arXiv:2303.11717v1 [cs.AI] 21 Mar 2023

生成式人工智能（AIGC）综述：从 GPT-4 到 GPT-5，ChatGPT 是否就是你所需要的一切?

Chaoning Zhang，韩国庆熙大学Chenshuang Zhang，韩国KAISTSHENG ZHENG，中国北京理工大学Yu Qiao，韩国庆熙大学CHENGHAO LI，韩国KAISTMENGCHUN ZHANG，韩国KAISTSUMIT KUMAR DAM，韩国庆熙大学CHU MYAET THWAL，韩国庆熙大学YE LIN TUN，韩国庆熙大学LE LUANG HUY，韩国庆熙大学DONGUK KIM，韩国庆熙大学SUNG-HO BAE，韩国庆熙大学LIK-HANG LEE，中国香港香港理工大学YANG YANG，中国电子科技大学HENG TAO SHEN，中国电子科技大学In So Kweon，韩国KAISTCHOONG SEON HONG，韩国庆熙大学

随着 ChatGPT 迅速走红，生成式人工智能（人工智能生成内容，简称 AIGC）凭借分析和创作文本、图像等内容的能力，成为各大媒体争相报道的焦点。在媒体铺天盖地的报道中，我们几乎不可能错过从某个角度一窥 AIGC 的机会。在人工智能从纯粹分析转向内容创作的时代，值得注意的是，搭载最新语言模型 GPT-4 的 ChatGPT，只是众多 AIGC 任务中的一种工具。ChatGPT 的能力令人印象深刻，许多人不禁想知道它的局限：GPT-5（或其他未来的 GPT 变体）能否帮助 ChatGPT 统一完成所有 AIGC 任务，

作者地址：Chaoning Zhang，韩国庆熙大学，chaoningzhang1990@gmail.com；Chenshuang Zhang，韩国 KAIST，zcs15@kaist.ac.kr；SHENG ZHENG，中国北京理工大学，zszhx2021@gmail.com；Yu Qiao，韩国庆熙大学，qiaoyu@ khu.ac.kr；CHENGHAO LI，韩国 KAIST，lch17692405449@gmail.com；MENGCHUN ZHANG，韩国 KAIST，zhangmengchun527@gmail.com；SUMIT KUMAR DAM，韩国庆熙大学，skd160205@khu.ac.kr；CHU MYAET THWAL，韩国庆熙大学，chumyaet@khu.ac.kr；YE LIN TUN，韩国庆熙大学，yelintun@khu.ac.kr；LE LUANG HUY，韩国庆熙大学，quanghuy69@khu.ac.kr；DONGUK KIM，韩国庆熙大学，g9896@khu.ac.kr；SUNG-HO BAE，韩国庆熙大学，shbae@khu.ac.kr；LIK-HANG LEE，中国香港香港理工大学，iskweon77@kaist.ac.kr；YANG YANG，中国电子科技大学，dlyyang@gmail.com；HENG TAO SHEN，中国电子科技大学，shenhengtao@hotmail.com；In So Kweon，韩国 KAIST，iskweon77@kaist.ac.kr；CHOONG SEON HONG，韩国庆熙大学，cshong@khu.ac.kr。

允许为个人或课堂使用而免费制作本作品的全部或部分内容的电子版或纸质副本，但前提是不得为营利或商业利益制作或分发副本，且副本必须附有本声明及首页的完整引文。对于本作品中由美国计算机协会以外的其他方所有的组成部分，必须遵守其版权规定。注明出处后，允许摘录。若要以其他方式复制、再版、发布至服务器或向列表重新分发，则须事先获得特定许可和/或支付费用。请通过 permissions@acm.org 申请许可。© 2022年 美国计算机协会。稿件已提交至美国计算机协会

如何实现多样化的内容创作？为回答这一问题，我们有必要全面梳理现有的人工智能生成内容任务。因此，我们的工作及时填补了这一空白，首次全面介绍人工智能生成内容，涵盖其技术与应用。现代生成式人工智能依托多种技术基础，包括从模型架构和自监督预训练到生成建模方法（如生成对抗网络和扩散模型）。在介绍基础技术之后，本文将根据输出类型，聚焦各类人工智能生成内容任务的技术发展，包括文本、图像、视频、三维内容等，展现 ChatGPT 未来的全部潜力。此外，我们还总结了人工智能生成内容在教育和创意内容等主流行业中的重要应用。最后，我们讨论当前面临的挑战，并展望生成式人工智能在不久的将来可能的发展方向。

CCS 概念：• **计算方法** → **计算机视觉任务**; 自然语言生成; 机器学习方法。

其他 关键词和短语：综述、生成式人工智能、人工智能生成内容、ChatGPT、GPT-4、GPT-5、文本生成、图像生成

美国计算机协会参考文献格式：

Chaoning Zhang、Chenshuang Zhang、Sheng Zheng、Yu Qiao、Chenghao Li、Mengchun Zhang、Sumit Kumar Dam、Chu Myaet Thwal、Ye Lin Tun、Le Luang Huy、Donguk kim、Sung-Ho Bae、Lik-Hang Lee、Yang Yang、Heng Tao Shen、In So Kweon 和 Choong Seon Hong。2022年。《生成式人工智能（人工智能生成内容）全面综述：从 GPT-4 到 GPT-5，ChatGPT 是否就是你所需要的一切？》1，1（2022年3月），共 56 页。https://doi.org/XXXXXXX.XXXXXXX

目录

摘要 1 目录 2 1 引言 3 2 概述 5 2.1 搜索兴趣所反映的受欢迎程度 5 2.2 它为何如此受欢迎？ 6 2.2.1 内容需求 6 2.2.2 技术条件 7 3 人工智能生成内容背后的基础技术 9 3.1 人工智能中的通用技术 9 3.1.1 骨干架构 9 3.1.2 自监督预训练 12 3.2 人工智能中的生成技术 13 3.2.1 基于似然的模型 14 3.2.2 基于能量的模型 15 3.2.3 两种明星模型：从生成对抗网络到扩散模型 15 4 人工智能生成内容任务：文本生成 17 4.1 文本到文本 18 4.1.1 聊天机器人 18 4.1.2 机器翻译 19 4.2 多模态文本生成 20 4.2.1 图像到文本 20 4.2.2 语音转文本 21 5 人工智能生成内容任务：图像生成 22 稿件已提交至美国计算机协会

5.1 图像到图像 22 5.1.1 图像复原 22 5.1.2 图像编辑 23 5.2 多模态图像生成 25 5.2.1 文本生成图像 25 5.2.2 会说话的人脸 25 6 人工智能生成内容任务：超越文本与图像 27 6.1 视频 27 6.2 三维生成 28 6.3 语音 28 6.4 图 29 6.5 其他 29 7 行业应用 30 7.1 教育 30 7.2 游戏与元宇宙 31 7.3 媒体 31 7.4 广告 32 7.5 电影 33 7.6 音乐 34 7.7 绘画 34 7.8 代码开发 35 7.9 手机应用与功能 35 7.10 其他领域 36 8 挑战与展望 36 8.1 挑战 36 8.2 展望 37 参考文献 38

