图像与视频生成调研报告

信息科学技术概论作业

姓名：汪苏轶

学号： 2500934024

二〇二五 年 十二 月

目录

1 引言

本文调研了图像与视频合成领域的前沿研究方向与进展。

图像与视频生成作为人工智能领域的重要分支，近年来取得了令人瞩目的进展。从初期的生成对抗网络（GANs）到如今的扩散模型（Diffusion Models）和自回归模型（Autoregressive Models），生成式 AI 技术已经从纯学术研究走向广泛的商业应用。特别是在多模态大模型的推动下，图像与视频生成技术不仅在视觉质量上大幅提升，在生成连贯性、时序一致性和物理规律模拟等方面也取得了显著突破。

在工业界，图像与视频生成技术已广泛应用于影视制作、游戏开发、广告营销和在线教育等领域，大大提升了内容创作的效率和质量。与此同时，学术界的研究重点也从单纯提升生成质量，转向解决长视频生成、物理规律模拟、细粒度控制等更具挑战性的问题。本调研报告将全面梳理当前图像与视频生成领域的技术路线、工业界应用现状、学术界前沿研究成果等。

2 图像与视频生成的核心技术方法

本节将介绍目前图像与视频生成领域的几种主流技术路线，每种都有其独特的优势和适用场景。这些方法在技术演进上存在着继承与发展的关系。

2.1 生成对抗网络（GANs）

生成对抗网络（Generative Adversarial Networks）[1]是推动图像与视频生成领域早期发展的奠基性技术。其核心在于同时训练两个“互为对手”的神经网络模型：一个生成器（Generator）和一个判别器（Discriminator），二者在博弈论中构成一个极小化极大博弈。

2.1.1 核心架构与数学原理

生成对抗网络由生成器和判别器两个主要组件组成。

通常来说，生成器与判别器都是深度神经网络。生成器的目标是学习将先验噪声变量 z（通常取自均匀分布或高斯分布）映射到数据空间，生成与真实训练数据 x 在统计上无法区分的样本 $G(z)$。判别器则是一个二分类器，其作用是估计一个给定样本来自真实数据分布 p_{data} 而不是生成器分布 p_g 的概率。对于一个输入的 x，判别器会输出一个标量 $D(x)$，表示 x为真实数据的置信度。

GAN 的训练过程可以被形式化为一个关于价值函数 $V(G,D)$的优化问题：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

对于判别器，$\max_D V(D,G)$ 意味着判别器希望最大化该价值函数：

- $\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]$：判别器需要为真实数据 x 输出高概率，即让 $D(x)$ 经可能接近 1，从而最大化这一项的值。

- $\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$：判别器需要为生成器生成的数据 $G(z)$ 输出低概率，从而最大化这一项的值。

总的来说，判别器即希望尽可能区分真实数据和生成器生成的数据。

对于生成器，$\min_G V(D,G)$ 代表生成器希望最小化这个被判别器最大化的价值函数，生成器只会影响第二项，即希望最大化 $D\big(G(z)\big)$ 从而最小化 $\log(1-D(G(z)))$。换句话说，生成器的目标是让判别器对其生成的样本 $G(z)$ 做出误判，认为它们是真实的。

2.1.2 主要挑战

GAN 的训练本质是一个复杂的动态系统，其目标并非寻找单一的最小值点，而是寻求生成器与判别器之间的纳什均衡。这种交替优化的过程极其脆弱，容易陷入不稳定状态[2]。

一种主要的问题式动态梯度问题，生成器和判别器的损失函数是高度耦合的。理想的训练状态是二者能力同步增长，达到一种精妙的平衡。然而，在实践中，一方很容易压倒另一方。

如果训练过程中判别器过早地变得过于强大，能够完美区分所有真实和生成样本，那么它反馈给生成器的梯度（即 $\nabla_{\theta_g}\log(1-D(G(z)))$）会趋近于零，即产生梯度消失。这将导致生成器无法从判别器获得有效的学习信号，从而使得训练陷入停滞。

反之，如果生成器进化过快，产生了足以“欺骗”当前判别器的样本，它可能会“走捷径”，专注于生成有限的几种成功模式，而忽略了数据的整体分布，这直接导致了模式崩溃。

判别器和生成器的损失值并不能可靠地指示训练进度或生成样本的质量。一个持续下降的生成器损失可能意味着它正在赢得“博弈”，但也可能意味着它正走向模式崩溃。因此，研究人员常常需要依赖定期可视化生成样本来主观判断训练状态，这凸显了训练过程缺乏鲁棒、可靠的收敛指标。

