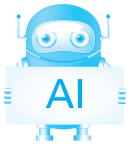
第 5 章



图像生成技术

本章介绍了图像生成领域的多种模型和技术，包括 VAE、GAN、扩散模型、自回归模型等，并讨论了它们在艺术创作、娱乐、医疗、广告、自动驾驶等多个领域的广泛应用。同时，还探讨了图像风格迁移、超分辨率重建、视频生成和医疗影像合成等重要应用方向，以及这些技术面临的挑战和未来发展方向，如提高生成质量、优化计算效率、增强模型泛化能力和解决伦理问题等。

5.1 图像生成的模型

生成式 AI 在图像生成领域已经取得显著进展，能够创造出逼真且富有创意的图像内容。这些技术不仅改变了艺术创作、设计和娱乐行业的工作方式，还为科学研究提供了新的工具。

(1) 在图像生成技术中，可以使用 VAE 模型，通过学习输入数据的概率分布来生成新样本。其中，编码器将图像映射到一个潜在空间做参数化分布，解码器则从该潜在空间中采样并重构原始图像。但 VAE 生成图像的质量可能不如其他方法高，尤其是在复杂数据集上；潜在空间的语义解释性较差。

(2) 在图像生成技术中，可以使用 GAN 模型，其中的神经网络生成器尝试创建看起来真实的假图像，而神经网络判别器则试图区分真实图像与生成的假图像。两者在训练过程中不断优化，使得生成器最终能够产生高质量的图像。GAN 能够在高维数据上生成非常逼真、细节丰富的图像，灵活性强，但训练过程不稳定，容易出现模式崩溃或梯度消失等问题。

另一方面，流模型生成的图像质量较高，尤其在低维数据上表现优异，但其计算成本高昂。

5.1.1 扩散模型

扩散模型是一类生成式模型，最初由索尔·迪克斯坦等 2015 年提出，并在随后的研究中得到显著发展。这类模型逐步向数据添加噪声，然后学习逆转这个过程来生成新的样本或者恢复原始图像，可以看作是对图像生成的一种去噪过程，其特点是生成的图像质量和多样性(图 5-1)都非常出色，尤其对于复杂的自然场景。而且它的训练相对稳定，不容易出现模式崩溃问题。

图 5-1　扩散模型生成的作品

扩散模型的一些应用案例如下。

（1）图像生成：用于艺术创作、风格迁移、超分辨率重建等领域。

（2）音频合成：生成音乐旋律、语音波形等。

（3）视频生成：创建连贯的视频序列。

（4）医学影像：增强低质量医学影像，或者生成合成的训练数据，以辅助诊断算法的开发。

（5）分子设计：帮助化学家设计新型药物分子结构。

尽管扩散模型成就显著，但仍面临一些挑战。

（1）计算成本较高：由于需要经过多个步骤才能完成一次完整的前向或反向过程，因此训练和推理的时间较长。

（2）优化效率：进一步提高模型训练的速度和效果是一个重要的研究方向。

（3）理论理解不足：对于扩散模型为什么能如此有效地工作，目前仍缺乏充分的理论解释。

5.1.2　自回归模型

自回归模型是一类用于处理时间序列数据和序列生成任务的统计模型，其核心思想是基于过去的观测值来预测未来的值。通过将当前值表示为先前值的线性组合加上噪声项进行建模，即逐像素地预测下一个像素的概率分布，从而逐步构建完整图像。其特点是生成的图像质量较高，特别是对于较小尺寸的图像；它提供了明确的概率解释，适合某些特定应用。在现代机器学习中，自回归模型不仅限于线性关系，还可以扩展到非线性情况，并广泛应用于 NLP、语音合成、图像生成等领域，因其直观性和有效性，在序列建模任务中占据重要地位。

自回归模型的一些应用实例如下。

（1）文本生成。

① 字符级 RNN：使用 RNN，特别是 LSTM 或 GRU 变体，来捕捉文本序列中的长期依赖关系。训练后的模型可以根据前面的字符预测下一个字符，从而生成连贯文本片段。

