配准。样条曲线/曲面（如B样条）以规则采样的控制点定义曲线形态，保证形变的平滑性与连续性，适用于骨骼、器官轮廓等相对规则结构的配准。

不规则采样: Landmark的位置由图像自身解剖特征决定，仅在“关键结构区域”，如病变边界、器官轮廓转折点采样，非关键区域无需设置锚点，其聚焦重要结构、降低冗余计算，适用于“局部形变显著、结构异质性强”的场景。

胸CT肺配准：在肺叶边界、支气管分叉等关键结构处设置Landmark，仅通过这些点的位移插值生成肺内稠密形变场，避免对无意义背景区域的多余计算；

脑图像配准：在脑沟回转折点、灰质核团中心等解剖标志点处采样，精准刻画脑内局部复杂形变，提升关键功能区的配准精度。

无论哪种采样方式，均是“以最少的Landmark数量实现最高的形变场精度”，规则采样适合全局平滑形变，不规则采样适合局部精细形变，实际应用中可结合两种方式，如全局规则格点+ 局部不规则补点，平衡效率与精度。

7.4.2 插值基函数的扩展：从空域到频域与谱域

传统插值多基于空域基函数，如样条、径向基函数，但随着配准场景的复杂化，基函数已扩展至频域与谱域，以应对“噪声干扰”“高维数据”等挑战。

表 7.2: 插值基函数的类型、特性与应用场景

基函数类型	形式	特性	适用场景
空域基函数	样条（TPS、B样条）、径向基函数（RBF）	直接在像素/体素空间计算，直观易解释，形变平滑性可控	低噪声、小形变的精细配准
频域基函数	傅里叶基（三角多项式）	通过离散傅里叶变换（DFT）滤除高频噪声，降低噪声对插值的干扰	高噪声图像的配准，需抑制噪声扰动
谱域基函数	图拉普拉斯特征向量	将图像转化为图模型，通过图矩阵特征分解生成正交基，实现低维映射	高维数据的配准，降低计算复杂度

频域基函数（傅里叶基）：先对图像进行离散傅里叶变换，将空域信号转换为频域信号。在频域中滤除高频噪声分量，仅保留低频有用信号，再通过逆傅里叶变换将信号映射回空域。最后基于频域处理后的信号进行插值，生成抗噪声的形变场。该方法尤其适用于低剂量CT、磁共振成像MRI等易受噪声干扰的图像配准。

谱域基函数（图拉普拉斯特征向量）：将图像的像素/体素视为图的节点，像素间相似度视为边的权重，构建图模型。对图的拉普拉斯矩阵进行特征分解，得到一组正交的特征向量；利用前k 个小特征值对应的特征向量刻画图的全局结构，滤除高频局部噪声进行插值，实现高维数据的低维形变场建模。典型算法包括“流形对齐（Manifold Alignment）”“函数映射（Functional Map）”，常用于三维表面网格的配准。

弹性体样条配准

弹性体样条（Elastic Body Splines, EBS）是一种基于物理力学模型的非刚性图像配准方法，它结合了弹性力学和样条插值的思想。将图像变形建模为弹性体在外力作用下的形变，并通过控制点驱动全局变形。

EBS 将图像视为弹性连续介质，在控制点施加外力时，整个图像会发生平滑的弹性变形。其数学模型基于弹性力学平衡方程，并利用径向基函数（Radial Basis

Functions, RBF）进行插值，确保变形场的光滑性和物理合理性。该方法**基于物理模型**，遵循线性弹性或St. Venant-Kirchhoff材料模型。通过**控制点驱动**，匹配的特征点计算变形场。具有**全局平滑性**，变形场满足弹性平衡方程，避免局部畸变。**解析解存在**，在某些简化条件下，可以直接计算变形场，无需迭代优化。EBS 的变形场$\mathbf{u}(\mathbf{x})$ 满足弹性平衡方程：

$$\mu\nabla^2\mathbf{u} + (\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) + \mathbf{f}(\mathbf{x}) = 0$$

其中$\mathbf{u}(\mathbf{x})$ 是位移场。μ 和λ 是Lamé弹性参数，控制材料的刚度。$\mathbf{f}(\mathbf{x})$ 是外力场，由控制点匹配误差驱动。假设有N 个控制点$\{\mathbf{p}_i\}$ 及其目标位置$\{\mathbf{q}_i\}$，外力可以表示为：

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{F}_i \delta(\mathbf{x}-\mathbf{p}_i)$$

其中$\mathbf{F}_i$ 是作用在第i 个控制点上的力。$\delta(\cdot)$ 是Dirac delta函数表示的点力模型。

在无限大弹性介质中，EBS的位移场基于Green函数方法的解析解可以表示为：

$$\mathbf{u}(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{G}(\mathbf{x}-\mathbf{p}_i)\mathbf{F}_i$$

其中$\mathbf{G}(\mathbf{r})$ 是弹性体的Green函数。对于三维情况，Green函数为：

$$\mathbf{G}(\mathbf{r}) = \frac{1}{8\pi\mu}\left[\frac{1}{|\mathbf{r}|}\mathbf{I} + \frac{\lambda+\mu}{\lambda+2\mu}\frac{\mathbf{r}\mathbf{r}^T}{|\mathbf{r}|^3}\right]$$

其中$\mathbf{I}$ 是单位矩阵。配准流程如下：

- **输入控制点**：给定源图像的控制点$\{\mathbf{p}_i\}$ 和目标图像的控制点$\{\mathbf{q}_i\}$。
- **计算外力**：通过最小化匹配误差$\mathbf{u}(\mathbf{p}_i) = \mathbf{q}_i - \mathbf{p}_i$，求解$\mathbf{F}_i$。
- **计算全局位移场**：利用Green函数插值所有点的位移。
- **应用变形**：将$\mathbf{u}(\mathbf{x})$ 作用于源图像，完成配准。

弹性体样条是一种基于弹性力学原理的非刚性配准方法，适用于需要物理合理变形的医学图像分析。它通过控制点驱动全局变形，并利用弹性Green函数计算平滑位移场。相比薄板样条，EBS 更适用于生物组织变形建模，但计算复杂度较高。该方法具有**物理合理性**，变形符合弹性力学规律，适用于生物组织。具有**全局平滑性**：避免局部扭曲，保持拓扑结构。**解析解存在**：无需迭代优化，计算高效。但是其**计算复杂度高**，当控制点较多时，矩阵求逆计算量大；**依赖控制点**，如果特征点不准确，配准效果下降；**参数选择敏感**，其中μ 和λ 需要调参。

表 7.3: EBS与TPS方法比较

特性	Elastic Body Splines (EBS)	Thin-Plate Splines (TPS)
物理基础	基于弹性力学（弹性体变形）	基于弯曲能量最小化（薄金属板）
控制参数	Lamé系数μ, λ	弯曲刚度参数
适用场景	更适合生物组织（如器官变形）	适用于一般非刚性配准
计算复杂度	较高（涉及弹性力学方程）	较低（解析解易计算）
拓扑保持性	较好（避免自交）	可能发生折叠

FFD非刚性配准方法

FFD非刚性配准的目标是估计从源图像到目标图像的最优变形变换，可以考虑**公共域映射**。定义两个变换函数T_1 和T_2，分别将源图像和目标图像映射至同一公共域，该公共域需满足“结构兼容性”，例如医学图像中器官轮廓的拓扑一致性。构建包含“相似性度量”与“正则化项”的目标函数，通过梯度下降等优化算法求解T_1和T_2 的最优参数。源图像到目标图像的最终映射通过“变换组合”实现，即先对源图像应用T_1 映射至公共域，再对公共域结果应用T_2 的逆变换T_2^{-1}，最终得到源图像到目标图像的直接变形场$T = T_2^{-1} \circ T_1$。

自由形变

自由形变（Free Form Deformations，FFD）基于控制网格的非刚性图像配准，通过调整网格控制点的位置来驱动图像变形，利用B样条（B-spline）基函数插值变形场。自由形变模型能够通过参数化控制网格实现图像的平滑、连续变形，适用于处理生物医学图像、遥感图像等复杂非刚性场景。而Jacobian行列式作为分析变形场物理合理性的关键工具，可定量评估变形过程中“体积拉伸/压缩”“折叠”等异常情况，如Jacobian 行列式小于0表示变形场存在折叠，不符合物理规律。将FFD与Jacobian 行列式结合，可在优化配准精度的同时，通过约束Jacobian 行列式范围，如强制其大于0剔除不合理变形，显著提升配准结果的鲁棒性与物理一致性。

控制网格定义 在图像上定义一个均匀的3D控制网格，网格点数为$(n_x + 3) \times (n_y + 3) \times (n_z + 3)$。每个控制点的位移$\mathbf{u}_{i,j,k}$ 影响其局部邻域的变形。

B样条基函数插值 变形场$\mathbf{T}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$ 由三次B样条基函数计算：

$$\mathbf{u}(\mathbf{x}) = \sum_{i=0}^{3} \sum_{j=0}^{3} \sum_{k=0}^{3} B_i(u) B_j(v) B_k(w) \mathbf{u}_{i',j',k'}$$

其中$\mathbf{x} = (x, y, z)$ 是图像中的点坐标。(i', j', k') 是控制点索引，由$\mathbf{x}$ 所在网格单元决

定。(u, v, w) 是归一化局部坐标，$u = \frac{x - x_{\min}}{\Delta x}$。$B_i(u)$ 是三次B样条基函数：

$$\begin{cases} B_0(u) = (1-u)^3/6 \\ B_1(u) = (3u^3 - 6u^2 + 4)/6 \\ B_2(u) = (-3u^3 + 3u^2 + 3u + 1)/6 \\ B_3(u) = u^3/6 \end{cases}$$

FFD 通常结合相似性度量和正则化项进行优化：

$$E = \text{Similarity}(I_{\text{ref}}, I_{\text{def}}) + \lambda \text{Reg}(\mathbf{u})$$

形变场的Jacobian行列式$J(\mathbf{x})$ 衡量变形场的局部体积变化率，用于分析体积膨胀（$J > 1$）、体积收缩（$0 < J < 1$）和折叠/自交（$J \leq 0$）。

变形场$\mathbf{T}(\mathbf{x}) = (T_x, T_y, T_z)$，其Jacobian矩阵为：

$$\mathbf{J} = \nabla \mathbf{T} = \begin{bmatrix} \frac{\partial T_x}{\partial x} & \frac{\partial T_x}{\partial y} & \frac{\partial T_x}{\partial z} \\ \frac{\partial T_y}{\partial x} & \frac{\partial T_y}{\partial y} & \frac{\partial T_y}{\partial z} \\ \frac{\partial T_z}{\partial x} & \frac{\partial T_z}{\partial y} & \frac{\partial T_z}{\partial z} \end{bmatrix}$$

$$J(\mathbf{x}) = \det(\mathbf{J}) = \begin{vmatrix} \frac{\partial T_x}{\partial x} & \frac{\partial T_x}{\partial y} & \frac{\partial T_x}{\partial z} \\ \frac{\partial T_y}{\partial x} & \frac{\partial T_y}{\partial y} & \frac{\partial T_y}{\partial z} \\ \frac{\partial T_z}{\partial x} & \frac{\partial T_z}{\partial y} & \frac{\partial T_z}{\partial z} \end{vmatrix}$$

由于FFD的变形场由B样条插值得到，其偏导数可解析计算：

$$\frac{\partial T_x}{\partial x} = 1 + \sum_{i=0}^{3} \sum_{j=0}^{3} \sum_{k=0}^{3} \frac{\partial B_i(u)}{\partial u} \cdot \frac{1}{\Delta x} \cdot B_j(v) B_k(w) \cdot u^x_{i',j',k'}$$