1 引言

生成式人工智能（人工智能生成内容，简称 AIGC）凭借 ChatGPT 或 DALL- E [343], 等引人入胜的工具成为新闻焦点，预示着一个人工智能新时代即将到来。在媒体铺天盖地的报道下，公众有了许多机会得以一窥 AIGC。然而，媒体报道的内容往往带有偏见，有时甚至具有误导性。此外，ChatGPT 强大的能力令人印象深刻，许多人也开始探究它的局限。就在最近，OpenAI发布了 GPT-4 [307]，其演奏表现较前代变体 GPT-3 有显著提升，还具备多模态生成能力，例如理解图像。GPT-4 在 AIGC 的支持下展现出强大能力，令人印象深刻，也让许多人不禁探问其局限：GPT-5（或其他 GPT变体）能否帮助下一代 ChatGPT 统一所有 AIGC任务? 因此，全面回顾生成式人工智能，有助于为人工智能驱动的内容创作这一不可避免的趋势奠定基础。更重要的是，我们的工作及时填补了这一空白。

传统人工智能的目标主要是执行分类 [263] 或回归 [227]。这种判别式方法使其主要用于分析现有数据。因此，传统人工智能也常被称为分析型人工智能。相比之下，生成式人工智能的特点是能够创建新内容。然而，生成式人工智能通常也需要模型在生成新内容之前，先理解一些现有数据（例如文本指令） [40, 342]。从这一角度看，分析型人工智能可视为现代生成式人工智能的基础，而二者之间的界限往往模糊不清。需要注意的是，分析型人工智能任务也会生成内容。例如，图像分类会生成标签内容 [216]。尽管如此，图像识别通常不被视为生成式人工智能，因为标签内容的维度较低。生成式人工智能的典型任务涉及生成高维数据，如文本或图像。这类生成内容也可用作合成数据，以缓解深度学习对更多数据的需求 [144]。第2节将概述生成式人工智能的流行程度及其背后的原因。

如上所述，生成式人工智能与传统人工智能的区别在于其生成的内容。由此可见，生成式人工智能在概念上与人工智能生成内容（简称 AIGC）相似 [304]。在描述基于人工智能的内容生成时，这两个术语通常可以互换使用。为简便起见，本文将内容生成任务称为 AIGC。例如，ChatGPT 是一种用于名为聊天机器人 [43], 的 AIGC 任务的工具；考虑到 AIGC 任务种类繁多，这只是冰山一角。尽管生成式人工智能与 AIGC 高度相似，但两者之间仍存在细微差别。AIGC 关注内容生成任务，而生成式人工智能还涵盖支撑各种 AIGC 任务发展的基础技术。本文将这些底层技术分为两类。第一类是生成建模技术，例如生成对抗网络 [124] 和扩散模型 [156], , 它们与用于内容创作的生成式人工智能直接相关。第二类主要包括人工智能技术，如骨干架构（例如 Transformer [443]）和自监督预训练（例如 BERT [87] 或 MAE [141]）。其中一些技术是在分析型人工智能的背景下开发的。然而，它们对于展现出有竞争力的演奏水平也变得至关重要，尤其是在具有挑战性的 AIGC 任务中。鉴于此，第 3 节总结了这两类底层技术。

在这些基础技术之上，众多 AIGC任务已成为可能，并可根据生成的内容类型进行简单分类。第 4、5 和 6 节总结了各类 AIGC任务的发展情况。具体而言，第 4 节和第 5 节分别聚焦于文本输出和图像输出。对于文本生成，聊天机器人 [43] 和机器翻译 [497] 是两项主要任务。有些文本生成任务也会以其他模态作为输入，我们主要关注图像和语音。对于图像生成，两项主要任务是图像复原和剪辑 [253]。近年来，文本生成图像受到了广泛关注。除上述两种主要输出类型（即文本和图像）外，第 6 节还介绍了其他输出类型，例如视频、三维、语音等。

随着技术的进步，人工智能生成内容在越来越多的任务中表现出令人满意的效果。例如，聊天机器人过去只能回答简单的问题。然而，近期的 ChatGPT 已展现出理解笑话并根据简单指令生成代码的能力。文本生成图像过去被认为是一项具有挑战性的任务；然而，近期的 DALL-E 2 [342]和 Stable Diffusion [357] 已能够生成照片级真实感图像。因此，将人工智能生成内容应用于行业的机会应运而生。第 7 节介绍了人工智能生成内容在各行各业的应用，包括娱乐、数字艺术、媒体/广告、教育等。随着人工智能生成内容在现实世界中的应用不断拓展，伦理问题等诸多挑战也随之出现，并在第 8 节中进行了讨论。除了当前面临的挑战，本文还展望了生成式人工智能未来可能的发展方向。

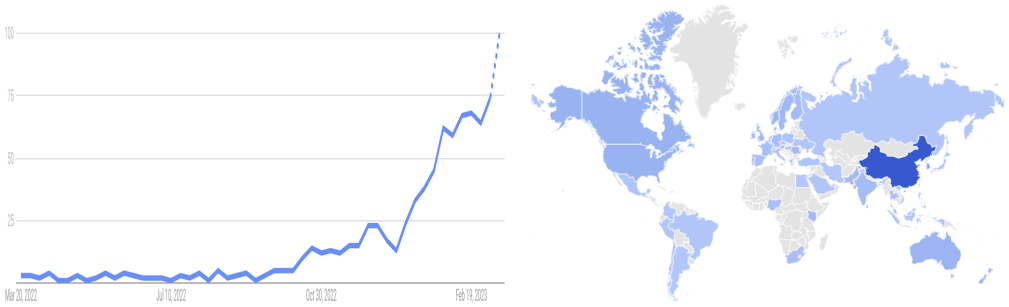


图 1. 生成式人工智能的搜索兴趣：时间趋势（左）和各地区的兴趣度（右）。右侧颜色的深浅表示兴趣排名。

总的来说，本文从生成内容（即 人工智能生成内容任务）的角度对生成式人工智能进行了综述，涵盖其基础技术、各项任务的技术发展、在行业中的应用以及社会影响。图 4 展示了论文结构概览。

2 概述

人工智能在内容创作中的应用历史悠久。1954 年，IBM 在其纽约总部首次公开演示了机器翻译系统。1957 年，第一首计算机生成的音乐以《Illiac Suite》之名问世。这些早期尝试和概念验证的成功让人们对人工智能的未来抱有很高的期望，促使各国政府和企业投入大量资源发展人工智能。然而，如此大规模的投资热潮并未带来预期的成果。此后，人工智能进入了一个被称为人工智能寒冬的时期，其发展及应用受到严重阻碍。进入 2010 年代，人工智能再次流行起来，尤其是在 AlexNet 于 2012 年成功完成 ImageNet 分类之后。进入 2020 年代，人工智能迈入了一个新时代：它不仅能够理解现有数据，还能够创造新的内容 [40, 342]。本节将聚焦生成式人工智能的流行程度及其流行的原因，对其进行概述。

2.1 搜索热度所反映的人气

衡量“某个词有多受欢迎”的一个良好指标是搜索兴趣。谷歌提供了一款很有前景的工具来可视化搜索频率，名为 Google 趋势。尽管其他搜索引擎也可能提供类似功能，但我们采用 Google 趋势，因为谷歌是世界上使用最广泛的搜索引擎之一。