② 仅限 Transformer 解码器模型：如 GPT 系列，这些模型仅包含解码器部分，利用自注意力机制有效地处理长距离依赖问题，并且能够在大规模语料库上预训练，以实现强大的文本生成能力。

（2）语音合成。DeepMind 提出了一种深度 CNN 架构 WaveNet，它采用因果卷积层来保证输出只依赖过去的时间步，实现了高质量的音频波形生成。WaveNet 可以直接从原始

音频信号中学习复杂的模式，支持多种声音类型的生成，如人类语音、乐器演奏等。

（3）图像生成。例如 PixelCNN/PixelRNN，这类模型将像素视为一维序列，按照扫描顺序（如左至右、上至下）依次生成每个像素的颜色值。尽管计算复杂度较高，但它们能够产生逼真的图片，尤其是在小尺寸图像上表现良好。

自回归模型本质上是条件概率模型，提供了对生成过程清晰的理解，有助于分析和调试。相比于一些复杂的 GAN，自回归模型通常更容易训练和稳定，尤其是在较小的数据集上。用户可以通过调整输入序列或引入额外条件变量（如类别标签）来指导生成过程，实现特定风格或内容的控制。

为了克服传统自回归模型的一些局限，研究者们提出了以下一些改进方案。

（1）Transformer 架构：通过引入自注意力机制，允许模型同时考虑所有位置之间的关系，从而更好地捕捉全局依赖。

（2）非自回归模型：尝试一次性生成整个序列，而不是逐个元素地生成，提高了速度，并减少了暴露偏差的影响。然而，这类模型往往需要特别设计，以保持生成质量。

（3）混合模型：结合自回归和非自回归的优点，例如，先使用非自回归模型生成粗略框架，再用自回归模型细化细节。

5.1.3　图像生成典型模型

在图像生成技术中，StyleGAN、BigGAN 和 DALL-E 等模型分别以高分辨率图像生成、大规模数据集上的卓越表现以及根据文本描述创造多样化图像的能力而闻名。

（1）StyleGAN：由 NVIDIA 开发，以其生成的高分辨率人脸图像而闻名，广泛应用于影视特效、游戏开发等领域。

（2）BigGAN：大规模 GAN 架构，在 ImageNet 大型数据集上展示了出色的图像生成能力。

（3）DALL-E：由 OpenAI 推出的图像生成模型，结合了 Transformer 架构和图像生成技术，能够根据文本描述生成对应的图像，支持多种风格和主题，创造独特的图像和艺术作品。

（4）Glow：一种基于流模型的图像生成框架，能够在保证高质量生成的同时实现快速推理。

（5）Stable Diffusion：一种开源的 AI 图像生成器扩散模型，支持多种风格和类型的图像创作，因高效性和易用性而受到广泛关注。

（6）Pixso AI：国产在线设计工具，集成 AI 功能，帮助设计师快速生成和编辑设计元素。

5.1.4　图像生成的应用场景

图像生成技术能够创建逼真的视觉内容或艺术化效果，广泛地应用于艺术创作、娱乐媒体、虚拟现实、游戏开发、广告设计、医疗影像合成以及自动化内容生成等多个行业和领域。

（1）艺术与设计。

① 创意辅助：艺术家和设计师可以利用图像生成技术快速生成概念图、纹理、图案等，激发灵感，并加速创作过程。作为辅助工具，它帮助艺术家探索新的创意方向，尝试不同风

格的表现形式，还可以自动生成具有特定艺术风格的作品，用于装饰、展览等多种用途。

② 风格迁移：将不同艺术作品的风格特点融合在一起，创造出独特的视觉效果，适用于绘画、摄影等多种形式的艺术创作。

（2）娱乐与媒体。

① 虚拟角色设计：用于创建游戏角色、电影角色或其他数字人物的形象，确保每个角色都有独特的外观和个性。创建更加丰富和互动的游戏环境，如 NPC 行为模拟、关卡设计等。通过 AI 驱动的虚拟人物进行直播或表演，提供全新的娱乐形式。