类似地计算其他偏导数，最终得到$J(\mathbf{x})$。

形变场的Jacobian行列式可用于:

- **检测折叠**：若$J(\mathbf{x}) \leq 0$，则变形场在该点发生自交。
- **量化体积变化**：$J > 1$：局部膨胀，如肿瘤生长；$0 < J < 1$：局部收缩。
- **正则化约束**：在优化中添加$\int (J(\mathbf{x}) - 1)^2 d\mathbf{x}$ 惩罚项，防止过度变形。

FFD是一种灵活的非刚性配准方法，而Jacobian行列式是分析变形场合理性的重要工具，两者结合可提高配准的鲁棒性和物理合理性。

7.5 基于统计约束的形变：从降维到物理合理性保障

医学图像配准的形变场通常具有高维特性，如128^3体数据的形变场维度达$128^3 \times 3$，直接优化易导致“过拟合”“计算爆炸”。形变需符合解剖生理规律，如不可逆、拓扑保持。基于统计约束的形变模型通过“维度压缩”与“物理规则约束”。配准需要考虑包括“子空间约束”与“任务特定约束”两类。

表 7.4: FFD与Jacobian行列式的特性比较

特性	FFD	Jacobian行列式
变形模型	B样条插值控制网格	衡量局部体积变化率
计算方式	基函数加权求和	变形场梯度矩阵的行列式
应用场景	非刚性配准、动画形变	检测折叠、量化体积变化
优化约束	控制点位移正则化	$J > 0$ 保证拓扑保持

7.5.1 子空间约束：高维形变场的维度消减

子空间约束的核心思想是通过统计分析将高维形变场投影到低维子空间，仅在子空间内优化形变参数，大幅降低计算复杂度并保证形变的合理性。

数据采集与形变场构建：收集N例同一解剖结构的医学图像，对每例图像与标准模板图像进行配准，得到N个稠密形变场$\{T_1, T_2, ..., T_N\}$，每个T_i 为高维向量，维度为D，如$D = 128^3 \times 3$。

子空间学习：对N个形变场进行主成分分析，计算协方差矩阵的特征值与特征向量，取前k 个最大特征值对应的特征向量$\{\phi_1, \phi_2, ..., \phi_k\}$，构成形变子空间$S = \text{span}\{\phi_1, ..., \phi_k\}$。

形变场表示与优化：任意新的形变场T可表示为子空间基的线性组合：

$$T = \sum_{i=1}^{k} \alpha_i \phi_i$$

其中α_i为系数，维度仅为k，远小于D；配准时仅需优化系数α_i，而非高维形变场T，实现维度消减。

子空间约束实现**计算效率提升**，形变参数维度从D降至k，大幅降低深度学习模型的训练难度与推理时间。

保障形变合理性，子空间由大量真实医学图像的形变场统计生成，仅在子空间内采样的形变场天然符合解剖规律，如脑形变的整体趋势，避免出现不符合生理的异常形变。

7.5.2 任务特定约束：可逆性与拓扑保持

除维度约束外，医学图像配准还需满足“任务特定的物理合理性约束”，包括可逆性约束与拓扑保持约束，二者均为保证解剖结构对应一致性的关键：

可逆性约束 要求形变场满足“双向一致性”，若存在从移动图像M到固定图像F的形变场T，即$F = T(M)$，则需存在从F到M的逆形变场T^{-1}，即$M = T^{-1}(F)$，且满足

$$T^{-1}(T(M)) = M, T(T^{-1}(F)) = F$$

可逆性的核心作用是避免“结构映射歧义”，实现方式包括：

对称形变建模：同时估计正向形变场$T_{M\to F}$与反向形变场$T_{F\to M}$，在损失函数中加入“对称约束项”

$$\|T_{F\to M}(T_{M\to F}(M)) - M\|_2^2$$

强制双向形变的一致性；

中间域映射：将M与F均映射到一个公共中间域C，即$C = T_{M\to C}(M) = T_{F\to C}(F)$，则反向形变可通过

$$T_{C\to M} = T_{M\to C}^{-1}, T_{C\text{ß}F} = T_{F\text{ß}C}^{-1}$$

间接保证可逆性。

拓扑保持约束 要求形变过程中解剖结构的拓扑关系不变，保证形变场为“双射函数”，即任意两个不同像素/体素经形变后仍为不同点，无多对一映射。医学图像中常用“雅可比行列式”实现拓扑约束：对形变场$T = (u, v, w)$计算其雅可比矩阵

$$J(T) = \begin{bmatrix} \partial u/\partial x & \partial u/\partial y & \partial u/\partial z \\ \partial v/\partial x & \partial v/\partial y & \partial v/\partial z \\ \partial w/\partial x & \partial w/\partial y & \partial w/\partial z \end{bmatrix}$$

计算雅可比行列式$\det(J(T))$，要求其值大于0（通常设置阈值$\epsilon > 0$，如$\det(J(T)) > 10^{-6}$），若$\det(J(T)) > 0$，则形变场为局部双射，保证拓扑不发生折叠。若$\det(J(T)) \leq 0$，则存在结构折叠，需通过正则项惩罚此类形变。

7.6 基于物理的形变模型

弹性体模型

弹性体模型是一种基于连续介质力学中弹性理论的图像配准方法，将图像视为可发生弹性形变的连续体。**弹性体假设**图像被模拟为弹性材料，在外力作用下可发生形变但仍保持结构连续。配准过程转化为最小化系统总能量的优化问题，通过求解位移场实现图像对齐。变形过程通过以下函数描述：

$$E(\mathbf{u}) = E_{\text{elastic}}(\mathbf{u}) + E_{\text{image}}(\mathbf{u})$$

其中$\mathbf{u}$ 为位移场，E_{elastic} 表示弹性变形能，E_{image} 为图像相似性度量项。

弹性变形能 通常采用线性弹性模型：

$$E_{\text{elastic}}(\mathbf{u}) = \int \left[\mu(\nabla\mathbf{u} + \nabla\mathbf{u}^T)^2 + \lambda(\nabla\cdot\mathbf{u})^2\right] d\mathbf{x}$$

其中μ 和λ 为Lamé系数，控制材料的弹性行为。

图像相似性能量 常用平方差和（SSD）、互相关（CC）或互信息（MI）等形式。

$$E_{\text{image}}(\mathbf{u}) = \int \left[I_1(\mathbf{x}+\mathbf{u}(\mathbf{x})) - I_2(\mathbf{x})\right]^2 d\mathbf{x}$$

弹性体模型采用**有限元离散**，将连续体划分为有限元网格。利用**欧拉-拉格朗日方程**，推导控制变形的偏微分方程。基于**迭代优化**求解，常用梯度下降或共轭梯度法进行数值求解。该方法可保持拓扑结构，防止折叠或穿透，产生平滑且物理合理的变形场，通过弹性参数灵活调节形变刚度。该模型适用于医学图像配准，能够保持解剖结构的合理性。

粘性流体流动模型

粘性流体流动模型将图像变形过程类比为粘性流体的流动，适用于处理大尺度形变。其类比**流体力学**，图像形变类似于粘性流体的流动行为；具有**时间连续性**，通过引入时间变量描述连续变形过程；具有**大变形适应**，能够有效处理高度非线性形变。基于Navier-Stokes方程：

$$\mu\nabla^2\mathbf{v} + (\mu+\lambda)\nabla(\nabla\cdot\mathbf{v}) + \mathbf{F} = 0$$

其中$\mathbf{v} = \dfrac{d\mathbf{u}}{dt}$ 是速度场，$\mathbf{F}$ 为图像衍生的驱动力：

$$\mathbf{F}(\mathbf{x}) = -\left[I_1(\mathbf{x}+\mathbf{u}(\mathbf{x})) - I_2(\mathbf{x})\right]\nabla I_1(\mathbf{x}+\mathbf{u}(\mathbf{x}))$$

时间离散采用欧拉方法：

$$\mathbf{u}_{t+\Delta t} = \mathbf{u}_t + \mathbf{v}_t\Delta t$$

粘性流体流动模型采用线性化近似，使用多重网格法加速大规模方程求解，迭代推进变形场。该方法支持大形变配准，允许拓扑结构变化，计算复杂度较高，适用于生长建模、运动分析等场景。值得注意流体模型更适合大形变，弹性模型更易于解释和参数调节，流体模型计算开销更大。

借鉴扩散过程中的“Demons”概念，将图像配准视为由图像梯度引导的双向推拉过程。基于光流法，假设图像强度守恒；同时施加“主动力”和“被动力”，推动图像双向对齐，每步迭代后对位移场进行高斯平滑以保持正则性。位移场更新公式为：

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \frac{(I_F - I_R\circ\mathbf{u}_n)\nabla I_R}{|\nabla I_R|^2 + (I_F - I_R\circ\mathbf{u}_n)^2}$$

该方法计算效率高，适用于中小程度形变，已有多种改进版本，例如对称Demons。

曲率配准

曲率配准（Curvature Registration）以曲率为正则项，确保变形场的光滑性和合理性。使用曲率算子（拉普拉斯算子）约束变形。最小化能量函数：

$$E(\mathbf{u}) = \text{相似性项} + \alpha \int (\Delta \mathbf{u})^2 d\mathbf{x}$$

对应的欧拉-拉格朗日方程为：

$$\Delta^2 \mathbf{u} + \mathbf{F}(\mathbf{u}) = 0$$

该方法变形场高度平滑；对噪声鲁棒；适用于需保持解剖连续性的场景。

微分同胚配准（Diffeomorphic Registration）原理

微分同胚配准（Diffeomorphic Registration）保证变换为光滑可逆的微分同胚映射，特别适用于大变形配准。通过速度场$\mathbf{v}$ 的时间积分构造变形场ϕ：

$$\phi = \exp(\mathbf{v})$$

在大变形微分同胚度量映射（LDDMM）框架中，定义能量泛函：

$$E(\mathbf{v}) = \int_0^1 \|\mathbf{v}_t\|_V^2 dt + \text{相似性项}$$

该方法严格保持拓扑结构，可处理大形变，计算复杂度高，广泛应用于脑图像分析。

表 7.5: 不同配准方法的特性比较

特性	Demons	Curvature	Diffeomorphic
变形程度支持	中/小	中/小	大变形
拓扑保持	不保证	部分保证	严格保证
计算效率	高	中等	低
参数数量	少	中等	多

这些方法在现代医学图像分析中都有广泛应用，选择取决于具体的应用需求和计算资源限制。

基于知识的模型

基于知识的模型利用**统计约束的几何变换**，通过统计先验减少变形场的自由度。使用**任务特异性约束**，包括：**逆一致性**，同时估计正反向变换并惩罚其差异；**对称性**，通过对称目标函数将图像映射到公共域；**拓扑保持**，确保变形场雅可比行列式处处为正（$\det(J(\mathbf{x})) > 0$），维持局部微分同胚特性。

7.7 跨模态匹配方法

跨模态匹配如RGB与红外、SAR与光学、医学CT与MRI的核心挑战在于“模态差异导致的特征分布偏移”与“带标注训练数据稀缺”，MatchAnything方法通过“数据策略创新”“模型架构优化”与“跨模态训练设计”，实现了多模态场景下的鲁棒匹配。跨模态匹配任务的根本瓶颈是**带标注训练数据的稀缺性**。一方面，跨模态数据，如同步采集的RGB- 红外图像对获取成本高，需专用传感器；另一方面，人工标注跨模态对应点，如医学图像中CT 与MRI 的解剖结构对应耗时且主观误差大，导致大规模高质量数据集难以构建。“数据合成”被用于为突破数据瓶颈 [16]。