随时间和地区变化的关注度。图 1（左）展示了生成式人工智能的搜索兴趣，表明过去一年搜索兴趣显著上升，尤其是在 2022年10月之后。进入 2013年，这一搜索兴趣达到新高。术语人工智能生成内容也呈现类似趋势，见图 2（左）。除了随时间变化的搜索兴趣，Google 趋势还提供按地区划分的搜索兴趣。生成式人工智能和人工智能生成内容的搜索热力图分别见图 1（右）和图 2（右）。对于这两个术语，主要热门地区包括亚洲、北美洲和西欧。最值得注意的是，在这两个术语的搜索兴趣排名中，中国均以 100 位居所有国家之首，其次是北美洲，约为 30，西欧约为 20。值得一提的是，一些规模较小但以科技为导向的国家，对生成式人工智能的搜索兴趣也非常高。例如，按国家搜索兴趣排名前三的国家是新加坡（59）、以色列（58）和韩国（43）。

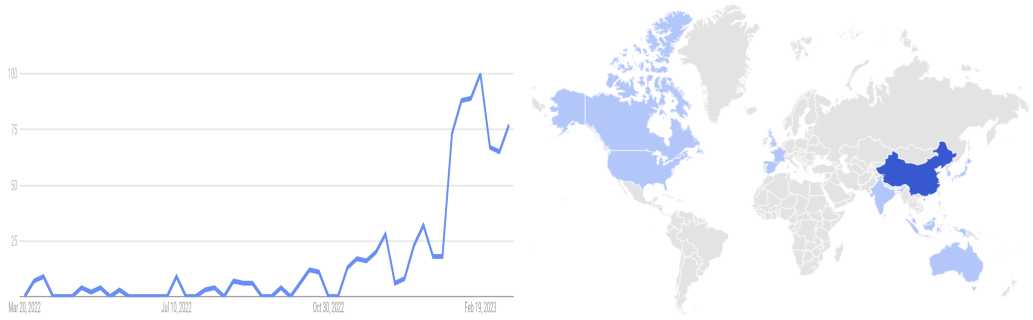


图 2. 人工智能生成内容的搜索兴趣：时间趋势（左）和地区兴趣（右）。右侧颜色深浅表示搜索兴趣排名。

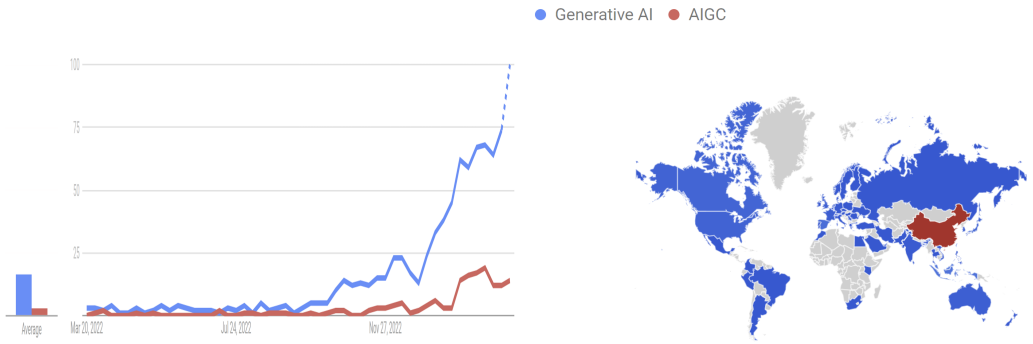


图 3. 生成式人工智能与人工智能生成内容的搜索兴趣对比：时间趋势（左）和地区兴趣（右）。

生成式人工智能 与 人工智能生成内容。 图 3 展示了生成式人工智能与人工智能生成内容的搜索兴趣对比。这里，我们将生成式人工智能与人工智能生成内容的兴趣比例定义为 GAI/AIGC。一个重要发现是，与生成式人工智能相比，中国更倾向于使用术语人工智能生成内容，GAI/AIGC 比例为 15/85。相比之下，美国的 GAI/AIGC 比例为 90/10。在包括俄罗斯和巴西在内的许多国家，这一比例为 100/0。总体而言，大多数国家更偏好生成式人工智能而非人工智能生成内容，因此生成式人工智能的总体搜索兴趣高于人工智能生成内容。中国成为采用术语人工智能生成内容的领先国家，原因尚不完全明确。一种可能的解释是，人工智能生成内容缩写为一个词，因此更容易使用。我们还在 Google 趋势上搜索了生成式人工智能和人工智能生成内容的中文版本，但目前的展示结果并不充分。

2.2 为什么它会变得流行?

过去一年，生成式人工智能的搜索兴趣激增，这主要归因于 Stable Diffusion 或 ChatGPT 等引人入胜的工具的出现。本文将探讨生成式人工智能为何如此流行，重点关注促成这类强大人工智能生成内容工具问世的因素。我们从两个角度总结其原因：内容需求和技术条件。

2.2.1 内容需求。互联网从根本上改变了我们与世界沟通和互动的方式，而数字内容在其中发挥着关键作用。在过去几十年里，网络上的内容也经历了多次重大变化。

在 Web 1.0 时代（1990年代至2004年），互联网主要用于访问信息和分享信息，网站也大多是静态的。用户之间几乎没有互动，主要的沟通方式是单向的：用户访问信息，但不贡献或分享自己的内容。这些内容主要以文本为基础，通常由相关领域的专业人士创作，例如新闻记者撰写新闻文章。因此，这类内容通常称为专业生成内容（PGC），后来又被另一类内容——用户生成内容（UGC） [214, 322, 427]所主导。与 PGC 不同，Web 2.0 [308] 中的 UGC 主要由用户在 Facebook [203], Twitter [257], YouTube [159], 等社交媒体平台上创作。与 PGC 相比，UGC 的数量明显更多，但质量可能较低。

我们目前正从 Web 2.0 迈向 Web 3.0 [363]。Web 3.0 以去中心化、无需中介为显著特征，同时也依赖一种超越专业生成内容（PGC）和用户生成内容（UGC）的新型内容生成方式，以解决产量与质量之间的权衡问题。人工智能被广泛视为解决这一问题的有力工具。例如，过去只有经过长期练习的用户才能绘制出质量尚可的图像。借助文生图工具（如 stable diffusion [357]），任何人都可以通过简单的文本描述创作绘画图像。用户的想象力与人工智能的执行能力相结合，使人们能够以前所未有的速度生成新类型的图像。除了图像生成，人工智能生成内容任务也有助于生成其他类型的内容。

人工智能生成内容带来的另一项变化是，内容消费者与创作者之间的界限变得模糊。在 Web 2.0 中，内容生成器和消费者通常是不同的用户。然而，在 Web 3.0 中，借助人工智能生成内容，数据消费者如今也能够成为数据创作者，因为他们可以利用人工智能算法和技术生成自己的原创内容，并对自己创作和消费的内容拥有更多控制权，从而利用自己的数据和人工智能技术，创作出符合自身特定需求和兴趣的内容。总体而言，向人工智能生成内容的转变有望彻底改变数据的消费与生产方式，让个人和组织能够更灵活地掌控自己创作和消费的内容。接下来，我们将探讨人工智能生成内容如今为何如此流行。