② VR/AR：生成逼真的虚拟环境、物体或生物，提升用户的沉浸感和交互体验，如游戏场景、虚拟旅游等，增强用户的沉浸感。

③ 影视特效制作：快速生成符合导演意图的高质量特效镜头、虚拟场景、视觉效果和场景氛围，减少实际拍摄的成本和难度，例如背景合成、特效制作等，节省后期制作时间。

④ 视频处理：实现老旧影片或低清视频的高清化，提升观看体验；也可用于实时视频通话中的画质增强。

（3）广告与营销。

① 个性化内容生成与定制：根据目标受众的特点快速生成定制化的广告素材，提高广告的相关性和吸引力，例如生成特定风格的产品图片或宣传海报。根据品牌调性和市场需求生成独特的产品包装、宣传海报等视觉材料。

② A/B 测试优化：快速生成多种版本的广告创意，用 A/B 测试找到最有效的设计方案。

③ 增强用户体验（UX）：为用户提供个性化的界面主题或背景图案，提升交互乐趣。

（4）医疗健康。

① 医学影像分析：提升 CT、MRI 等医学成像设备获取的图像分辨率，帮助医生更准确地诊断疾病，发现细微病变，更好地理解和预测疾病，例如生成更多的 CT 扫描图像或 X 光片，辅助诊断和治疗计划。

② 手术模拟与训练：生成详细的虚拟患者模型（三维重建图像），供外科医生练习复杂的手术操作，降低实际手术风险。

③ 康复训练：创建个性化的康复方案，帮助患者在家完成专业的物理治疗课程。

④ 疾病检测：生成更多样化的病变图像，帮助医生识别早期症状或难以察觉的微小变化，提高诊断的准确性。

⑤ 病理分析：通过合成不同阶段的病理切片图像深入理解疾病的发展过程，指导个性化治疗策略。

⑥ 放射治疗计划：优化放射治疗剂量分布图，确保肿瘤区域得到充分照射的同时最大限度地保护周围正常组织。

⑦ 数据增强：为训练机器学习模型提供更多样化的数据集，尤其是针对罕见病或特殊病例，提高模型的鲁棒性和泛化能力。

⑧ 减少辐射暴露：利用合成图像代替真实的 CT 扫描图像或 X 光片，减少患者接受的辐射剂量，特别是儿童和孕妇等敏感群体。

（5）自动驾驶。

① 场景模拟：模拟各种驾驶场景，测试和改进车辆的安全性能，包括天气变化、交通状

况等因素。

② 数据增强：为训练自动驾驶算法提供多样化的数据集，提高系统的鲁棒性和泛化能力。

（6）时尚与零售。

① 虚拟试衣：通过生成用户穿不同服装的效果图提供在线购物参考，改善用户体验。

② 产品展示：快速生成高质量的产品设计草图或渲染图，加速研发流程，无需实物拍摄即可呈现多种视角和细节，用于电商平台的商品展示。

③ 消费电子：应用于智能手机、平板计算机等设备，即使在不理想的拍摄条件下也能获得高质量的照片和视频。

（7）建筑与房地产。

① 建筑设计可视化：生成建筑外观和内部空间的效果图，帮助客户直观地理解设计方案，促进销售。

② 房产营销：生成虚拟的室内装饰方案，让潜在买家提前体验未来的居住环境。

（8）教育与培训。

① 互动学习材料：生成生动的教学资源，如教学插图、科学实验动画、历史场景重现等，使学习更加有趣和有效。

② 职业技能训练：模拟真实工作环境，如工厂生产线、医院急诊室等，培养学生的专业技能。

（9）科学研究。

① 数据可视化：生成复杂数据的图形表示，帮助研究人员更清晰地理解实验结果，发现新的模式或趋势。

② 模拟实验：在无法直接观察或实验的情况下，通过生成图像推测可能的结果，如天文学。

③ 卫星遥感：改善卫星拍摄的照片质量，用于地理测绘、环境监测等领域，提供更高精度的数据支持。

（10）安全与监控。

① 异常检测：结合图像生成技术与监控系统实时生成正常情况下的预期图像，对比实际画面，以识别异常行为或事件。

② 隐私保护：在不泄露个人身份信息的前提下生成模糊处理后的监控图像，既保持了监控的有效性，又保护了隐私。

③ 监控系统：从低分辨率监控摄像头捕获的画面中提取更多有用信息，辅助安防工作，如人脸识别、车牌识别等。

5.2　图像风格迁移

图像风格迁移是图像生成技术中的一个重要应用，它将一张图片的艺术风格应用到另一张图片的内容上，从而创造出新视觉效果的图像（图 5-2）。这项技术结合了内容图像的结构信息和风格图像的纹理、颜色及笔触特征，广泛应用于艺术创作、设计和个人化内容生成等领域。