跨模态刺激信号生成 通过像素对齐的图像翻译网络，如CycleGAN [59]、Pix2Pix [21]合成多模态图像对：以单模态数据为输入，生成外观与源模态差异显著但几何结构完全一致的目标模态图像。该策略迫使模型脱离“模态特异性特征”，转而学习“模态无关的结构特征”，提升跨模态泛化能力。

训练数据多样性增强 混合多源无标注/弱标注数据，覆盖广泛场景与模态类型，其中多视图重建数据，提供精确几何约束，辅助学习三维结构匹配。未标记视频序列，利用帧间时序连续性生成伪标签，解决无标注问题。使用单图像数据集，通过图像变换扩充数据量，增强模型对场景变化的适应性。该策略避免模型过拟合于单一数据源/场景，如仅训练医学图像导致无法适配遥感场景，提升对未知跨模态任务的兼容性。

对于**多视图几何数据**，基于深度图，例如MegaDepth的稀疏深度、ScanNet的稠密深度实现跨视图像素投影，生成亚像素级真实匹配点对。利用**深度投影**，将源视图像素通过深度信息投影至相邻目标视图，计算理论对应点坐标。**噪声过滤**通过“置信度掩码”，如深度误差小于阈值、投影点在目标视图有效区域内，剔除遮挡、深度噪声导致的错误对应点。**亚像素优化**对理论对应点进行局部像素搜索，如归一化互相关窗口匹配，提升对应点精度至亚像素级。

数据利用策略：从粗到细与多源联合

针对视频数据无标注的问题，采用从粗到细与多源联合：

- **粗匹配**：基于光流估计（如RAFT）或稀疏特征跟踪（如SuperGlue）建立相邻帧间的初始对应关系，生成短时序轨迹。
- **细优化**：通过长期轨迹关联（如Bundle Adjustment）与一致性约束（如重投影误差过滤），剔除噪声对应点，生成跨远帧的高精度伪真值，解决“长时序漂移”问题。

多源数据联合训练 融合不同数据源的互补优势，构建多任务联合优化目标：其中场景重建数据提供几何监督，提升匹配精度。视频数据提供时序监督，增强动态场景适应性。单图像数据提供多样性监督，提升泛化能力。通过多任务损失函数，如几

何损失+时序损失+相似性损失的联合优化，使模型同时学习不同数据的特性，避免单一数据训练的局限性。

表 7.6: MatchAnything方法

数据	跨模态图像合成（GAN/扩散模型）→解决标注稀缺 多源数据混合（重建+视频+单图像）→解决多样性不足
训练	从粗到细伪标签生成→释放无监督视频数据价值 多任务联合优化→融合不同数据源互补特性
应用	零修改适配现有无检测器模型→降低使用门槛 单一权重跨多模态任务零样本泛化→减少微调需求

基础模型架构选择

MatchAnything采用**Transformer的密集/半密集匹配架构**，类似LoFTR、MatchFormer放弃传统“特征检测-描述-匹配”的两阶段流程，直接处理像素级或patch级特征匹配。该方法对低纹理、重复模式场景更鲁棒，传统检测器在低纹理区域易失效，而Transformer通过全局注意力捕捉长程依赖，可学习全局结构匹配。适合端到端预训练，密集匹配架构支持从像素级监督，如伪标签到任务级监督的端到端优化，无需人工设计特征描述子。具有多模态兼容性，通过共享Transformer编码器将不同模态图像映射至同一特征空间。

模态对齐生成 采用扩散模型（如Stable Diffusion）或GAN（如CycleGAN）生成高质量跨模态图像对。像素级几何对齐通过条件生成模型的“空间注意力机制”，强制生成图像与源图像的几何结构完全一致，避免合成过程中的结构偏移；生成的目标模态需涵盖不同传感器类型、不同环境条件，确保模型适应各类模态差异。此外，逐步提升模型对跨模态差异的适应能力。先在大规模单模态数据上训练模型的基础匹配能力，学习通用结构特征，随后引入模态差异较小的跨模态数据，通过“模态对齐损失”微调模型；最后引入模态差异大的跨模态数据，通过“跨模态匹配损失”优化模型，完成跨模态能力的迁移。

7.8 实例：可变形2D-3D配准

可变形2D-3D配准技术通过建立三维容积图像与二维X光片之间的密集对应关系，可用于基于二维影像的患者特异性三维结构重建与形态变化追踪，在纵向治疗评估和回顾性研究中具有重要意义。与锥形束CT （CBCT）相比，侧位头颅X光片（LC）因辐射低、成本低，成为正畸临床中常用的影像工具。基于模板的可变形2D-3D配准有望进一步减少辐射暴露，但LC作为二维投影，在恢复三维结构方面存在固有歧义，重建问题高度不适定。

传统2D-3D配准通过优化变形模板生成的DRR（数字重建放射影像）与目标X光片之间的图像相似度实现，通常依赖耗时的非线性优化和渲染过程。为提升效率，统计模型被用于约束配准参数空间，深度学习方法则借助神经网络直接预测配准变换，降低了计算成本。然而，基于单张LC的配准与重建仍面临三重挑战：一是深度学习严重依赖大量成对三维-二维数据，采集困难且增加辐射风险；二是统计形状模型提供有用先验但结果过于平滑，缺乏细节；三是常用的强度相似度度量缺乏对解剖结构语义一致性的显式约束。

针对以上问题，Jiang 等人 [22]提出一种弱监督学习框架，能够从单一LC图像生成体素级配准场，实现高质量的三维颅面结构重建。该方法基于统计配准场模型（SRM）产生与LC一致的基础变形，并利用配准场优化器进一步增强结构细节，克服传统统计方法的过度平滑问题。框架整合了基于SRM的容积先验与可微分渲染器，通过自监督进行配准场预测，并引入局部结构的弱监督标注以增强语义一致性，从而提升重建的结构保真度和稳健性。在合成及临床LC图像上的实验表明，本方法在颅面CBCT重建任务中优于现有主流技术。消融实验验证了配准场优化器和语义约束机制的有效性。

基于模板的2D-3D配准旨在最大化变形模板体积的数字重建放射影像（DRR）与目标二维X光片之间的图像相似度。传统方法涉及大规模变换参数的非线性优化和容积渲染过程，具有耗时特性 [9, 28, 35, 55, 60]。为减轻在线计算负担，统计学模型被引入2D-3D配准领域，通过在低维参数空间中求解容积变换 [3, 52, 53, 56]。近期基于深度学习的配准方法通过建立非线性映射至配准场，有效降低了在线计算成本 [17, 25, 31]。然而，基于单张LC 的深度可变形2D-3D配准与结构重建仍面临三重挑战：

首先，现有深度学习配准方法依赖大量配对的三维容积图像与二维X光片进行学习，这加剧了数据收集负担与辐射暴露风险 [31, 42]。虽可采用合成数据缓解数据需求，但仍需额外努力处理合成数据与临床数据之间的域间差异。其次，统计形状模型通过形状先验在低维参数空间促进2D-3D配准 [3, 53]，但基于统计模型的配准往往过于平滑，难以精细建模颅面结构。第三，虽然常用变形体积DRR与目标二维X光片之间的强度相似度评估跨维度配准and，但该方法缺乏对解剖结构跨维度语义对应的显式考量。

本工作提出了一种弱监督学习框架，用于生成目标二维LC与CBCT之间的体素级配准场，从而实现患者特异性三维颅面结构重建，并显著降低容积图像采集中的辐射风险。统计配准场模型（SRM）能产生与LC一致的容积变形，尽管结果平滑且缺乏结构细节。引入配准场优化器来处理精细颅面结构的体素级变形，通过粗细化配准场克服统计模型方法的平滑变形问题。该可变形2D-3D 配准框架整合了基于SRM 的容积先验与体素级优化器，利用可微分体积渲染器和自监督学习预测配准场。进一步通过局部结构的弱监督实现语义一致性约束，从而促进高保真度2D-3D配准和容积重建中的稳健结构解析。在临床LC与CBCT图像上的定性与定量实验表明，

该方法在患者特异性颅面结构重建方面优于最先进的基于单张X 光片的重建方法。通过系统的消融研究验证了配准场优化器与语义一致性约束的有效性。

2D-3D配准框架

精化2D-3D配准框架用于估计患者特异性颅面结构。给定包含m个像素的LC图像$I \in \mathbb{R}^m$ 与包含n 个体素的模板体积$T \in \mathbb{R}^n$，目标是推断最优可变形配准场$\phi \in \mathbb{R}^{3n}$。映射函数$f_\Theta : (I, T) \to \phi$ 由参数为Θ 的深度神经网络参数化实现。本框架由基于SRM的粗配准和体素级配准场优化器构成：首先，SRM回归器预测配准场子空间中的低维坐标α；基于SRM的解码器重建粗配准场$\phi_c \in \mathbb{R}^{3n}$；配准场优化器将经目标LC 特征嵌入增强的$\phi_c$作为输入，预测体素级细化配准场$\phi_f \in \mathbb{R}^{3n}$。空间变换器与可微分体积渲染器实现自监督学习，减轻了容积图像采集中的辐射风险。通过下颌骨图谱实现弱监督，以增强目标LC 与模板体积间的语义对应关系。

优化器网络的输入$\hat{X}$包含粗位移场ϕ_c和多尺度LC特征X两部分。在体素v 处的优化器输入为$\hat{X}(v) = (\phi_c(v), X(p))$，其中关联投影像素$p = \pi(v + \phi_c(v))$。输入的第一分量是位移向量$\phi_c(v) \in \mathbb{R}^3$，第二分量是关联像素$p$ 处的多尺度LC特征嵌入$X(p)$。

7.9 小结

医学图像配准技术经历了从传统数值优化、到传统机器学习、再到深度学习与基础模型的发展历程。当前研究呈现出以下趋势：

- **表征学习**：从手工特征转向数据驱动的深度特征学习。
- **模型架构**：从迭代优化转向端到端的深度学习模型，并进一步向利用大规模预训练基础模型的方向演进。
- **约束利用**：更加注重融入物理规律、解剖学知识等约束，以保证配准结果的合理性与可解释性。
- **跨模态能力**：通用基础模型正在突破模态限制，朝向能够处理多种模态图像的通用匹配算法发展。

图像配准经历了从“数值迭代”到“特征驱动的无训练优化”。图像配准（如术前术后影像对齐、多模态影像融合）的技术逻辑，也因大模型的介入实现了“降维打击”：

传统阶段：通过数值迭代优化求解形变参数，形变规模可大可小，既可以是刚性变换、全局affine 变换，也可以是像素级独立位移，参数规模与图像像素数一致，计算成本极高。

深度学习阶段：构建端到端模型，从一对输入图像直接预测形变场，需大量“配准对”数据训练模型。

大模型阶段：直接复用预训大模型的“模态无关特征”，无需训练任何网络参数，通过特征相似性匹配，直接优化图像间的形变场，计算逻辑与传统数值优化一致，但因特征具备强语义性，配准精度与鲁棒性远超传统方法。

随着大规模多模态数据的积累以及基础模型技术的不断成熟，医学图像配准技术将继续向着更高效、更精准、更通用的方向发展。

第八章　三维重建

假设有图像f，存在一个与之对应的测量值g，该测量值呈现为模糊图像的形态，变换函数为h。图像重建的目标是求解成像系统的逆函数，即求出h的逆。一旦成功求解h的逆，便能够依据测量值g 重建原始图像f。当前测量值g 可能会受到噪声n 的污染。即便噪声n看似幅度较小，然而若直接对受噪声污染的测量值$g+n$ 应用逆向成像函数h^{-1}，却无法重现原始图像f。面临的主要问题是测量仪器通常具有有限性、离散性以及可量化的特点，而物理成像系统则恰恰相反，它具有不确定性、非线性，并且难以进行精确度量。