2.2.2 技术条件。谈到人工智能生成内容技术，人们首先想到的往往是机器（深度）学习算法，却忽视了它的两个重要条件：数据访问和计算资源。

数据访问方面的进展。深度学习是指利用数据训练模型的实践。模型演奏表现很大程度上取决于训练数据的规模。通常，训练样本越多，模型演奏表现越好。以图像分类为例，包含超过 100 万张图像的 ImageNet [83] 是一个常用于训练模型和验证演奏表现的数据集。生成式人工智能通常需要更大的数据集，尤其是文本生成图像等具有挑战性的 AIGC任务。例如，DALL-E [343]的训练使用了约 2.5 亿张图像。另一方面，DALL-E 2 [342], 使用了约 6.5 亿张图像。ChatGPT 基于 GPT3 [40]构建，而 GPT3 [40]的训练部分使用了 CommonCrawl 数据集；该数据集在筛选前包含 45TB 的压缩纯文本，筛选后则为 570GB。其他数据集，如 WebText2、Books1/2 和维基百科，也参与了 GPT-3 的训练。能够访问如此庞大的数据集，主要归功于互联网。

计算资源的进步。 推动人工智能生成内容发展的另一个重要因素是计算资源的进步。早期的人工智能算法运行在中央处理器上，无法满足训练大型深度学习模型的需求。例如，AlexNet [216] 是首个在完整 ImageNet 数据集上训练的模型，其训练使用了图形处理器（GPU）。图形处理器最初用于渲染电子游戏中的图形，但如今已广泛应用于深度学习。图形处理器具有高度并行化的特点，执行矩阵运算的速度远快于中央处理器。Nvidia 是图形处理器制造领域的领先企业。其 CUDA 的计算能力不断提升

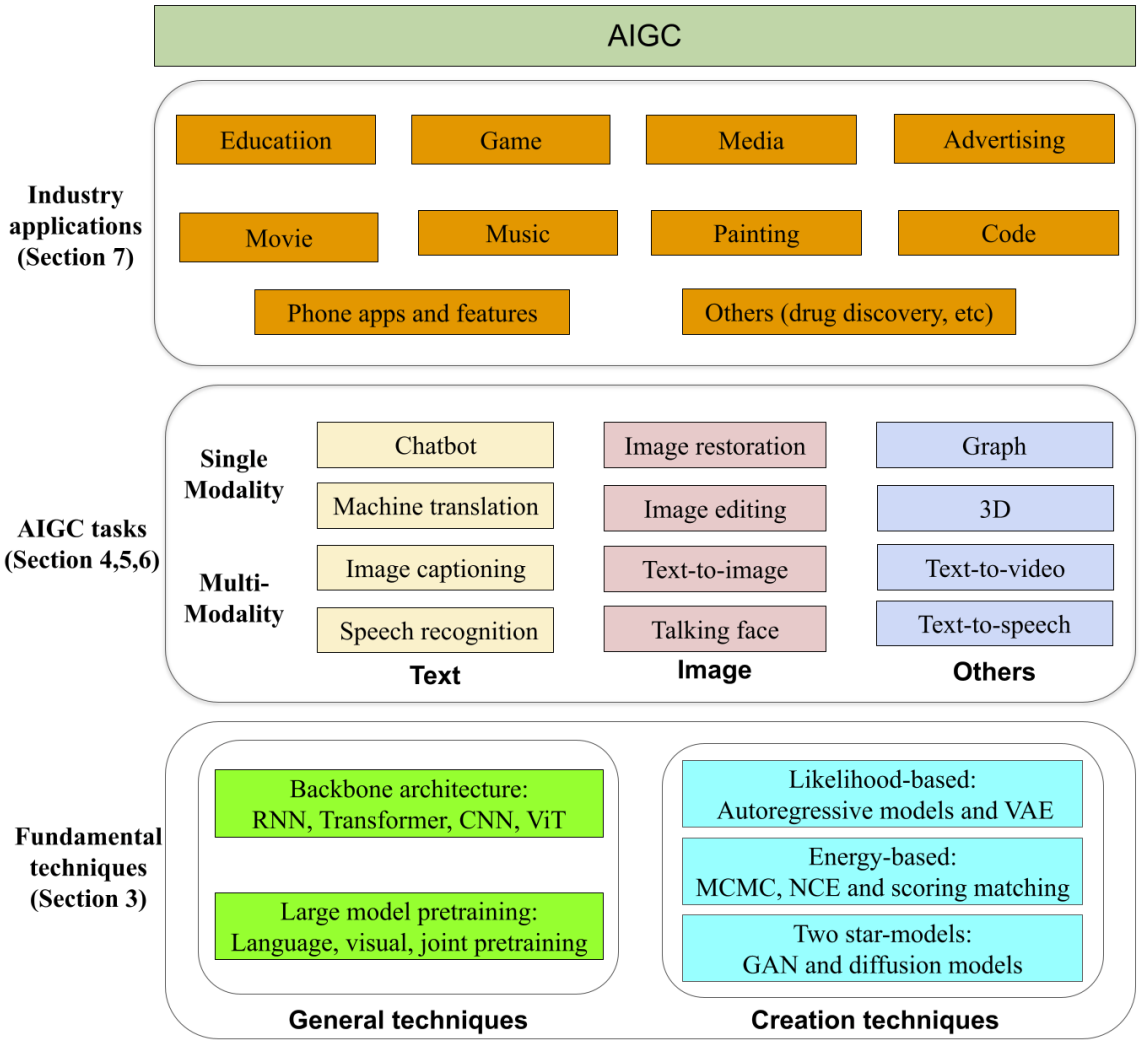


图 4. 生成式人工智能（人工智能生成内容）概览：基础技术、核心人工智能生成内容任务和工业应用。

从2006年的首款支持 CUDA 的图形处理器（GeForce 8800），到近期计算能力提升数百倍的图形处理器（Hopper）。图形处理器的价格从几百美元到几千美元不等，具体取决于核心数量和记忆容量。张量处理单元（TPU）是谷歌专门设计、用于加速神经网络训练的专用处理器。张量处理单元可在 Google Cloud Platform 上使用，价格因使用量和配置而异。总体而言，计算资源的价格正趋于更加亲民。

3 人工智能生成内容背后的基础技术

在本文中，我们将人工智能生成内容视为一组利用人工智能方法生成内容的任务或应用。在介绍人工智能生成内容之前，我们先考察其背后的基础技术；从技术层面来看，这些技术属于生成式人工智能的范畴。在此，我们将基础技术大致分为两类：生成技术和生成技术。具体而言，生成技术是指能够生成各种内容的技术，例如生成对抗网络和扩散模型。与此同时，通用技术无法直接生成内容，但对人工智能生成内容的发展至关重要，例如 Transformer架构。本节将简要概述人工智能生成内容所需的技术。

3.1 人工智能通用技术

AlexNet [216], 取得非凡成功后，深度学习引发了广泛关注，并在某种程度上成为人工智能的代名词。与传统的基于规则的算法不同，深度学习是一种数据驱动方法，通过随机梯度优化模型参数。深度学习能够获得更优特征表示，其成功有赖于更出色的骨干架构和更多数据，这极大地推动了人工智能生成内容的发展。

3.1.1 骨干架构。 自然语言处理（NLP）和计算机视觉（CV）是深度学习的两大主流领域，相关研究显著改进了骨干架构，并推动改进后的骨干网络在其他领域（如语音领域）得到广泛应用。在自然语言处理领域，Transformer [443] 已取代循环神经网络（RNN） [281, 285] ，成为事实上的标准骨干网络。在计算机视觉领域，视觉 Transformer（ViT） [97]也展现出强大能力，与传统的卷积神经网络（CNN）形成对照。下文将简要介绍这些主流骨干网络的工作原理及其代表性变体。