图 5-2　图像风格迁移示例

5.2.1　基本原理

图像风格迁移是通过深度学习算法将一张图像的内容与另一张图像的风格相结合，生成一张既保留原图内容又体现指定艺术风格的新图像。具体来说，它通常利用 CNN 提取内容图像的高级特征和风格图像的纹理特征，并在这两个域之间找到平衡，以创造视觉上和谐的结果。

图像风格迁移的一些基本概念如下。

（1）内容图像：提供主要形状和对象布局的基础图像。

（2）风格图像：提供视觉风格（如色彩、纹理、笔触等）的参考图像。

（3）目标图像：最终生成的图像，保留了内容图像的主要结构，但采用风格图像美学特征。

主要的实现方法如下。

（1）CNN：最常用的实现方式是基于预训练的深度 CNN，如 VGGNet。通过分析不同层的激活值来捕捉图像的内容和风格特征。

（2）内容损失：衡量生成图像与内容图像在高级特征表示上的差异，确保两者的结构相似性。

（3）风格损失：衡量生成图像与风格图像在低级特征（如纹理、颜色分布）上的统计相似性，通常通过对特征图进行 Gram 矩阵计算来实现。

其优化过程主要是：初始时，使用内容图像作为起点或随机噪声图像。使用梯度下降法最小化内容损失和风格损失之和，逐步调整像素值，以接近理想的效果。

5.2.2　代表性算法

图像风格迁移的代表性算法是由加蒂等提出的基于 CNN 的“神经风格迁移”，它通过优化生成图像的内容和风格损失函数来融合内容图像与风格图像的特征。

（1）加蒂等2016年提出了经典的风格迁移方法，首次展示了如何利用深度学习模型有效地分离并重组图像的内容和风格特征，在学术界和工业界产生了深远影响，并启发了许多后续研究。

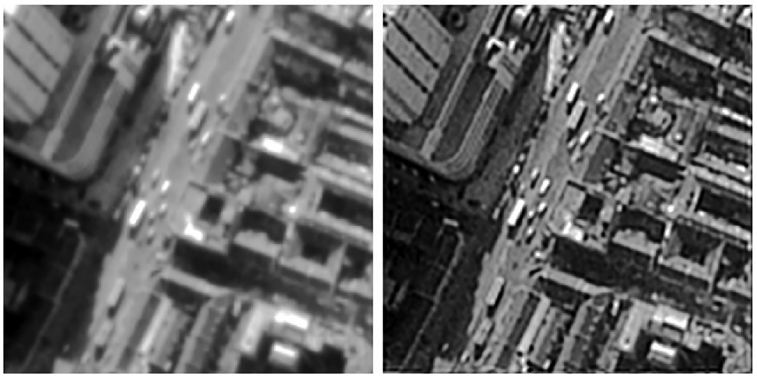
（2）快速样式传输：为了加速传统风格迁移的速度，约翰生等提出了快速风格迁移算法，该方法训练一个前馈网络，直接从输入图像生成带有所需风格的结果，大大提高了处理效率。它适用于实时应用场景，如移动设备上的滤镜应用。

（3）AdaIN：该方法简化了风格迁移的过程，仅需对单个风格图像进行适应性归一化操作即可实现高质量的风格转换。特别适合多风格切换任务，因为可以在推理阶段轻松更换不同的风格参数。

（4）CycleGAN和其他无监督方法：当没有成对的训练数据时，CycleGAN及其变体可以通过对抗训练机制学习两个域之间的映射关系，实现跨域风格迁移。该算法应用范围广泛，包括照片编辑、视频处理以及医学影像分析等。