传统的医学图像重建方法主要包括直接求逆方法、变分方法以及Tikhonov正则化方法。在稀疏化重建领域，常见的方法包含压缩感知与稀疏字典学习。在传统重建方法中，如果仅有少量稀疏的采样点，重建工作将无法开展。但在当下，借助图像分布先验，即便使用稀疏采样点，甚至达到仅有一个采样视图的极端情况，也能够实现三维图像的重建。

将机器学习用于医学图像重建中，需要学习优化成像过程中的诸多参数，包括对正则化系数进行优化，以获取更为理想的正则化效果；学习数据的置信项，从而增强数据的可靠性；构建去噪函数，用于消除图像中的噪声干扰；以及学习前向投影模型，以精确模拟成像过程。基于机器学习的医学图像重建的最根本的任务，是学习图像到图像之间的回归函数，这与自然图像的重建任务在本质上具有一致性。

8.1　传统生物医学图像三维重建方法

8.1.1　FBP方法

经典的重建算法如滤波反投影（Filtered Back - Projection，FBP）算法，在重建过程中，通过离散化或者运用合成算子，生成线性成像矩阵h。若要直接求逆，就需要计算线性算子的逆矩阵，即求解h^{-1}。

CT在进行不同角度的X射线投射时，会生成正弦图（sinogram）。将这个正弦图进行反向投影，并对多个角度的投影结果进行累加，即可生成三维的CT 图像。若当前有一个简单的3×3二维图像，在某一角度进行投影后，其对应数值分别为3、1、2。进行反向投影时，会将这些数值填充到投影光线所经过的所有像素上。例如数值3 会

填充到对应行的所有像素位置，使该行像素值均为3。

FBP算法对图像在某一角度θ上的投影进行傅里叶变换，接着进行滤波处理，随后再进行反向傅里叶变换，最后对变换结果进行积分。当积分区间取0 到π 时，即可生成对应的二维图像。这一过程主要基于中心切片定理。该定理表明，对于一个二维图像，其X射线穿透形成的积分成像属于一维成像。对一维成像进行傅里叶变换并滤波，再进行反向傅里叶变换后求积分，能够覆盖整个二维图像的傅里叶变换，进而通过反向傅里叶变换得到对应的二维图像。这也解释了在实际应用中要求进行0到180度的积分计算。因为若仅考虑一个采样点，它仅对应二维图像傅里叶变换中经过中心的一个片段，若要覆盖整个傅里叶变换区域，就需要进行0 到180 度的反向投影计算。更为直观地说，反向投影就是将某一角度投影后的数值，沿着反向光线上的所有像素进行填充。

当投影角度在0至180度范围内旋转，利用正弦图中的记录数据进行反向投影操作时，会生成类似“柱子”的结构，该结构每条线上的数值完全一致。若仅考虑单一采样点对应的投影图像，其信息极为有限，看似并无实际意义。然而，将0至180度范围内所有角度的反向投影结果进行累加，便能够逐步构建出具有实际价值的图像信息。依据采集角度，对图像轮廓进行“涂抹”的过程中，将某一角度投影所得数值，沿该角度投影光线所经过的所有像素进行填充。但生成的图像往往存在模糊现象。这种模糊是由反向投影累积求和所导致的，该模糊效应被称为1/r模糊。可采用滤波手段来消除该模糊。通常选用斜坡滤波（Ramp Filter），其计算效率较高，在频率域中对应一个乘法函数，能够借助快速傅里叶变换高效实现。斜坡滤波能够有效校正1/r模糊效应。

直接方法在医学图像重建中存在明显的局限性。从理论层面分析，直接方法若要实现CT 图像的重建，前提是获取完整的数据，即涵盖所有视点上的投影图像。然而，当采样点数量减少时，其重建质量会急剧恶化，甚至难以完成重建任务。特别是在当前备受关注的小采样点、稀疏采样点，乃至单采样点或正交采样点的情况下，FBP算法完全无法满足三维图像重建的需求。

8.1.2 变分方法

假设待重建的图像为f，数据项表示为$g = Hf$，其中H 代表成像系统，该问题通常是一个病态问题，成像系统H不一定可逆。若从最小二乘意义考虑，可以构建法线方程来求解f。计算H的伪逆，从而得到最小二乘意义下的最优图像。

Tikhonov正则化

原始重建问题由于H不可逆而呈现病态，通过添加正则化项，可将其转化为良态问题。在这个转化过程中，存在正则化参数λ。从线性方程组的角度来看，λ值越大，系统的良态性能越好，即系数矩阵的条件数越小。但从图像重建的实际效果而

言，随着λ 增大，数据项在重建过程中所占的比重逐渐降低，这将导致重建质量显著下降。反之，若λ 值过小，趋近于0，那么问题将再次回归到病态，在数值求解过程中会面临诸多困难。因此，在实际计算中，需要谨慎选择合适的λ 值，以确保问题既非病态，又能保证图像重建具有较高的质量。当引入蒂霍诺夫正则化项后，所得到的法方程会增加额外项，即$\lambda L^T L$，其中L 为线性算子，这使得方程组的条件数得到改善，成为一个条件良好的可逆矩阵问题。

从贝叶斯理论的角度来看，成像过程中图像f和噪声均被视为随机变量。通常假定噪声服从零均值的高斯分布，其方差记为σ^2。在此情形下，求解图像f实际上是在已知测量值g 的条件下，求解图像f 出现的概率。根据贝叶斯公式，后验概率可表示为：

$$P(f|g) = \frac{P(g|f)P(f)}{P(g)}$$

该问题通过最小均方误差法和最大后验概率法求解。在特定的噪声高斯分布假设下，两种方法的求解结果一致。最小均方误差法的目标函数是使$\|\hat{f} - f\|_2^2$1最小，其中$\hat{f}$ 为估计图像。

最大后验概率法的目标函数由两项构成，第一项是数据项$\|g - Hf\|_2^2$，第二项是正则化项。假设f 服从零均值的高斯过程，其协方差矩阵为C 且可逆，那么正则化项可表示为$f^T C^{-1} f$。在理想情况下，该项应足够小。此类问题同样可通过最小二乘法求解，在最小二乘意义下的最优解需解一个线性方程组，其系数矩阵为$H^T H + \sigma^2 C^{-1}$。

8.1.3 迭代方法

在迭代方法中，目标函数G包含两项：第一项为数据项$\|g - Hf\|_2^2$，第二项为正则化项$\lambda\|Lf\|_2^2$ ，其中L 为对f 施加的线性算子，可采用梯度下降法或共轭梯度法求解。在梯度下降法中，需确定负梯度方向，并寻找最优步长进行迭代求解，也可预先设定步长进行计算。

8.1.4 基于稀疏化的图像重建方法

该方法将图像f表示为一组非零系数。对图像进行傅里叶变换、离散余弦变换或小波变换后，能够得到一组非零系数来表征图像。在稀疏化过程中，正则化范数使用零范数或一范数，以促进图像的稀疏性。若采用零范数，其代表非零元素的个数k，且k远小于图像的像素总数。

压缩感知是稀疏重建方法，其目标函数由两部分组成：第一部分是数据项，要求测量值g 与Hf 之间的差距以二范数衡量足够小；第二部分体现稀疏性，即f的零范数需满足$\|f\|_0 \leq k$，k远小于图像像素个数。通过引入拉格朗日算子，可将该问题转化为无条件优化问题，优化目标为f。此时，目标函数的第一部分为数据项，第二部分为关于f 的正则化项，f的正则化可选用零范数或一范数。

在更一般的情况下，对f的正则化约束需施加线性算子L，使得$\|Lf\|_0 \leq k$，添加拉格朗日乘子λ后，目标函数为

$$\|g - Hf\|_2^2 + \lambda\|Lf\|_1$$

图像f 可表示为$L\alpha$的线性组合，组合系数为α，正则化针对α进行，即$\lambda\|\alpha\|_1$。通过直观比较不同稀疏图像重建方法，如离散傅里叶变换、离散余弦变换以及小波变换后的结果可以发现，DB的信噪比最高。当参数数量从1300 增加到2600 时，信噪比显著提高，这与我们的直观认知相符，即使用更多的非零系数能够捕捉图像更多细节，从而提升重建图像的信噪比。以经过模糊和下采样处理的图像为例，在重建过程中对比使用L2范数和L1范数作为正则化范数的效果，发现采用L1范数，即稀疏化能够显著改善重建质量。在对经过模糊和下采样处理的图像进行重建时，采用系数化的L1 范数，其信噪比可达10，而使用L2范数时，信噪比仅为7.9。这清晰地表明，运用基于L1范数的稀疏化方法能够有效提升重建质量。

8.2 基于学习的方法

在CT重建领域，基于机器学习的方法已成为主流选择。通过学习前向模型，即物理成像系统，将图像f 转换为测量值g，记为$g = h(f)$ 。从成对的训练数据，便可进行有监督学习以构建该前向模型。针对不同模态的图像，前向模型的学习侧重点各有不同。对于显微图像，通常假设其前向模型为卷积托普利兹矩阵，即对角线元素为常数的矩阵。在正电子发射断层扫描（PET）图像重建中，精确度量原点位置至关重要，这将用于生成前向模型中的行向量。而在磁共振成像（MRI）中，线圈灵敏度是关键因素。可通过使用具有均匀灵敏度的线圈收集数据，进而估计线圈灵敏度。

在基于机器学习的重建方法中，正则化权重的确定尤为关键。通常采用交替优化策略来寻找最优的正则化权重λ 。交替优化在已知重建图像的情况下，需要确定目标函数中λ的取值，使重建图像与真实图像之间的差异最小化。在给定λ值时，优化目标函数以计算重建图像。

目标函数一般由两部分构成：第一项为重建的数据项，反映测量值与重建图像经成像投影后的一致性；第二项是以λ为系数，对f施加线性算子L后的p范数作为正则化项，通过优化目标函数可得到最优的重建图像f 及对应的最优权重λ 。

考虑学习与正则化项对应的势能函数，记为$\Phi(Lf)$ 。首先需确定最优的线性算子L，在重建的目标函数框架下，通过使$\|Lf\|_e$最小来确定最优的L。通过带有不等式约束条件的优化问题求解图像h，即要求重建图像与目标图像之间的差异小于σ 。

8.2.1 稀疏字典学习

在稀疏字典学习里，图像可表示为字典中码的线性组合。若将码记为$\mathbb{C}$，权重记

为α，则组合后可得到重建图像。在学习过程中，需要寻找码的最优组合系数α，同时字典本身也是可学习的。对于组合系数α的正则化，通常采用p范数，p一般取零范数或一范数，以此促进系数的稀疏化，使最终重建图像尽可能依赖字典中较少的码。这里的字典可以是由小图像块构成，例如15×15 的图像块。通过确定最优的α，能够将字典中的码组合成最终的目标图像。

从更广义的角度看重建问题，对于带有参数θ的函数f_θ，以及与之相关的线性算子L_θ，求解最优的正则化形式，本质上是一个交替优化问题。在给定正则化形式下求解最优的图像函数f，再根据得到的f优化参数θ ，如此反复迭代，直至得到最优的图像f以及最优的正则化形式$\Phi(L_\theta f)$ 。