循环神经网络架构。循环神经网络主要用于处理时间序列数据，例如语言或音频。普通循环神经网络有三层：输入层、隐藏层和输出层。循环神经网络中的信息流有两个方向。第一个方向是从输入层到隐藏层，再到输出层。循环神经网络的循环特性体现在第二种信息流中，即沿时间方向的信息流。除相应的输入外，当前隐藏状态还取决于时间t 的隐藏状态，取决于时间 $t-1$的隐藏状态。这种双向信息流设计很好地处理了序列顺序问题，但序列变长时会遭遇梯度爆炸或消失。为缓解长期依赖问题，长短期记忆网络（LSTM） [158] 应运而生，它引入了细胞状态，犹如高速公路，便于信息沿序列方向流动。LSTM 是缓解梯度消失/爆炸问题最常用的方法之一。然而，LSTM 使用三种门控，因此复杂度较高，且需要更多记忆。门控循环单元（GRU） [65]通过合并细胞状态与隐藏状态，并用所谓的更新状态取代遗忘门和输入门，简化了 LSTM。酉循环神经网络 [18] 通过实现酉矩阵来处理梯度问题。门控正交循环单元 [184] 则兼具门控和酉矩阵的优点。双向循环神经网络 [376] 通过在细胞中捕捉过去和未来的信息来改进普通循环神经网络，也就是说，时间 t 的状态是根据时间 $t-1$和 $t+1$计算得出的。根据任务不同，循环神经网络可以采用输入和输出数量各异的多种架构：一对一、多对一、一对多和多对多。多对多架构可用于机器翻译，也称为序列到序列（seq2seq）模型 [413]。注意力机制于 [24] 被引入，使模型的解码器能够查看编码器的每个词元，并根据其重要性自动确定相应权重。

Transformer。 不同于带注意力机制的序列到序列模型 [24, 267, 315], ，一种新型架构摒弃了序列到序列架构，并声称只需注意力即可 [443]。这种注意力称为自注意力，所提出的架构称为 *Transformer* [443] （见图 5）。标准的 Transformer 由一个编码器和一个

解码器组成，并基于残差连接 [143] 和层归一化 [22]构建。除 Add & Norm 模块外，Transformer 还有两个核心组件：多头注意力和前馈神经网络（又称多层感知机）。注意力模块采用多头设计，其中自注意力以缩放点积的形式定义如下：

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}.})V \tag{1}$$

与循环神经网络不同，循环神经网络通过依次输入句子信息来构建位置信息，而 Transformer 通过构建全局依赖关系获得强大的建模能力，但也会丢失位置偏差信息。因此，需要位置编码来使模型感知输入信号的位置信息。位置编码有两种类型。固定位置编码由不同频率的正弦函数和余弦函数表示。可学习位置编码由一组可学习参数组成。Transformer 已成为自然语言处理任务中的事实标准方法。

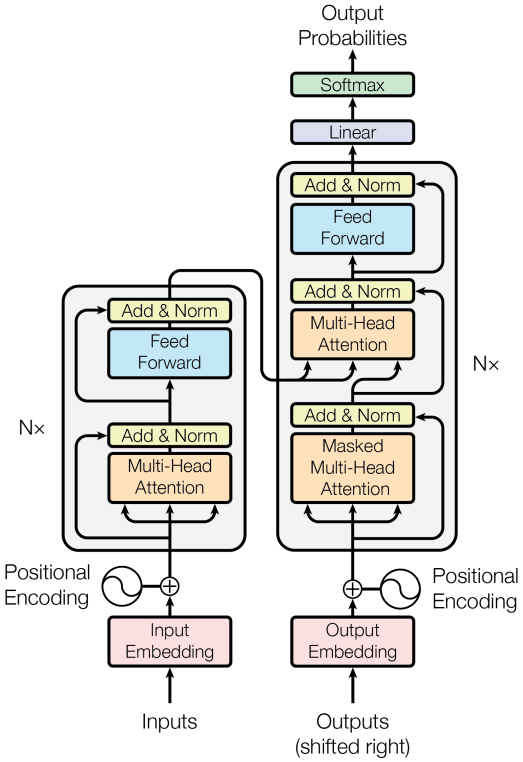


图5. Transformer 结构（图片取自 [443]）。

卷积神经网络架构。在介绍了自然语言处理领域的循环神经网络和 Transformer 后，我们开始了解计算机视觉领域的两种主流骨干网络，即卷积神经网络和视觉Transformer。卷积神经网络已成为计算机视觉领域的标准骨干网络。卷积神经网络的核心在于其卷积层。卷积层中的卷积核（也称滤波器）是一组用于处理图像的共享权重参数，其灵感源自生物视觉皮层细胞。卷积核在图像上滑动，并与图像中的像素值进行相关运算，最终得到。稿件已提交至美国计算机协会

特征图，从而实现图像的特征提取。GoogleNet[417], 凭借其 Inception 模块允许在每个模块中选择多种卷积滤波器尺寸，增加了卷积核的多样性，从而提升了卷积神经网络的性能。ResNet[143] 是卷积神经网络发展中的一个里程碑，它引入了残差连接，使训练更加稳定，并使模型能够通过更深层次的建模取得更好的性能。此后，残差连接成为卷积神经网络中的标准配置。为了拓展 ResNet 的工作，DenseNet[165] 在所有前置层与后续层之间建立密集连接，从而增强了模型的建模能力。EfficientNet[418] 采用一种缩放方法，使用一组固定的缩放系数统一调整卷积神经网络架构的宽度、深度和分辨率，从而提高模型效率。