2.1.3 小结

生成对抗网络提供了一个强大而灵活的框架，用于学习高维、复杂的真实世界数据分布。尽管其训练过程充满挑战，但通过不断的理论创新和架构改进，GAN 已经成为人工智能领域，特别是计算机视觉和生成模型中最重要的范式之一，持续推动着内容生成技术的边界。其核心思想“通过对抗性竞争来驱动学习”的影响已远超其本身，渗透到机器学习的诸多其他领域。

2.2 扩散模型（Diffusion Models）

扩散模型[3]是一类基于似然最大化的生成模型，其核心思想是通过学习一个可逆的马尔可夫链，将简单的高斯分布逐步转化为复杂的数据分布。该框架主要包含两个过程：前向的扩散过程和反向的生成过程。

2.2.1 扩散模型核心原理

扩散模型的核心原理主要分为前向过程和反向过程两部分。

前向过程，也称为扩散过程。这是一个固定的、预先定义好的过程。它通过在原始数据 $\mathbf{x}_0$（来自真实数据分布 $q(\mathbf{x}_0)$）上连续添加高斯噪声，经过 T 步后，将其转化为一个各向同性的高斯白噪声 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$。这个过程是一个马尔可夫链，其每一步的转移概率由下式给出：

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

其中，$\beta_t \in (0,1)$ 是一个预先定义的、随时间步 t 变化的**噪声调度**。$\sqrt{1-\beta_t}$ 是为了保持数据的总体方差稳定。

一个关键的设计是，我们可以通过重参数化技巧，直接从原始数据 $\mathbf{x}_0$ 采样得到任意时间步 t 的噪声数据 $\mathbf{x}_t$：

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

其中：

- $\alpha_t = 1 - \beta_t$；
- $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$；
- $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

这个闭式解极大地简化了训练过程，因为它允许我们随机采样任意时间步 t 并进行损失计算，而无需逐步执行整个前向链。

反向过程则是学习式去噪。生成数据的目标是逆转上述过程：从一个纯噪声 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 开始，逐步去噪，最终得到来自真实数据分布 $\mathbf{x}_0$。如果我们知道真实的反向转移概率 $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$，我们就可以完成这一逆转。

然而，$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 依赖于整个数据分布，是难以直接计算的。因此，我们使用一个参数化的神经网络（通常是 U-Net 架构）来学习这个反向过程，定义为 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$。
我们构建一个生成式的马尔可夫链，其起始分布为 $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$，每一步的转移由下式给出：

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$

这里，神经网络 μ_θ 和 Σ_θ 负责预测给定当前噪声图像 $\mathbf{x}_t$ 和时间步 t 时，前一步的均值和方差。

2.2.2 模型的训练与生成流程

扩散模型的训练通过最大化变分下界来实现。经过数学推导，一个关键发现是：最小化 VLB 等价于让模型预测在前向过程中所添加的噪声 ϵ。

因此，在实践中，我们通常使用一个简化的、重加权的均方误差损失：

$$L_{\text{simple}} = \mathbb{E}_{t\sim[1,T], \mathbf{x}_0\sim q(\mathbf{x}_0), \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \|^2]$$

其中：

- t 从 $\{1,2,\ldots,T\}$ 中均匀采样。

- $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ 是加噪后的图像。
- ϵ_θ 是神经网络，其任务是预测出在前向过程中添加到 $\mathbf{x}_0$ 上的噪声 ϵ。

训练流程可概括为：

1. 从训练集中随机抽取一张干净图像 $\mathbf{x}_0$。
2. 随机选择一个时间步 t。
3. 从标准高斯分布中采样噪声 ϵ。
4. 根据闭式解计算加噪后的图像 $\mathbf{x}_t$。
5. 让神经网络 ϵ_θ 根据 $\mathbf{x}_t$ 和 t 预测噪声。
6. 计算预测噪声 $\epsilon_\theta(\mathbf{x}_t, t)$ 与真实噪声 ϵ 之间的均方误差，并通过梯度下降更新网络参数。

训练完成后，我们可以通过以下迭代去噪过程从模型中生成新样本：

1. 从 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ 中采样一个随机噪声 $\mathbf{x}_T$。
2. 从 $t = T$ 开始，逐步迭代至 $t = 1$：
   a. 使用训练好的网络 ϵ_θ 预测噪声：$\hat{\epsilon}_t = \epsilon_\theta(\mathbf{x}_t, t)$。
   b. 根据反向转移概率 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 采样前一时刻的图像 $\mathbf{x}_{t-1}$。这通常涉及以下计算：