5.3　超分辨率重建

超分辨率（super resolution，SR）重建是图像生成技术中的一个重要领域，旨在从低分辨率（low resolution，LR）图像中恢复出高分辨率（high resolution，HR）图像（图5-3）。通过增强图像的细节和清晰度，这项技术广泛应用于医疗影像、卫星遥感、视频处理、监控系统以及消费级电子产品等多个领域。



(a) 原分辨率为1m的测试影像　　(b) 超分辨率处理的结果

图5-3　超分辨率重建示例

5.3.1　基本原理

SR重建是通过算法融合多帧低分辨率图像，或利用深度学习模型推测并生成更高分辨率的图像，恢复细节信息，提升图像的清晰度和质量。

一些主要概念如下。

（1）LR图像：即原始输入图像，具有较低的空间分辨率，通常表现为模糊或细节丢失。

（2）HR图像：即期望输出的图像，具有更高的空间分辨率，能够展现更多的细节和更清晰的视觉效果。

SR 重建的主要挑战如下。

（1）信息缺失：从数学角度来看，这是一个不适定问题，因为 LR 图像丢失了高频成分，直接放大无法恢复这些信息。

（2）多解性：同一张 LR 图像可能对应多个不同的 HR 版本，选择最合理的解决方案是一个关键问题。

5.3.2 基于学习的方法

随着深度学习的发展，基于学习的 SR 重建方法取得了显著进展，主要分为以下几类。

（1）基于 CNN。

① SRCNN（超分辨率卷积神经网络）：最早的深度学习 SR 模型之一，它将 SR 问题建模为端到端学习任务，通过训练一个三层卷积网络来直接映射 LR 图像到 HR 图像。

② VDSR（非常深的超分辨率）：扩展了 SRCNN 的思想，使用更深的网络结构，以捕捉更复杂的特征表示，进一步提高了重建质量。

③ EDSR（增强的深度超分辨率）：移除了不必要的模块（如批归一化层），并增加了网络深度和宽度，实现了更好的性能。

（2）基于递归网络。

① DRCN（深度递归卷积网络）：引入递归机制，允许网络重复利用先前层的信息，从而增强了对上下文的理解能力。

② DRRN（密集连接的残差网络）：结合了密集连接和残差学习的概念，使得网络能够更好地传播梯度，避免了深层网络中的梯度消失问题。

（3）基于 GAN。

① SRGAN（超分辨率生成对抗网络）：首次将 GAN 应用于超分辨率重建，通过对抗训练使生成器不仅关注像素级别的准确性，还注重图像的感知质量（如纹理、颜色等）。判别器负责区分真实 HR 图像与生成的假 HR 图像，迫使生成器产生更加逼真的结果。

② ESRGAN（增强的超分辨率生成对抗网络）：改进了 SRGAN，采用了新的损失函数（如相对感知损失）和优化策略，进一步提升了生成图像的真实感和细节表现。

（4）基于注意力机制。

① RCAN（残差通道注意力网络）：引入通道注意力机制，自动调整不同特征通道的重要性，使得网络能够聚焦于最具代表性的部分，提高重建效果。

② SAN（空间注意力网络）：利用空间注意力机制捕捉图像中的局部依赖关系，增强对复杂结构的理解。

5.4 视频生成

视频生成是一个复杂且多样的领域，它结合了计算机视觉、机器学习和深度学习等技术，旨在从静态图像或少量帧中生成连贯的视频序列（图 5-4）。视频生成不仅在娱乐产业中有广泛应用，还为医疗、安全监控、自动驾驶等多个行业提供了新的工具和支持。

图 5-4　视频生成示例

5.4.1　基本原理

视频生成是通过深度学习模型，尤其是 GAN 或 VAE，从 LR 视频帧中预测并生成 HR 的对应帧，从而提升整个视频的分辨率和细节清晰度。

一些重要概念如下。

(1) 视频生成：从给定的数据(如单张图像、文本描述、关键帧等)创建连续的视频帧序列，这些帧之间具有时间上的连贯性和空间上的合理性。

(2) 时空一致性：确保生成的每一帧都与前后帧保持逻辑上的联系，形成自然流畅的动作或场景变化。

5.4.2　主要方法

视频生成的主要方法是利用深度学习模型，如 GAN、VAE 和基于 Transformer 的架构，从静态图像或 LR 视频中预测并合成连续、连贯的 HR 视频帧序列。