8.2.2 自编码器

自编码器通过构建编码器E和解码器D，在无监督的条件下进行训练。期望估计的函数f 的分布与自编码器所学习到的图像分布一致。在训练自编码器时，要求图像编码为稀疏编码，即对f经编码器E 编码后得到的结果，其零范数需满足$\|E(f)\|_0 \leq k$ 。传统的正则化方式被生成模型所替代。此时，目标函数的第一项为数据项，体现重建图像f经成像投影后与测量值g的一致性；第二项则基于生成模型构建正则化项，要求重建图像f 经编解码后与自身一致，即符合生成模型所刻画的图像分布。

以CT重建为例，若给定一张二维X光片，期望重建对应的三维图像。在深度学习框架下，若拥有成对的三维图像及其对应的二维投影X光片图像，便可借助卷积神经网络构建图像到图像的回归模型。然而，该回归模型与常见的图像变换或处理回归模型存在差异，主要体现在输入为二维图像，输出为三维图像，这涉及跨维度的特征嵌入变换。使用卷积神经网络处理此类端到端回归模型时，面临诸多挑战。一方面，在跨维度特征嵌入变换过程中，难以保证二维图像与三维图像特征嵌入之间的准确对应；另一方面，若采用基于卷积模型的解码器重建三维图像，空间复杂度较高。在医学图像处理领域，三维图像分辨率通常并不高，常见的处理图像分辨率为128×128×128 。这并非因为医学图像本身分辨率低，实际上，常用的CT 或核磁图像在获取时，分辨率可达0.1 - 0.3毫米甚至更高。如两年前出现的光子技术CT设备，其分辨率较传统CT设备提高了9 倍。即便如此，基于卷积神经网络直接处理如此高分辨率的图像仍困难重重。

8.2.3 隐式神经表示

隐式神经表示在自然图像或自然场景的三维表示中应用广泛，在医学图像领域，也可用于CT 标识与重建。与自然图像不同，对最终图像有贡献的体素多分布于物体表面，内部体素贡献有限且分布局限，因此可利用特殊数据结构（如八叉树）充分发挥其稀疏性，加速三维图像绘制。但医学图像不同，其外层与内层体素均对最终

图像绘制有重要贡献，这也是医学图像能够呈现人体内部底层解剖结构的原因。隐式神经表示应用于医学图像时，优势显著，因其基于坐标系统，可生成任意分辨率的三维图像，有效避免了基于卷积神经网络回归模型的高时空代价问题。

若采用隐式神经表示能够规避基于卷积神经网络的重建方法所面临的高时空代价问题。隐式神经表示在场景重建方面，多聚焦于特定场景的构建。对于医学图像领域，为确保重建精度，隐式神经表示往往针对特定个体进行建模，其泛化模型的性能仍有待验证。在自然图像领域，已有研究通过动态调整策略来提升模型在时间维度上对物体形变的泛化能力。医学图像场景中，考虑个体间解剖结构存在差异，借助隐式神经表示可实现对不同个体的有效重建。

基于深度学习的医学图像重建还可通过图像配准的方式实现。从一张二维X光片出发，估计与之对应的模板三维图像的形变场，进而生成该X光片对应的三维图像。运用统计先验模型具有显著优势，例如可大幅降低参数空间维度，通过少量非零系数即可描述三维图像。若采用线性子空间进行参数化表示，这些非零系数实际上就是子空间的坐标，其维度可被压缩至较低水平。

基于隐式神经表示的模型训练方式主要有两种：若有成对的CT 图像及对应的X光片数据，可对MLP模型进行预训练。若缺乏此类成对数据，多数情况下，临床采集的成对数据稀缺，现有的成对数据多为合成数据，即通过对三维图像进行体绘制生成指定采样点的X 光片，可通过体绘制在二维平面上比对生成的三维图像与输入的二维X光片的一致性。将三维坐标输入MLP 估计对应位置的灰度值以生成三维图像，再通过标准的投影变换，如Radon变换生成投影X 光片，在训练时仅需将投影图像与输入X 光片进行比对并最小化差异，即可优化坐标系统。

8.2.4 生成模型

扩散模型能够对三维图像的先验分布进行建模，可从随机噪声生成三维图像。对于CT重建而言，直接基于噪声进行重建并不可行。因为针对特定患者的CT重建，必须从该患者对应的X 光片。基于现有生成模型的CT 重建，通常需要进行有条件的在线图像推理。

在深度神经网络用于CT重建的实践中，从二维X光片到三维图像的端到端回归函数，深度神经网络在此过程中犹如一个黑盒。给定成对的训练图像后，即可对其参数进行优化。目标函数同时融合了传统FBP算法的目标函数以及深度神经网络自身的目标函数。

生成对抗网络（GAN）

与常规GAN有所不同，其生成器需实现从二维X光片到三维图像的跨维度图像生成。其中，关键问题在于如何将二维图像特征转换为三维图像特征，即特征变换问题。现有方法多采用线性变换，如通道特征复制、特征变形（reshape）或运用特

定变换算子，以实现二维图像特征向三维图像特征的嵌入。然而，这些方法存在明显缺陷。二维图像特征嵌入中的每个像素，蕴含着其在特定感受野内的上下文信息。但通过通道复制或特征嵌入变形变换生成的三维图像特征，与二维图像特征之间缺乏有效的语义对应关系。

卷积神经网络

基于CNN的重建网络，还可用于从哺乳动物的X光片重现其CT图像，在此过程中常利用特征通道复制将二维图像特征转化为三维图像特征。但此类重建方法存在两个主要问题：其一，二维与三维特征嵌入之间的语义含义不明确；其二，受CNN自身特性影响，重建图像存在平滑效应。为克服平滑效应，可引入对抗学习，在重建过程中增加特定患者的噪声信息，从而生成更多细节。但这也可能导致重建图像信噪比下降，因为生成的细节未必与特定患者的真实情况相符。为提升从X光片重建CT图像的精度，可构建特定个体的端到端重建模型。针对某一个体，基于少量采样视图，极端情况下为单采样视图及成对数据，训练回归模型进行CT 重建。然而，该模型存在明显局限性，虽能从少量采样点重建CT图像，但仅适用于特定个体，对新个体的X光片无法实现有效重建。

基于CNN的端到端模型中，存在语义对应问题。在从二维X光片经特征嵌入生成二维图像特征后，在通过解码器转换为三维图像特征并解码的过程中，现有方法多依赖特征通道复制、特征重塑或变换操作，缺乏从二维图像像素到三维图像体素的明确对应关系。针对这一问题，可通过构建语义对应映射函数加以解决。为简化该映射函数，可对特征嵌入进行离散化及量化处理，使二维图像对应单通道特征编码，三维图像也对应单通道特征编码，从而构建起清晰的映射关系。由于X光片是三维CT图像的积分投影，二维X光片上的一个像素与经过该点的X射线上的所有体素相关。通过构建这种显式映射，可直观地确立二维图像像素与三维图像体素之间的对应关系。

在机器学习和深度学习的框架下，若有充足的监督数据，便可构建回归模型，且模型参数可通过深度神经网络进行优化。但在实际应用中，临床场景下的成对监督数据往往匮乏。这并非意味着医院采集的医学图像数量不足，实际上，医院采集了大量医学图像。然而，受科研伦理和医学伦理限制，涉及人体的数据使用极为谨慎。若缺乏成对数据，可尝试其他途径。采用统计模型，通过少量非零系数描述三维图像。无需成对数据，仅从一张X 光片出发，结合先验统计模型，即可估计形变场，进而生成X 光片对应的三维图像。

在优化模型时，常通过在二维平面上对比体绘制后的X光片与输入X光片的一致性来实现。然而由于训练过程未使用与X光片成对的三维图像，其三维约束依赖预先设定的统计模型。而统计模型的泛化能力存疑，难以确保能有效处理所有临床获取的二维X 光片。为应对输入X光片与形变后三维图像间的细微差异，可采用反馈机制，通过对变形场进行残差估计，以提升重建质量。

8.3 实例：物理驱动的自监督光场显微重建网络

光场显微镜（Light-field microscopy, LFM）及其变体技术为活体高速三维成像提供了强大工具，然而其实际应用受限于重建方法在速度、保真度与泛化性之间的固有矛盾。物理驱动的自监督重建网络——SeReNet适用于非扫描式和扫描式LFM（sLFM），能够在毫秒级时间内实现近衍射极限分辨率的三维重建。

SeReNet利用光场四维信息先验，在重建质量上优于当前主流深度学习方案，并可用在强噪声、光学像差和样本运动等复杂条件下。其处理速度较传统迭代断层扫描技术提升700 倍，大幅提高了数据吞吐量。可选轴向微调模块可进一步提升轴向分辨率，但会轻微影响泛化性能。通过在活细胞、斑马鱼胚胎与幼体、线虫及小鼠脑部等多种样本进行验证。集成SeReNet的sLFM系统实现了24小时连续三维亚细胞成像，成功捕获超过30万个细胞的大规模细胞间动态过程，为免疫反应和神经活动等研究提供了有力工具。

目前已有CARE、VCD-Net、HyLFM-Net等有监督深度学习方法用于降低光场计算成本，但这些方法在多样本或复杂成像条件下存在分辨率低、泛化差的问题，尤其对sLFM数据优化不足。有监督方法对真值数据的依赖也限制了其应用，训练数据获取困难、多样性有限，导致模型在未知数据上表现不佳。尽管已有研究将物理模型嵌入网络以提升可解释性，但由于复杂成像环境中多角度数据间存在相位相关性，这类方法难以适用于LFM的四维测量场景。将光场数据简单视为二维图像序列会丢失相位信息，导致性能甚至不及传统迭代算法。

基于波动光学的精确点扩散函数（PSF）建模是实现高分辨率重建的关键。因此，开发不依赖监督数据、兼顾速度与精度的LFM重建算法仍是核心挑战。SeReNet [26]通过将成像物理过程嵌入网络架构，实现了完全自监督训练，无需训练数据对，通过最小化重建结果的前向投影与原始测量间差异进行优化。该方法避免了过度估计，充分利用四维数据的高自由度，有效应对噪声、运动及像差等问题。

经数值模拟与实验验证，SeReNet在速度、分辨率、数据通量、泛化性及鲁棒性方面均优于当前最先进方法，同时适用于非扫描LFM与sLFM，重建速度较迭代方法提升近三个数量级，在不同样本和显微镜数据上均表现出一致的性能。

训练SeReNet

SeReNet的训练流程包括数据预处理、网络重建与自监督投影三个主要阶段。在训练之前，原始光场数据需经过TW-Net进行样本运动校正和preDAO模块进行光学像差估计与校正。网络的核心由三个模块组成：深度分解模块执行数字重聚焦，生成初始焦堆栈；去模糊与融合模块通过多层3D卷积结构将焦堆栈转换为高保真的3D体积数据；自监督模块则利用4D波动光学PSF对重建体积进行前向投影，并通过比较投影结果与原始输入计算损失函数，驱动网络参数优化。

在sLFM中，扫描光场图像（M）定义于四维相空间，包括二维空间坐标$\mathbf{x} = (x_1, x_2)$ 和二维角度坐标$\mathbf{u} = (u_1, u_2)$。sLFM的实际测量可表示为$M(\mathbf{x}, \mathbf{u}_j)$，其中$j$ 为角度索引。角度$\mathbf{u}_j$ 对应的点扩散函数为$H(\mathbf{x}, z, \mathbf{u}_j)$，$z$ 为轴向坐标。传统迭代层析方法通过不断减小第j 个角度测量$M(\mathbf{x}, \mathbf{u}_j)$ 与估计体积$\hat{g}_k(\mathbf{x}, z)$ 在相应角度下的前向投影$P_k(\mathbf{x}, \mathbf{u}_j)$ 之间的误差实现重建，其中投影积分写为：

$$P_k(\mathbf{x}, \mathbf{u}_j) = \int_z [\hat{g}_k(\mathbf{x}, z) * H(\mathbf{x}, z, \mathbf{u}_j)]\, dz \tag{8.1}$$