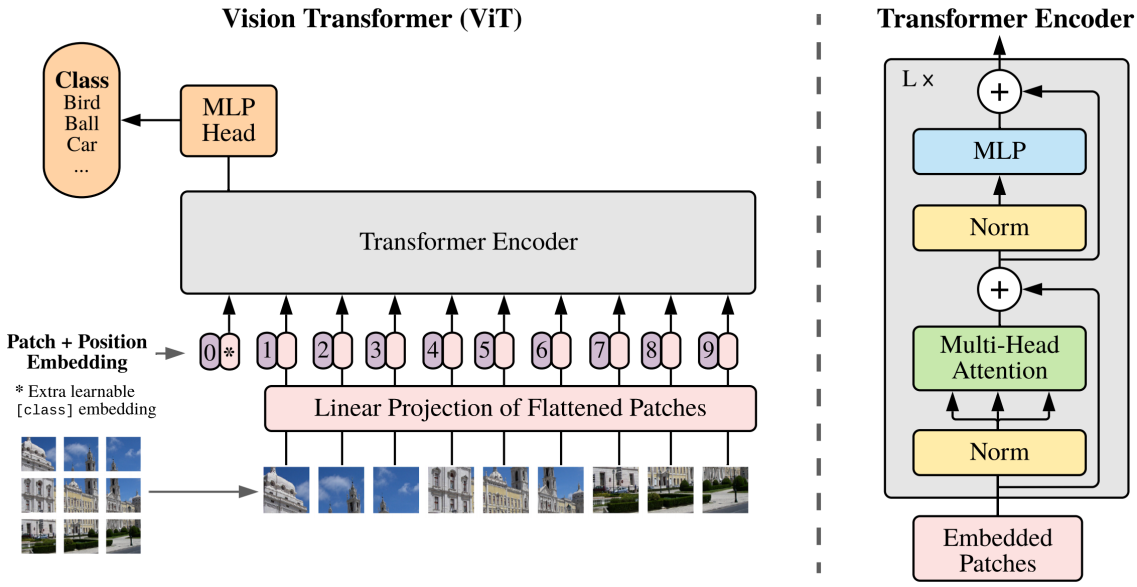


图 6. 视觉Transformer 结构（图片取自 [97]）。

视觉Transformer 架构。受 Transformer 在自然语言处理领域取得成功的启发，许多研究尝试将 Transformer 应用于计算机视觉领域，而 ViT[97] （见图 6）是首个此类模型。ViT 首先将图像展平为一系列二维图像块，并在序列开头插入一个类别标记，以提取分类信息。经过嵌入位置编码后，标记嵌入被输入标准 Transformer。这种简单有效的 ViT 实现使其具有很强的可扩展性。Swin [261] 通过在更深层合并图像块来构建层次化特征图，从而高效处理图像分类和密集识别任务；同时，由于仅在各个局部窗口内计算自注意力，它降低了计算复杂度。DeiT[430]采用教师-学生策略进行训练，并通过引入蒸馏标记，降低 Transformer 模型对大量数据的依赖。CaiT[431] 引入类别注意力，以有效增加模型深度。T2T[508] 通过标记融合有效地实现模型的局部化，并借助卷积神经网络先验递归地将相邻标记聚合为一个标记，从而引入层次化深窄结构。凭借置换等变性，Transformer 摆脱了卷积神经网络的平移不变性，能够建模长程依赖且归纳偏置更少，因此比卷积神经网络更强大，也更容易迁移到下游任务。在

当前大模型和大型数据集范式下，Transformer 已逐渐取代卷积神经网络，成为计算机视觉领域的主流模型。

3.1.2 自监督预训练。 与改进骨干架构并行，深度学习也受益于自监督预训练，因为它能够利用更大规模的（未标注）训练数据集。在此，我们总结与人工智能生成内容最相关的预训练技术，并根据训练数据类型（例如语言、视觉和联合预训练）对其进行分类。

语言预训练。 语言预训练方法主要有三种类型。第一种通过掩码对编码器进行预训练，代表性工作是 BERT [87]（见图 7）。具体而言，BERT 根据未掩码的词元预测被掩码的语言词元。掩码后再预测的预训练任务与下游任务之间存在显著差异，因此，像 BERT 这样的掩码语言建模方法很少在未经微调的情况下用于文本生成。相比之下，自回归语言预训练方法适用于少样本或零样本文本生成。GPT系列 [40, 338, 339] 是最受欢迎的一类，它采用解码器而非编码器。具体而言，GPT-1 [338] 是该系列的首个模型，GPT-2 [339]和 GPT-3 [40]则进一步研究了海量数据和大模型在迁移能力方面的作用。基于 GPT-3，ChatGPT 取得了空前成功，近期吸引了广泛关注。此外，还有一类语言模型采用了与原始 Transformer 一样的编码器和解码器。B ART [226] 通过多种方式对输入加入噪声，并预测原始的干净输入，其方式类似于去噪自编码器。MASS [400] 和 PropheNet [332] 遵循 BERT 的做法，将掩码序列作为编码器的输入，再由解码器以自回归方式预测被掩码的词元。T5 [340] 则将被掩码的词元替换为一些随机词元。

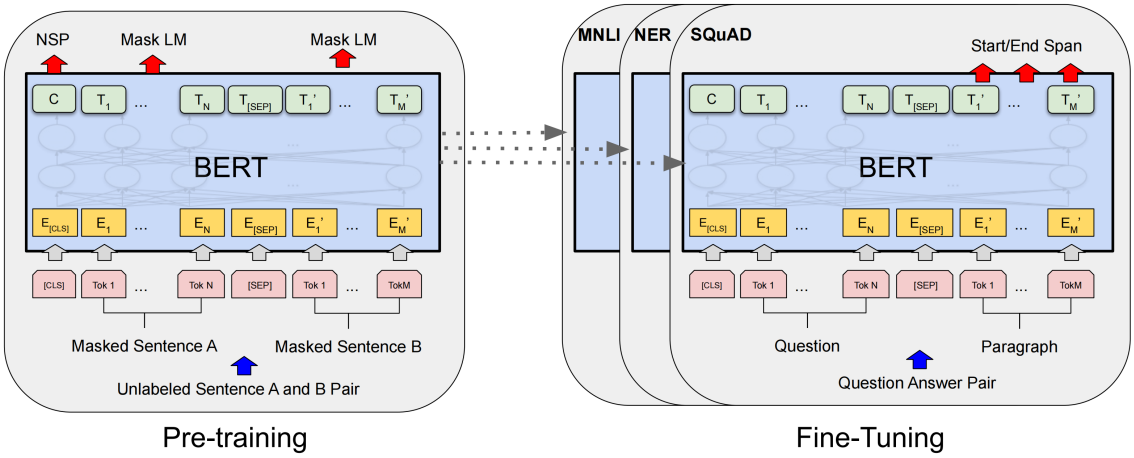


图 7. BERT 结构（图片来源： [87]）。

视觉预训练。 为了在预训练期间更好地学习视觉数据的表征，自监督学习（SSL）已得到广泛应用，我们将其称为视觉自监督学习。视觉自监督学习经历了三个阶段。早期研究着重设计各种前置任务，例如拼图任务 [303] 或预测旋转角度 [121]。与从头训练相比，此类预训练能在下游任务上取得更好的演奏表现，这推动了对比学习方法的发展 [54,142, 520]。对比学习采用联合嵌入，最小化增强后图像之间的表征距离。稿件已提交至美国计算机协会

以学习增强不变表征。纯联合嵌入中的表征可能会坍缩为常数，而与输入无关；因此，对比学习还会同时最大化表征与负样本之间的距离。SimSiam [55]和 BYOL [129]也探索了无负样本的联合嵌入方法。文献 [521]探讨了 SimSiam 如何在没有负样本的情况下发挥作用。受 BERT 在自然语言处理预训练中的成功启发，BEiT [30]将掩码建模应用于视觉领域，其成功依赖于预训练的变分自编码器来获取视觉词元。掩码自编码器（MAE） [141] （见图 8）将其简化为端到端的去噪框架，通过未掩码图像块预测被掩码图像块。MAE 的表现优于对比学习和无负样本的联合嵌入方法，已成为视觉自监督学习框架的一种新变体。感兴趣的读者可参阅 [519]了解更多详情。