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_t\right) + \sigma_t \mathbf{z}$$

   其中 $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，$\sigma_t$ 与方差调度相关的标准差。
3. 最终，$\mathbf{x}_0$ 为生成的图像。

2.2.3 模型优势与挑战

与 GAN 的对抗性训练相比，扩散模型的目标函数是简单的回归损失，训练过程更稳定，不易出现模式崩溃。与 GAN 不同，扩散模型提供了可处理的变分下界，使得我们可以定量地评估模型的似然性能。

扩散型能够产生多样性和保真度都极高的样本，在许多基准测试中超越了 GAN[4]。

但是扩散模型采样速度慢，生成一张图像需要进行 T 次（通常需要数千次）神经网络前向传播[3]，这非常耗时，且大量的迭代步骤导致推理成本高昂。

扩散模型通过一个参数化的去噪过程来学习数据的底层结构，提供了一个强大且理论坚实的生成建模框架。尽管存在采样速度的挑战，但通过 DDIM、LDM 等后续研究，这一问题已得到有效缓解。目前，以 Stable Diffusion 为代表的扩散模型，已成为推动 AIGC 领域发展的核心驱动力，在图像、视频、音频生成等方面展现出巨大的潜力。

2.3 其他模型

除了较为早期的生成对抗网络和目前较为主流的扩散模型两种模型以外，还有很多其他的生成式 AI 模型都具有优秀的表现。本节将简单介绍其他几种模型。

2.3.1 自回归模型（Autoregressive Models）

自回归模型[5]将图像或视频生成视为序列预测问题，通过对数据分布进行因式分解，逐个生成图像块或视频帧。

自回归模型基于链式法则，将生成过程分解为一系列条件概率的乘积，通常使用 Transformer 架构来建模这些依赖关系。在视频生成中，这种方法能够自然处理因果依赖，适合流式生成场景。

早期自回归模型如 VideoGPT 使用 3D VQ-VAE 将视频压缩为离散表示，然后使用 Transformer 建模这些表示的分布。最新研究如 InfinityStar 提出了时空金字塔建模方法，将视频分解为"首帧（外观信息）"和"后续片段（动态信息）"，通过下一尺度/片段预测统一图像和视频生成任务。这种方法能够"将静态外观和动态信息解耦"，有效提升了生成效率和质量。

2.3.2 归一化流（Normalizing Flows）

归一化流[6]是另一种基于似然的生成模型，通过一系列可逆变换将简单分布转换为复杂数据分布。与扩散模型相比，归一化流支持单步生成和精确似然计算，具有显著的效率优势。

STARFlow-V[7] 是归一化流在视频生成领域的代表，STARFlow-V 采用全局-局部架构，在时空潜在空间中操作，将因果依赖限制在全局潜在空间中，同时保留丰富的帧内局部交互。这种设计缓解了时间上的错误累积问题，这是标准自回归扩散模型生成的常见缺陷。

STARFlow-V 还引入了流-得分匹配，使模型配备了一个轻量级因果去噪器，以自回归方式提高视频生成的一致性。同时，它采用视频感知 Jacobi 迭代方案，将内部更新重新构造为可并行迭代而不破坏因果关系。得益于其可逆结构，同一模型原生支持文本到视频、图像到视频以及视频到视频生成任务。

3 图像与视频生成目前在工业界的应用

工业界的图像与视频生成模型注重实用性、效率和商业化价值，已形成多元化的竞争格局。

3.1 工业界主流图像与视频生成模型

- OpenAI Sora：作为视频生成领域的风向标，Sora 基于 DiT 架构，能够生成长达 1 分钟的高质量、高一致性视频。Sora 不仅能够生成复杂场景，还展现了对物理规则的动态模拟能力，代表了当前视频生成的最高水平。
- Runway Gen-2：作为较早涉足视频生成的工具，Runway Gen-2 支持文字转视频、图像转视频和视频风格化等多种能力，被广泛应用于短视频创意平台。
- Pika Labs：专注于风格多样性与动作丰富性，特别适合制作动漫、卡通、科幻风格内容，深受二次创作用户欢迎。