这里$*$ 表示在$\mathbf{x}$平面上的二维卷积操作。尽管该方法不受监督，但其迭代优化极为耗时。SeReNet则通过学习从$M(\mathbf{x}, \mathbf{u})$到体积估计$\tilde{g}(\mathbf{x}, z)$ 的端到端映射，实现了高效重建。网络结构包含约19.5万个参数，由深度分解、去模糊融合和自监督三大模块构成。深度分解模块借鉴数字重聚焦思想，通过对多角度图像进行位移编码提取深度信息：

$$g_a(\mathbf{x}, \mathbf{u}, z) = M(\mathbf{x} - \alpha z\mathbf{u}, \mathbf{u}) = M(\mathbf{x}, \mathbf{u}) * \delta(\mathbf{x} - \alpha z\mathbf{u}) \tag{8.2}$$

其中$\delta(\cdot)$为脉冲函数，α为缩放因子，$\alpha\mathbf{u}$代表多角度PSF的斜率图。该操作实现了轴向层面的粗分离。去模糊融合模块将角度维度信息重整至通道维，通过九层3D卷积与三层线性插值实现非相干孔径合成与高保真重建：

$$\tilde{g}(\mathbf{x}, z) = f_\theta(g_a(\mathbf{x}, \mathbf{u}, z))$$

其中θ为待优化的网络参数。自监督模块每次随机选取21个角度，利用波动光学PSF对输出体积$\tilde{g}(\mathbf{x}, z)$进行前向投影，计算与实测数据之间的损失：

$$\sum_j \text{loss}_{\text{NLL-MPG}}(\tilde{P}(\mathbf{x}, \mathbf{u}_j), M(\mathbf{x}, \mathbf{u}_j))$$

其中$\tilde{P}(\mathbf{x}, \mathbf{u}_j) = \int_z [\tilde{g}(\mathbf{x}, z) * H(\mathbf{x}, z, \mathbf{u}_j)]\, dz$。该损失函数在训练中引导网络收敛，而其角度随机选择策略有利于节省显存并提高效率。

NLL-MPG 损失函数

对于给定真实体图像$g(\mathbf{x}, z)$，sLFM 的理想成像过程表示为：

$$M'(\mathbf{x}, \mathbf{u}_j) = \int_z g(\mathbf{x}, z) * H(\mathbf{x}, z, \mathbf{u}_j), dz,$$

其中$M'(\mathbf{x}, \mathbf{u}_j)$ 表示无噪声的空间-角度测量。然而，在实际应用中，各种类型的噪声不可避免。其中，读出不确定性遵循高斯分布，而光子散粒噪声遵循泊松分布。因此，实际观测$M(\mathbf{x}, \mathbf{u}_j)$ 更好建模为遵循混合泊松-高斯（MPG）分布，可表示为：

$$M(\mathbf{x}, \mathbf{u}j) = \text{Poisson}(\tau M'(\mathbf{x}, \mathbf{u}j)) + n,$$

其中Poisson$(\cdot)$ 是泊松过程，τ 是单位时间到达的光子数，$n \sim \mathcal{N}(\mu, \sigma^2)$ 是遵循均值为μ、方差为σ^2 的高斯分布的噪声。在SeReNet 的训练过程中，自监督模块中的前

向投影$\tilde{P}(\mathbf{x},\mathbf{u}j)$ 可视为估计$\tilde{g}(\mathbf{x},z)$ 的理想空间- 角度测量。若算法收敛，$\tilde{P}(\mathbf{x},\mathbf{u}j)$ 将等价于$M'(\mathbf{x},\mathbf{u}j)$，噪声模型可表达为：

$$M(\mathbf{x},\mathbf{u}_j)=\text{Poisson}(\tilde{P}(\mathbf{x},\mathbf{u}_j))+n.$$

$p(M)$ 是$M(\mathbf{x},\mathbf{u}j)$ 的概率密度函数，在MPG 噪声模型中泊松和高斯分量独立。给定$\tilde{P}$、μ、σ^2 时M 的似然是MPG 噪声的联合概率函数，推导为：

$$p(M|\tilde{P},\mu,\sigma)=\prod_{\mathbf{x},\mathbf{u}}\sum_{\tau=0}^{+\infty}\frac{(\tilde{P})^{\tau}\exp(-\tilde{P})}{\tau!}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(M-\mu-\tau)^2}{2\sigma^2}\right).$$

遵循最大似然原则，基于负对数似然并使用斯特林近似处理阶乘算子。损失函数可公式化为：

$$\begin{aligned}&\text{lossNLL-MPG}=\\&-\log\left(\prod_{\mathbf{x},\mathbf{u}}\sum_{\tau=0}^{+\infty}\frac{(\tilde{P}/\tau)^{\tau}\exp(\tau-\tilde{P})}{\sqrt{2\pi\tau}}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(M-\mu-\tau)^2}{2\sigma^2}\right)\right).\end{aligned}$$

高斯参数μ 和σ 是相机内在属性，与样本强度无关但受读出时间影响。我们根据特定曝光时间表征μ 和σ，使用提供的无照明暗图像或LF 测量的角落视图。

SeReNet 与DINER 的模型架构差异

DINER与SeReNet均受NeRFs的启发，但模型架构与性能有显著差异。DINER以多层感知机为核心，其首先初始化三个可训练的3D哈希表，随后通过MLP 将其映射为最终输出体积。该模型的所有参数需针对每一个新的光场显微（LFM）数据集进行重新训练，导致其泛化能力差且重建速度缓慢。在自监督训练中，DINER通过直线传播规则将预测体积投影为多角度图像，并与输入图像计算损失。

相比之下，SeReNet采用CNN结构，其在CNN前端引入深度分解模块，通过深度编码卷积增强对深度信息的提取能力。此外，SeReNet 在训练中融合了多角度波动光学点扩散函数（PSFs），更准确建模显微成像过程，改善了收敛效果与重建精度。SeReNet仅训练网络参数，不绑定特定场景，因此具备优异泛化能力，能够直接应用于真实生物数据，推理速度可达毫秒级，显著优于DINER。

8.4 小结

三维重建是构建从二维测量数据，如CT重建中的二维X光片到对应三维图像的映射模型，可通过有监督学习构建端到端模型。配准方法以三维图像作为参考模型，通过估计二维X光片对应的三维图像形变，也可得到目标三维图像。

数据匮乏是首要难题，构建重建模型需要优化大量参数，然而可用数据有限。虽可通过数据增强或仿真手段扩充数据量，但这些模拟数据与临床真实图像间存在

域差异，难以完全契合实际需求。此外，重建过程需处理跨维度映射，这涉及语义对应问题。二维X光片由各类解剖结构的积分投影构成，结构重叠、边界模糊，要从中重建出精确的三维图像，并清晰区分各解剖结构，与真实情况高度吻合，即便拥有大量训练数据，仍是极具挑战性的病态问题。

当前基于单采样点、正交采样点或少量采样点进行三维图像重建，其精度难以保证。依据中心切片定理，CT 重建需覆盖0 - 180 度的积分图像，但在小采样点CT重建工作中，仅从少量采样点重建可信的三维结构，因问题病态性强，从单一积分图像难以唯一确定三维图像，并在重建时保持结构细节。

基于基础模型的三维重建利用**大规模深度估计/表面法线估计模型**，使用在数百万张图像-深度图对上训练的基础模型，从单张图像预测出稠密的深度图。再通过相机参数将深度图反向投影到三维空间，形成点云或网格。但生成的几何可能不够精细，且通常需要后续的多视角融合或优化步骤来改善质量。

基于文本到3D的生成模型中，分数蒸馏采样（Score Distillation Sampling, SDS）利用像Stable Diffusion这样的文生图基础模型作为“老师”，指导3D表示，如NeRF、高斯溅射的优化过程。通过SDS技术，让3D模型在不同角度下渲染的图片都与文本提示词在扩散模型眼中看起来“合理”。其无需3D训练数据，可通过文本自由生成。但过程计算量大、耗时较长。

基于多视图生成模型使用专门训练来生成多视角一致图像的基础模型，如Zero-1-to-3, MVDream。输入一张图片和指定的相机变换角度，模型可以生成该物体在新视角下的图像，再使用传统的SFM/MVS算法对这些生成的多视角图像进行三维重建。结合了生成模型的理解能力和传统几何重建的精度，能产生高质量的三维模型。

端到端的单图到3D基础模型训练一个庞大的模型，像文生图模型一样，直接从一张图片或文本，生成高质量的3D资产。基础模型编码了关于物体结构、材质、光照的常识，能补全被遮挡的部分，生成合理且完整的三维模型。可处理各种常见物体和场景，而非仅限于特定类别。许多相关模型已开源，推动了技术的快速普及和应用。

医学图像关注“内部三维构造”的分析，除生物医学图像处理外，其他领域诸如地质勘探的地层构造分析、考古学的文物内部结构探测也会关注“内部复杂结构”，但医学图像的核心差异在于对“内部三维解剖/病理结构”的精准分析需求，这与自然图像侧重“表面形态”的分析逻辑截然不同：**自然图像的三维重建**多关注“表面轮廓”，如从照片重建建筑外观。**医学图像的三维重建**需覆盖“内部多层结构”，如从CT重建肺部时，需同时呈现肺实质、支气管、血管、结节等内部组织，且需保证结构的空间位置与解剖学一致，为临床诊断提供精准依据。

这种特殊性也决定了医学图像处理技术需突破“跨模态适配”“小样本泛化”等关键问题，特别是不同模态的医学图像表观差异极大，且临床数据常因隐私限制、标注成本高而规模有限。

第九章　总结

本讲义以基于生物医学图像处理为核心，融合计算机视觉与图像计算的基础理论与实践技术，覆盖“传统方法→深度学习→多模态大模型”等技术。

跨模态、跨场景的图像对象

生物医学图像不仅包含临床常见的医学影像，还包括生物研究场景的显微图像与特殊场景的自然关联图像，其发展与成像技术的进步紧密相关：医学影像：自1895年X射线发现后，CT、MRI、超声、PET等多模态影像逐步进入临床，构成生物医学图像处理的核心对象，其特点是“关注人体内部解剖结构与病理变化”；显微图像：包括病理切片、细胞荧光图像、电子显微镜图像等，聚焦“亚细胞级、分子级结构”，是基础医学研究的关键数据源；自然关联图像：如皮肤病变照片、内窥镜视频帧等，虽成像原理与自然图像类似，但需结合医学语义进行分析。这些图像的共同特点是“模态多样性”与“维度复杂性”，既包含二维图像，也涵盖三维体数据，且不同模态的成像机制差异极大，这也决定了处理技术需具备“跨模态适配能力”。

从基础任务到多模态融合

课程围绕计算机视觉的核心任务展开，同时追踪技术范式的迭代，包括基础视觉任务：图像分割、目标检测、图像配准、三维重建，这些任务是生物医学图像处理的“基石”；技术范式演进包括：