(1) 基于 CNN。

① VideoGAN：扩展了图像 GAN，通过引入 3D 卷积层来处理视频数据的空间和时间维度，生成逼真的视频片段。

② MoCoGAN(动作和内容 GAN)：将视频分解为动作和内容两个部分，分别用不同的子网络建模，从而更好地控制生成视频的风格和动态特性。

(2) 基于 RNN。

① LSTM/GRU：利用 LSTM 或 GRU 来捕捉视频帧之间的长期依赖关系，适用于较短序列的预测任务。

② ConvLSTM：结合卷积操作和 LSTM 的优点，专门用于处理具有时空结构的数据，如视频。

(3) 基于 Transformer 架构。

① ViViT(视频视觉变换器)：基于 Transformer 的架构，可以同时处理视频的空间和时间特征，展现出色的泛化能力和表达力。

② TimeSformer：进一步优化了 Transformer 在视频理解中的应用，特别强调了对时间信息的有效编码。

(4) 基于流模型。使用流模型来学习视频帧的概率分布，允许精确地估计数据点的确

切对数似然，并支持高效的采样过程。

（5）基于扩散模型。DDPM（去噪扩散概率模型），通过逐步向图像添加噪声，然后学习逆转这一过程，以恢复原始图像，这种方法也可以应用于视频生成，提供高质量的结果。

5.4.3 代表性算法

视频生成的代表性算法举例如下。

（1）MoCoGAN：提出了一种分离运动和内容的方法，使得用户可以通过调整输入参数来定制生成视频的特定方面。

（2）Vid2Vid：由 NVIDIA 开发，可以从给定的关键帧或草图自动生成完整的视频，广泛应用于影视特效、游戏开发等领域。

（3）Text-to-Video：结合了文本到图像生成技术和视频合成，可以根据自然语言描述直接生成相应的视频内容，如故事叙述或虚拟旅行体验。

（4）CIPS-3D：一种最新的三维感知生成对抗网络，能够在保持物体形状一致性的前提下生成不同视角下的视频。

（5）Runway Gen-2 ：支持基于文本到视频的生成，可创建具有特定主题或风格的短片。

（6）Reelskit：专为短视频创作者设计，利用 AI 快速生成吸引人的视频内容。

5.5 医疗影像合成

图像生成技术应用在医疗影像合成中快速发展，通过人工智能和深度学习方法来创建或增强医学图像，以辅助诊断、治疗规划和研究（图 5-5）。

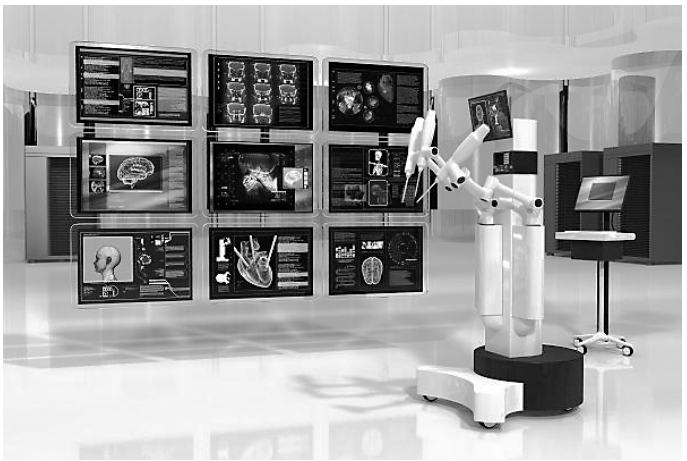


图 5-5 医疗影像合成

5.5.1 基本原理

医疗影像合成的是利用深度学习模型（如 GAN、VAE），从现有医学图像数据中学习特征，进而生成新的或增强的医学影像，以辅助诊断、治疗规划或研究。