- 生数科技 Vidu Q1：作为业内首个"高可控视频大模型"，Vidu Q1 能够接受空间布局信息作为输入，显著提升了视频生成的可控性。该模型在多主体细节可控、音效同步可控和画质增强方面均取得成效，推动 AI 视频生成走向"高可控"时代。
- 腾讯混元视频与字节即梦：腾讯的混元视频大模型整合了自研扩散技术与语义建模，而字节跳动的"AIGC 视频平台即梦"则专注于商业化落地。字节跳动还推出了 InfinityStar 自回归模型，是"首个在 VBench 上超越扩散模型的离散自回归视频生成器"，在单张 GPU 上生成 5 秒 720p 视频仅需不到 1 分钟，比同尺寸 DiT 方法"快了一个数量级"。
- 智象未来 HiDream-I1：该模型采用 Sparse DiT 架构和对抗蒸馏技术，在生成效果和运行速度之间找到了平衡，成为"首个登顶 Artificial Analysis 图像竞技场榜单的中国自研生成式 AI 模型"。
- 快手可灵：从 2023 年推出至今已迭代至可灵 2.1，在画质质量、动态质量、美学表现、运动合理性以及语义理解等方面不断提升。可灵团队还与香港城市大学合作提出了 VANS 模型，开创了"视频作为答案"的新任务范式，使 AI 能够直接生成视频而不仅仅是文字来回答问题。

模型名称	开发公司	核心技术	主要特点	应用场景
Sora[8]	OpenAI	DiT(扩散 Transformer)	1 分钟生成、物理规律模拟	创意内容、影视制作
Vidu Q1	生数科技	扩散模型+可控生成	高可控性、空间布局控制	广告、专业视频制作
InfinityStar	字节跳动	自回归+时空金字塔	高效生成、单 GPU 快速推理	社交平台、UGC 内容
混元视频	腾讯	扩散模型+语义建模	多任务支持、3D 生成	游戏、影视、广告
可灵 2.1	快手	扩散模型+优化算法	平衡质量与效率、美学表现	短视频、社交平台
HiDream-I1	智象未来	Sparse DiT+对抗蒸馏	图像生成领先、提示理解强	创意设计、图像生成

表 3.1.1 工业界主流模型特点对比

3.2 图像与视频生成模型在各领域的应用

图像与视频生成技术在工业界已经广泛应用于多个领域：

- 影视行业：生成式 AI 已经初步在影视行业落地，应用于剧本生成、角色/场景建模、动画生成、后期配音和剪辑调色等环节。例如，通过 AI 生成分镜和预览，大幅降低了前期制作成本。
- 游戏行业：AI 技术赋能游戏角色建模、贴图动画、场景生成和剧情设计，同时通过 NPC 的个性化交互增强玩家体验。生成式 AI 还能动态调整关卡难度和道具属性，创造更丰富的游戏体验。

- 电商营销：商家可以通过输入产品描述文字或静态图片，让 AI 自动生成产品展示视频，显著节省人力成本并提升转化率。
- 教育行业：结合 PPT 文案和音频内容，AI 可以生成配套教学视频，提升学习体验与效率。VANS 模型更进一步，能够根据用户当前进度生成定制化的程序性教学视频，例如展示具体的烹饪步骤或领带打法。

4 总结

图像与视频生成技术正以前所未有的速度发展，从初期的简单视觉合成逐步走向能够模拟物理规律、理解复杂指令的高阶生成。在技术路线上，扩散模型当前占据主导地位，但自回归模型和归一化流等替代方法因其在效率和长序列生成方面的潜力而受到持续关注。工业界已形成多元化的模型生态，专注于不同应用场景和商业化路径；学术界则致力于探索基础架构创新、因果生成、语义规划等前沿方向。

未来，图像与视频生成领域将朝着更高质量、更长时长、更强可控性和更高效率的方向发展，同时需要解决物理规律模拟、伦理安全和内容认证等挑战。随着技术的不断成熟，图像与视频生成有望成为未来人机交互的核心媒介，重塑内容创作和知识传递的方式。

参考文献

[1] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[2] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[4] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.

[5] Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks[C]//International conference on machine learning. PMLR, 2016: 1747-1756.

[6] Dinh L, Krueger D, Bengio Y. Nice: Non-linear independent components estimation[J]. arXiv preprint arXiv:1410.8516, 2014.

[7] Gu J, Shen Y, Chen T, et al. STARFlow-V: End-to-End Video Generative Modeling with Normalizing Flow[J]. arXiv preprint arXiv:2511.20462, 2025.

[8] Brooks T, Peebles B, Holmes C, et al. Video generation models as world simulators[J]. OpenAI Blog, 2024, 1(8): 1.