传统方法：基于形态学算子、阈值分割、边缘检测的手工规则方法。

机器学习方法：基于SVM、随机森林的监督分类，基于K-Means的无监督聚类，依赖手工设计特征。

深度学习方法：以CNN、Transformer（ViT）为核心，实现端到端的语义学习，解决复杂分割、配准问题。

多模态大模型：以CLIP、SAM为代表，通过“图文对齐”实现零样本分割，大幅降低标注依赖。

特别地课程重点关注“语言模型与图像处理的融合”，尽管传统图像处理与语言无关，但多模态大模型的出现让“文本增强图像表征”成为可能（如用医学报告引导影像分割），这也是当前技术的核心。

表 9.1: 基于AI图像处理技术发展

时间阶段	核心技术	代表技术	应用场景
1950s-1970s	计算机基础+数字图像雏形	简单灰度变换、边缘检测	X射线片的初步降噪与增强
1980s-1990s	传统机器学习+早期神经网络	SVM分类、Hopfield神经网络、形态学分割	细胞计数、简单器官阈值分割
2010s-2020s	深度神经网络（CNN/Transformer）	U-Net分割、ResNet特征提取、3D 重建模型	肿瘤精准分割、多模态影像配准
2020s-至今	多模态预训大模型	SAM零样本分割、CLIP图文对齐、Medical SAM	临床快速分割、罕见病影像分析

需特别注意的是，CNN并非近10年新发明，早在1980s的“新感知机”中就已具备卷积、池化等核心特征，但其真正爆发是在2012年AlexNet证明“深度CNN+大数据”的有效性后，这一技术迁移至生物医学领域，彻底改变了医学影像的分析范式。我们尝试回答两个问题：

问题1：AI医生能否取代人类医生？

当前不能取代，但已成为关键辅助工具，核心原因在于“技术局限性”与“医疗场景的特殊性”。

技术局限性：AI模型的“泛化能力”与“可解释性”不足，例如基于某医院CT数据训练的肿瘤分割模型，在面对其他医院的低剂量CT时可能出现漏诊；且AI无法解释“为何判定为肿瘤”，仅能输出概率结果，缺乏临床决策所需的“逻辑链条”。

医疗场景特殊性：医疗诊断需结合“影像+病史+体征”的多维度信息，AI目前仅能处理影像数据，无法整合患者的既往病史、生活习惯等非影像信息。同时，医疗决策涉及伦理责任，当前AI的错误风险无法由机器承担，需人类医生最终把关。

尽管AI可生成影像分析报告，但临床最终报告仍需影像科医生审核签字。达芬奇手术机器人等设备也需外科医生操作，AI仅负责“定位精度优化”“误差校正”等辅助功能。

问题2：生物医学图像处理是否已沦为纯工程问题？

当前仍有大量研究问题，工程化是技术落地的手段而非终点，核心原因在于“数据瓶颈”与“任务复杂性”：**数据瓶颈未突破**：AI模型依赖“大规模、高质量、均衡化”的数据，但生物医学数据存在“伦理约束”，患者隐私限制数据公开、“标注成本高”、“类别不均衡”等问题，如何在数据有限的情况下提升模型性能仍是核心研究方向。**复杂任务待解决**：临床存在大量“未被满足的需求”，如“动态影像的实时分析”（手术中实时追踪病灶）、“多模态数据的融合分割”、“未知异常的检测”，这些任务无法通过“单纯调参+ 加算力”解决，需突破现有模型架构。**工程化的挑战**：即便有成熟模型，如何将其落地到临床设备、满足实时性要求、适配不同医院的设备差异，仍需解决“模型压缩”“跨设备泛化”等工程问题，而这些工程问题本身也依赖技术创新。

生物医学图像处理的临床应用场景：贯穿诊疗全流程

生物医学图像处理并非孤立的技术，而是深度融入临床诊疗的“全链条”，从图

像采集到术后评估，均发挥关键作用：

图像采集阶段：成像与三维重建 将原始扫描数据转化为可分析的图像/体数据，或从二维图像重建三维结构。CT扫描后通过“滤波反投影”算法重建肺部三维模型，为后续分割、诊断提供基础；内窥镜手术中，通过多视角图像重建手术区域的三维环境，辅助医生判断解剖位置。

影像分析阶段：病灶检测与量化自动识别影像中的异常区域，实现“检测-分割-分级”的一体化分析。用Medical SAM分割MRI中的脑肿瘤，同时计算肿瘤体积、边缘不规则度，为“良恶性判断”提供量化依据；病理切片图像中，自动计数肿瘤细胞数量，评估病变严重程度。

诊断与治疗阶段：决策支持与手术辅助 为临床医生提供诊断参考，优化治疗方案，降低手术风险。肺癌诊疗中，AI分析CT影像后输出“病灶位置、分期建议”，辅助医生制定手术/化疗方案；神经外科手术中，AI 实时追踪脑组织位移，避免手术器械损伤功能区。

术后评估阶段：疗效监测与随访 对比术前术后影像，评估治疗效果，监测复发风险。例如在肝癌介入治疗后，AI对比治疗前后的CT影像，分析肿瘤血供变化（如“强化区域缩小”提示治疗有效），同时定期随访影像，及时发现复发病灶。

大模型的技术范式变革：从“训练模型”到“复用模型”

自2019年GPT系列问世以来，大模型虽仅发展数年，却已彻底改变生物医学图像处理的技术思路，核心转变是“从‘从头训练模型’到‘复用预训模型’”。

大模型出现前：“算力+数据”的双重依赖在大模型普及前，基于深度学习解决生物医学任务需攻克两大难关：算力瓶颈：训练一个3D U-Net需单块20GB以上显存的GPU，复杂模型甚至需多卡集群。数据瓶颈：需大规模高质量标注数据，而临床数据常因隐私、标注成本高而稀缺，导致模型易过拟合、泛化能力差。

彼时的技术重心是“如何设计更高效的网络架构”“如何通过数据增强扩充样本”，以在有限资源下提升模型性能。

大模型出现后：“迁移学习”成为核心范式。大模型的出现让“迁移学习”从“可选方案”变为“必选路径”，核心逻辑是“复用预训模型的通用表征，适配特定生物医学任务”：预训模型的本质：多在海量自然图像或通用多模态数据上训练，具备“模态无关”“语义丰富”的特征提取能力，例如，CLIP的视觉编码器可识别“圆形结构”“边界突变”等通用特征，这些特征在医学图像中同样适用；**模型迁移**：将预训大模型的权重作为目标任务模型的初始参数；**特征迁移**：冻结预训模型的特征提取模块，仅训练下游任务的小模块；**提示迁移**：无需训练，仅通过设计合理提示，驱动大模型完成特定任务。

从“自然图像大模型”到“医学专用大模型”为进一步适配生物医学场景，研究者通过“医学数据微调”将自然图像大模型改造为“医学专用大模型”：- 例：将自然图像预训的CLIP，用CT、MRI、病理切片等医学图像+对应的文本报告（如

“右肺上叶磨玻璃结节”）微调，得到“Medical CLIP”；将SAM用医学影像标注数据微调，得到“Medical SAM”；- 目标：提升模型对医学语义的理解能力，解决自然图像大模型对医学特殊结构识别精度不足的问题。

未来展望：从“辅助工具”到“协同伙伴”

尽管当前生物医学图像处理仍面临诸多挑战，但技术发展趋势已逐渐清晰，未来将向“更高精度、更低依赖、更强协同”方向演进：

标注成本降低：伪标签与半监督学习预训大模型可生成“伪标注”，医生仅需少量修正即可得到高质量标签，大幅降低标注成本；同时，半监督/无监督学习技术将进一步发展，让模型在“少量标注+大量无标注数据”下实现高精度分割，突破数据瓶颈。

多模态融合深化：影像+文本+病史 未来的模型将不再局限于“图像分析”，而是整合“影像数据+电子病历+医学文献”的多维度信息，例如，结合患者的“糖尿病病史”与CT影像，更精准地判断肺部感染的类型，实现“个性化诊疗”。

临床可解释性提升：从“黑箱”到“透明化” 通过“注意力可视化”“特征归因分析”等技术，让AI模型解释“为何做出该判断”，同时结合解剖学知识图谱，让输出结果符合临床逻辑，增强医生对AI的信任度。

边缘计算与实时化：适配临床场景 将大模型压缩为“轻量级模型”，部署到CT机、内窥镜等边缘设备，实现“扫描完成即出分析结果”的实时性要求；同时，开发便携式AI辅助设备，让技术下沉到基层医院，提升医疗资源可及性。

本课程不仅覆盖生物医学图像处理从传统方法到多模态大模型的核心技术，无论是选择传统形态学方法进行快速分割，还是用预训大模型实现零样本分析，最终都需围绕“临床需求”展开。未来，随着AI技术的进一步发展，生物医学图像处理将从“辅助工具”逐步转变为医生的“协同伙伴”，但技术的进步始终需以“安全、可靠、伦理”为前提。对于学习者而言，不仅需掌握模型与算法的实现细节，更需理解技术的局限性与临床场景的复杂性，才能让技术真正服务于医学，推动精准医疗的发展。

当前AI在生物医学图像处理中的核心能力经过多年发展，AI已从“辅助工具”逐步升级为“具备全栈能力的技术体系”，具有全栈式处理能力：端到端闭环AI可覆盖从“图像采集”到“报告生成”的全流程，无需人工干预：直接辅助临床决策。通过对比学习、图文对齐等技术，AI可整合“影像、文本、基因”等多模态数据，提升分析精度：预训大模型的特征具有“通用性”，可适配多种生物医学任务，无需为每个任务单独训练特征提取器：基于预训大模型的“小样本学习”，可在标注数据极少的场景下实现高精度任务：可实现该罕见病的病灶分割，解决了“罕见病数据稀缺”的行业痛点，为小众疾病的诊疗提供技术支持。

当前，众多科技公司与医疗企业已推出针对生物医学图像处理的商业化、开源解决方案，覆盖多场景需求：GE医疗的Carestream AI可辅助病理切片的肿瘤细胞

计数；Google Health 的眼底影像分析系统可筛查糖尿病视网膜病变；IBM Watson Health聚焦多模态影像的综合诊断；开源工具3D Slicer的深度学习扩展模块集成了Medical SAM，支持医生通过交互提示快速分割器官；MONAI （Medical Open Network for AI）提供标准化的医学深度学习框架，方便研究者开发定制化模型。

尽管大模型已成为当前的技术主流，但“传统数值计算、传统机器学习、深度学习、预训大模型”仍处于“协同并存”状态，例如，在临床实践中，医生会先用传统阈值分割快速预览CT 硬组织区域，再用Medical SAM 精细分割肿瘤，最后用传统边缘检测修正分割边界。

AI将进一步向“更高精度、更强泛化、更可解释”发展，例如，通过“注意力可视化”让AI解释“为何判定为肿瘤”，如“该区域灰度值突变、边缘不规则，符合肿瘤特征”；通过“联邦学习”在保护数据隐私的前提下，实现多中心数据联合训练，提升模型的域外泛化能力。

最终，AI的目标不是“取代医生”，而是通过“技术赋能”让优质医疗资源更可及，例如，在基层医院，AI 可辅助非专科医生完成初步影像分析，缩小“城乡医疗差距”，推动精准医疗的普及。

参考文献

[1] Amit Aflalo, Shai Bagon, Tamar Kashti, and Yonina C. Eldar. Deepcut: Unsupervised segmentation using graph neural networks clustering. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 32–41, 2022.

[2] Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light, Science & Applications*, 12, 2022.

[3] Nóra Baka, Bart L. Kaptein, Marleen de Bruijne, Theo van Walsum, J. E. Giphart, Wiro J. Niessen, and Boudewijn P. F. Lelieveldt. 2d-3d shape reconstruction of the distal femur from stereo x-ray imaging using statistical shape models. *Medical image analysis*, 15 6:840–50, 2011.