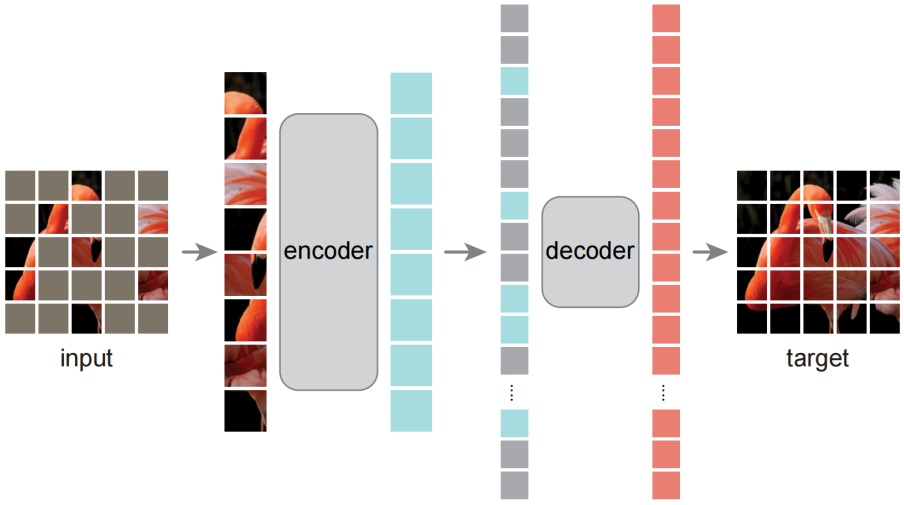


图 8. MAE 结构（图片来源： [141]）。

联合预训练。借助从互联网收集的大型图文对数据集，多模态学习 [29, 487]在学习数据表征方面取得了前所未有的进展，其中跨模态匹配 [115]处于前沿。对比预训练被广泛用于在同一表征空间中匹配图像嵌入和文本编码 [180, 336, 507]。CLIP [336] （见图 9）是这一方向的开创性工作，并被用于众多文本到图像模型，例如 DALL-E 2 [342],UPainting [241], DiffusionCLIP [206]。ALIGN [180] 通过噪声文本监督扩展了 CLIP，使图文数据集无需清洗即可扩展到更大规模（从 4 亿增至 18 亿）。Florence [507] 进一步拓展了跨模态共享表征的范围，从粗粒度场景到精细物体，从静态图像到动态视频等。因此，学到的共享表征更加通用，并展现出更优异的性能 [507]。

3.2 人工智能中的生成技术

深度生成模型（DGMs）是一类利用神经网络生成样本的概率模型。早期的生成建模尝试主要聚焦于使用自编码器进行预训练 [28, 154, 365]。一种带有掩码的自编码器变体已发展成为主流的自监督学习框架，感兴趣的读者

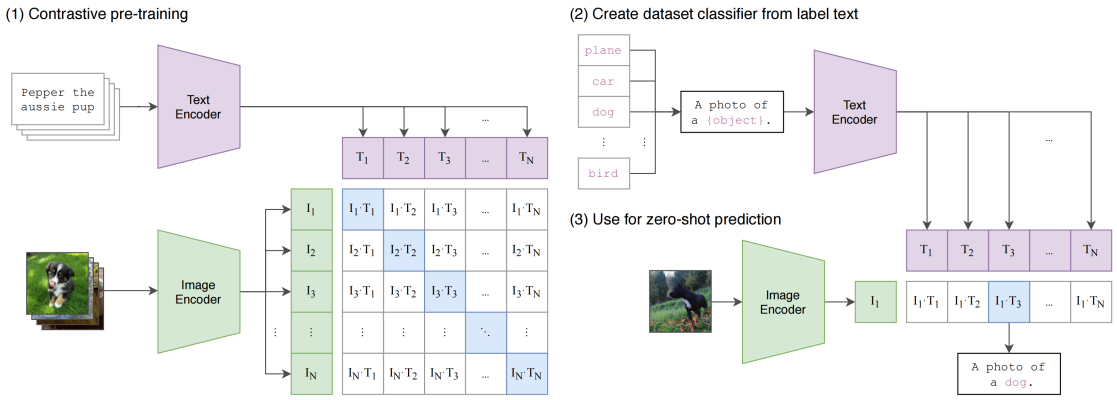


图 9. CLIP 结构（图片取自 [336]）。

可以查阅关于掩码自编码器的综述 [519]。除非另有说明，本综述中深度生成模型的应用场景仅考虑生成新数据。生成的数据通常具有高维特征，因此，预测样本的标签不被视为生成建模，而是判别建模，尽管从技术上讲，标签之类的内容也属于生成结果。

目前已出现众多深度生成模型，可分为两大类：基于似然的模型和基于能量的模型。基于似然的概率模型，例如自回归模型 [126] 和流模型 [90],，具有可解析的似然，因此可以直接根据观测（训练）数据的对数似然来优化模型权重。变分自编码器（VAE） [210], 的似然无法完全解析，但可以优化一个可解析的下界，因此 VAE 也被归入基于似然的模型，这类模型定义了归一化概率。相比之下，基于能量的模型 [128, 153] 以未归一化概率（又称能量函数）为特征。由于不受归一化常数可解析性的约束，基于能量的模型在参数化方面更加灵活，但训练起来较为困难 [403]。值得注意的是，尽管生成对抗网络和扩散模型的出发点不同，但它们与基于能量的模型有着密切联系。下文将分别介绍各类基于似然的模型，然后讲解如何训练基于能量的模型，以及生成对抗网络和扩散模型背后的机制。

3.2.1 基于似然的模型。 **自回归模型**。 自回归模型学习序列数据的联合分布，并将序列中前一时间步的变量作为输入，预测序列中的每个变量。如公式 2 所示，自回归模型假设联合分布 $p_\theta(x)$ 可以分解为条件分布的乘积。

$$p_\theta(x) = p_\theta(x_1)p_\theta(x_2|x_1)...p_\theta(x_n|x_1, x_2, ..., x_{n-1}), \tag{2}$$

尽管二者都依赖于先前的时间步，但自回归模型与循环神经网络架构有所不同：在循环神经网络中，先前的时间步以隐藏状态的形式呈现，而在自回归模型中，则作为输入提供给模型。换言之，自回归模型可以看作一个以前面所有时间步的变量作为输入的前馈网络。早期研究使用不同的函数来估计条件分布，从而对离散数据进行建模，例如，全可见 sigmoid 信念网络（FVSBN）中的逻辑回归 [114]，以及神经自回归分布估计中的单隐藏层神经网络。稿件已提交至美国计算机协会

（NADE） [221]。后续研究进一步将建模范围扩展到连续变量 [437, 438]。自回归方法已广泛应用于多个领域，包括计算机视觉（PixelCNN [441]和 PixelCNN++ [373]）、音频生成（WaveNet [440]）以及自然语言处理（Transformer [443]）。

变分自编码器。自编码器是一类模型：它们首先通过编码器将输入映射到低维潜在层，然后通过解码器重构输入。整个编码器—解码器过程旨在学习底层数据模式并生成未见样本 [310]。变分自编码器（VAE） [210]是一种自编码器，它从潜在空间 z 中学习数据分布 $p(x)$，即 $p(x) = p(x|z)p(z)$，其中 $p(x|z)$ 由解码器学习。为了获得$p(z)$，VAE [210] 采用贝叶斯定理，并由编码器近似后验分布 $p(z|x)$。VAE 模型通过正则项 [13] 朝着似然目标进行优化。

3.2.2 基于能量的模型。由于似然易于计算，自回归模型和流模型可以直接根据数据的对数似然优化参数。这会迫使模型受到特定形式的约束。例如，自回归模型需要分解为条件概率的乘积，而流模型必须采用可逆变换。

基于能量的模型将概率指定为一个未知归一化常数下的值，因此也称为未归一化概率模型。为不失一般性，假设基于能量的模型仅涉及一个变量 $\boldsymbol{x}$，并将其能量记为 $E_\theta(\boldsymbol{x})$。其概率密度计算如下：

$$p_\theta(\boldsymbol{x}) = \frac{\exp(-E_\theta(\boldsymbol{x}))}{z_\theta} \tag{3}$$