[4] Yin Cai, Yin Cai, M. Julius Hossain, Jean-Karim Hériché, Antonio Zaccaria Politi, Antonio Zaccaria Politi, Nike Walther, Birgit Koch, Birgit Koch, Malte Wachsmuth, Bianca Nijmeijer, Moritz Kueblbeck, Marina Martinic-Kavur, Rene Ladurner, Rene Ladurner, Stephanie Alexander, Jan-Michael Peters, and Jan Ellenberg. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature*, 561:411 – 415, 2018.

[5] Jeremy G. Carlton, Hannah Jones, and Ulrike S. Eggert. Membrane and organelle dynamics during cell division. *Nature Reviews Molecular Cell Biology*, 21:151–166, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[7] Shiyi Cheng, Sipei Fu, Yumi Mun Kim, Weiye Song, Yunzhe Li, Yujia Xue, Ji Yi, and Lei Tian. Single-cell cytometry via multiplexed fluorescence prediction by label-free reflectance microscopy. *Science Advances*, 7, 2020.

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021.

[9] Gendrin Christelle et al. Monitoring tumor motion by real time 2d/3d registration during radiotherapy. *Radiotherapy and Oncology Journal of the European Society for Therapeutic Radiology and Oncology*, 102(2):274–280, 2012.

[10] Emil M. Christiansen, Samuel J. Yang, David M. Ando, Ashkan Javaherian, Gabriel Skibinski, Scott Lipnick, Elen Mount, Alison O¡¯Neil, Kevan Shah, Alice K. Lee, et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173:792–803.e19, 2018.

[11] Jan Oscar Cross-Zamirski, Praveen Anand, Guy B. Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3802–3811, 2023.

[12] Jan Oscar Cross-Zamirski, Elizabeth Mouchet, Guy B. Williams, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Label-free prediction of cell painting from brightfield images. *Scientific Reports*, 12, 2021.

[13] Omri Efroni, Dvir Ginzburg, and Dan Raviv. Spectral teacher for a spatial student: Spectrum-aware real-time dense shape correspondence. In *2022 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2022.

[14] Hongqing Han, Mariia Dmitrieva, Alexander Sauer, Ka Ho Tam, and J. Rittscher. Self-supervised voxel-level representation rediscovers subcellular structures in volume electron microscopy. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1873–1882, 2022.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021.

[16] Xingyi He He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. *ArXiv*, abs/2501.07556, 2025.

[17] Benjamin Hou, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. Predicting slice-to-volume transformation in presence of arbitrary subject motion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 296–304. Springer, 2017.

[18] Luzhe Huang, Yuzhu Li, Nir Pillar, Tal Keidar Haran, William Dean Wallace, and Aydogan Ozcan. A robust and scalable framework for hallucination detection in virtual tissue staining and digital pathology. *Nature biomedical engineering*, 2024.

[19] Jaroslav Icha, Michael Weber, Jennifer C Waters, and Caren Norden. Phototoxicity in live fluorescence microscopy, and how to avoid it. *BioEssays*, 39, 2017.

[20] Sara Imboden, Xuanqing Liu, Marie C. Payne, Cho-Jui Hsieh, and Neil Y. C. Lin. Trustworthy in silico cell labeling via ensemble-based image translation. *Biophysical Reports*, 3, 2023.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2016.

[22] Yikun Jiang, Yuru Pei, Tianmin Xu, Xiaoru Yuan, and Hong yan Zha. Toward semantically-consistent deformable 2d-3d registration for 3d craniofacial structure estimation from a single-view lateral cephalometric radiograph. *IEEE Transactions on Medical Imaging*, 44:685–697, 2024.

[23] YoungJu Jo, Hyun-Soo Cho, Wei Sun Park, Geon Kim, DongHun Ryu, Young Seo Kim, Moosung Lee, Sangwoo Park, Mahn Jae Lee, Hosung Joo, HangHun Jo, Seon-Gyeong Lee, Sumin Lee, Hyun-Seok Min, Won Do Heo, and Yongkeun Park. Label-free multiplexed microtomography of endogenous subcellular dynamics using generalizable deep learning. *Nature Cell Biology*, 23:1329 – 1337, 2021.

[24] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768, 2024.

[25] Haofu Liao, Wei-An Lin, Jiarui Zhang, Jingdan Zhang, Jiebo Luo, and S. Kevin Zhou. Multiview 2d/3d rigid registration via a point-of-interest network for tracking and triangulation (point^2). *CoRR*, abs/1903.03896, 2019.

[26] Zhi Lu, Manchang Jin, Shuai Chen, Xiaoge Wang, Feihao Sun, Qi Zhang, Zhifeng Zhao, Jiamin Wu, Jingyu Yang, and Qionghai Dai. Physics-driven self-supervised learning for fast high-resolution robust 3d reconstruction of light-field microscopy. *Nature Methods*, 22:1545 – 1555, 2025.

[27] Zhi Lu, Siqing Zuo, Minghui Shi, Jiaqi Fan, Jingyu Xie, Guihua Xiao, Li Yu, Jiamin Wu, and Qionghai Dai. Long-term intravital subcellular imaging with confocal scanning light-field microscopy. *Nature Biotechnology*, pages 1–12, 2024.

[28] Primoz Markelj, Dejan Tomazevic, Bostjan Likar, and Franjo Pernus. A review of 3d/2d registration methods for image-guided interventions. *Medical image analysis*, 16(3):642–661, 2012.

[29] Luke Melas-Kyriazi, C. Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8354–8365, 2022.

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[31] Shun Miao, Z Jane Wang, and Rui Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE Transactions on Medical Imaging*, 35(5):1352–1363, 2016.

[32] Zeinab Navidi, Jun Ma, Esteban A. Miglietta, Le Liu, A.E. Carpenter, Beth A. Cimini, Benjamin Haibe-Kains, and Bo Wang. Morphodiff: Cellular morphology painting with diffusion models. *bioRxiv*, 2024.

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023.

[34] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15:917–920, 2018.

[35] Pietro Perona, Takahiro Shiota, and Jitendra Malik. Anisotropic diffusion. In *Geometry-Driven Diffusion in Computer Vision*, pages 73–92. 1994.

[36] William A. Prinz, Alexandre Toulmay, and Tamas Balla. The functional universe of membrane contact sites. *Nature Reviews Molecular Cell Biology*, 21:7–24, 2020.

[37] Yair Rivenson, Tairan Liu, Zhensong Wei, Yibo Zhang, and Aydogan Ozcan. Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. *Light, Science & Applications*, 8, 2018.

[38] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydn, Jonathan E. Zuckerman, Thomas Chong, Anthony E. Sisk, Lindsey Westbrook, William D. Wallace, and Aydogan Ozcan. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3:466 – 477, 2018.

[39] Nico Scherf and Jan Huisken. The smart and gentle microscope. *Nature Biotechnology*, 33:815–818, 2015.

[40] Luca Scorrano, Maria Antonietta De Matteis, Scott Emr, Francesca Giordano, Gy?rgy Hajn¨®czky, Beno?t Kornmann, Laura L. Lackner, Tim P. Levine, Luca Pellegrini, Karin Reinisch, et al. Coming together to define membrane contact sites. *Nature Communications*, 10:1287, 2019.

[41] Srijit Seal, Maria-Anna Trapotsi, Ola Spjuth, Shantanu Singh, Jordi Carreras-Puigvert, Nigel Greene, Andreas Bender, and Anne E Carpenter. Cell painting: a decade of discovery and innovation in cellular imaging. *Nature methods*, pages 1–15, 2024.

[42] Daniel Toth, Shun Miao, Tanja Kurzendorfer, Christopher A. Rinaldi, Rui Liao, Tommaso Mansi, Kawal Rhode, and Peter Mountney. 3d/2d model-to-image registration by imitation learning for cardiac procedures. *International Journal of Computer Assisted Radiology and Surgery*, 13(8):1141–1149, 2018.

[43] David A. Van Valen, Takamasa Kudo, Keara M. Lane, Derek N. Macklin, Nicolas T. Quach, Mialy Defelice, Inbal Maayan, Yu Tanouchi, Euan A. Ashley, and Markus W. Covert. Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS Computational Biology*, 12, 2016.

[44] Alex M. Valm, Sarah Cohen, Wesley R. Legant, Justin Melunis, Uri Hershberg, Eric Wait, Andrew R. Cohen, Michael W. Davidson, Eric Betzig, and Jennifer

Lippincott-Schwartz. Applying systems-level spectral imaging and analysis to reveal the organelle interactome. *Nature*, 546:162–167, 2017.

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[46] Ritvik Vasan, Alexandra J. Ferrante, Antoine Borensztejn, Christopher L. Frick, Philip Garrison, Nathalie Gaudreault, Saurabh S. Mogre, Fatwir S Mohammed, Benjamin Morris, Guilherme G. Pires, Daniel Saelid, Susanne M. Rafelski, Julie A. Theriot, and Matheus P. Viana. Interpretable representation learning for 3d multi-piece intracellular structures using point clouds. *Nature Methods*, 22:1531 – 1544, 2025.

[47] Matheus P. Viana et al. Integrated intracellular organization and its variations in human ips cells. *Nature*, 613:345 – 354, 2023.

[48] Zhengyang Wang, Yaochen Xie, and Shuiwang Ji. Global voxel transformer networks for augmented microscopy. *Nature Machine Intelligence*, 3:161 – 171, 2020.

[49] Håkan Wieslander, Ankit Gupta, Ebba Bergman, Erik Hallström, and Philip John Harrison. Learning to see colours: Biologically relevant virtual staining for adipocyte cell images. *PLoS ONE*, 16, 2021.

[50] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[51] Ronald Xie, Kuan Pang, Gary D Bader, and Bo Wang. Maester: Masked autoencoder guided segmentation at pixel resolution for accurate, self-supervised subcellular structure recognition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3292–3301, 2023.

[52] Weimin Yu, Chengwen Chu, Moritz Tannast, and Guoyan Zheng. Fully automatic reconstruction of personalized 3d volumes of the proximal femur from 2d x-ray images. *International journal of computer assisted radiology and surgery*, 11(9):1673–1685, 2016.

[53] Weimin Yu, Moritz Tannast, and Guoyan Zheng. Non-rigid free-form 2d–3d registration using a b-spline-based statistical deformation model. *Pattern Recognition*, 63:689–699, 2017.

[54] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Brian D. Piening, Carlo Bifulco, Mu-Hsin Wei, Hoifung Poon, and Sheng Wang. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *Nature methods*, 2024.

[55] Guoyan Zheng. Effective incorporating spatial information in a mutual information based 3d–2d registration of a ct volume to x-ray images. *Computerized Medical Imaging and Graphics*, 34(7):553–562, 2010.

[56] Guoyan Zheng. Personalized x-ray reconstruction of the proximal femur via intensity-based non-rigid 2d-3d registration. In *Medical Image Computing and Computer-Assisted Intervention*, pages 598–606, 2011.

[57] Donghao Zhou, Chunbin Gu, Junde Xu, Furui Liu, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Repmode: Learning to re-parameterize diverse experts for subcellular structure prediction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3312–3322, 2022.

[58] Donghao Zhou, Chunbin Gu, Junde Xu, Furui Liu, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Repmode: learning to re-parameterize diverse experts for subcellular structure prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3312–3322, 2023.

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[60] L Zollei, E Grimson, Alexander Norbash, and W Wells. 2d-3d rigid registration of x-ray fluoroscopy and ct images using mutual information and sparsely sampled histogram estimators. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[61] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *ArXiv*, abs/2304.06718, 2023.