其中，z_θ 是所谓的归一化常数，定义为 $z_\theta = \int \exp(-E_\theta(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}$。$z_\theta$ 是一个难以计算的积分，使得优化基于能量的模型颇具挑战性。

马尔可夫链蒙特卡罗和噪声对比估计。早期优化基于能量的模型的方法，尝试使用马尔可夫链蒙特卡罗（MCMC）方法估计对数似然的梯度，但这需要繁琐地绘制随机样本。因此，一些研究致力于提高 MCMC 的效率，其中一项代表性工作是朗之万马尔可夫链蒙特卡罗 [128, 316]。尽管如此，执行 MCMC 以获取所需结果仍需要大量计算，而对比散度（CD） [153] 是一种广受欢迎的降低计算量的方法，它通过近似计算实现，并衍生出多种变体：持久对比散度 [425], 、平均场对比散度 [468], 和多网格对比散度 [116]。另一类研究通过噪声对比估计（NCE） [137], 优化基于能量的模型，该方法将概率模型与另一种噪声分布进行对比。具体而言，它优化以下损失函数：

$$\mathbb{E}_{p_d}\left[\ln \frac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + q_\phi(\boldsymbol{x})}\right] + \mathbb{E}_{q_\phi}\left[\ln \frac{q_\phi(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + q_\phi(\boldsymbol{x})}\right], \tag{4}$$

分数匹配。在优化基于能量的模型时，另一种广受欢迎的无 MCMC 方法是最小化模型与观测数据之间对数概率密度的导数。对数概率密度函数的一阶导数称为分布的乐谱 （$s(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$），因此，这种方法通常称为分数匹配。遗憾的是，数据的分数函数 $s_d(\boldsymbol{x})$ 无法获得。人们已尝试多种方法来缓解这一问题，其中一种代表性方法称为去噪分数匹配 [446]。去噪分数匹配利用噪声样本近似数据的乐谱。模型将噪声样本作为输入，并预测其中的噪声。因此，可以通过迭代去除噪声，从噪声中采样得到干净样本 [374, 401]。

3.2.3 两种明星模型：从生成对抗网络到扩散模型。说到深度生成模型，你最先想到的是什么？答案取决于你的背景，不过，生成对抗网络绝对是最常被提及的模型之一。

生成对抗网络（GAN） [124] 由 Ian J. Goodfellow 及其团队于2014年首次提出。2016年，Yann Lecun 称其为“过去10年机器学习领域最有趣的创意”。作为生成高质量图像的开创性工作，生成对抗网络一直被广泛视为图像合成这一艰巨任务的事实标准模型。近年来，这种长期主导地位受到一类名为扩散模型 [156]的新型深度生成模型的挑战。扩散模型取得的巨大成功始于图像合成，并扩展至视频、音频、文本、图等其他模态。鉴于它们在生成式人工智能发展中的重要影响，我们将在介绍其他深度生成模型系列之前，先概述生成对抗网络和扩散模型。

生成对抗网络。 生成对抗网络的架构如图 10 所示。生成对抗网络由两个网络组件构成：判别器（D）和生成器（G）。D 用于区分真实图像与 G 生成的图像，而 G 则旨在欺骗 D。给定一个潜变量 $z \sim p_z$，G 的输出为 G(z)，构成一个概率分布p_g。生成对抗网络的目标是使p_g 逼近观测数据分布 p_{data}。这一目标通过对抗学习实现，可将其解释为一个极大极小游戏 [375]:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{data}} \log[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z} \log\left[1 - D(G(\mathbf{z}))\right]. \tag{5}$$

其中，D 经过训练后能够最大化为真实图像和生成图像赋予正确标签的概率，并用于指导 G 的优化，使其生成更逼真的图像。由于对抗训练的特性，生成对抗网络可能存在训练不稳定、生成多样性较低等缺点。生成对抗网络与自回归模型的根本区别在于，生成对抗网络学习隐式数据分布，而后者学习由模型结构施加的先验所支配的显式分布。

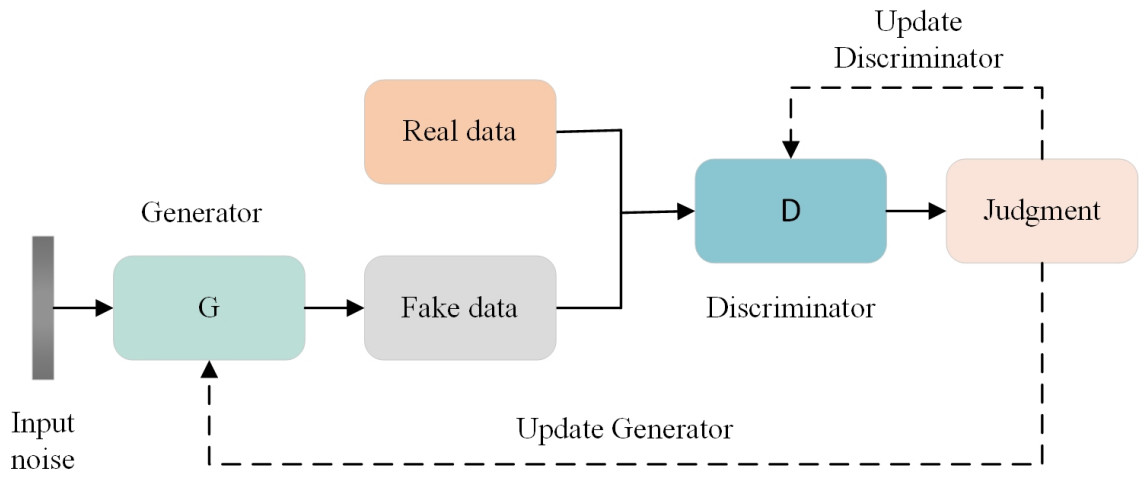


图 10. GAN结构示意图。

扩散模型。 扩散模型是一种特殊形式的层次化变分自编码器，过去几年间其应用呈现爆发式增长 [45, 73, 245, 320, 435]。扩散模型（图 11）也称为去噪扩散概率模型（DDPM）或基于分数的生成模型，它们能够生成与训练数据相似的新数据 [156]。受非平衡热力学启发，DDPM 可定义为一个参数化的马尔可夫链。稿件已提交至美国计算机协会

该马尔可夫链包含多个扩散步骤，用于逐步向训练数据中添加随机噪声，并学习逆转扩散过程，从纯噪声中构造所需的数据样本。

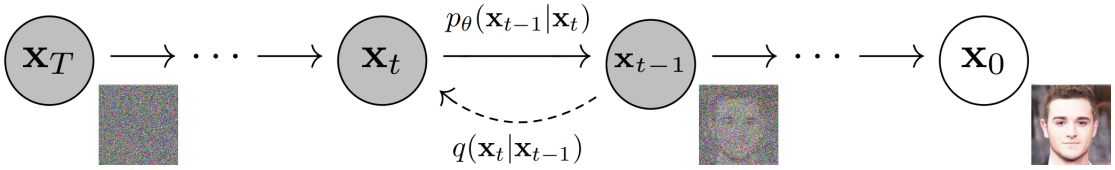


图 11. 用于图像生成的扩散模型（图片取自 [156]）。

在前向扩散过程中，去噪扩散概率模型通过连续添加高斯噪声来破坏训练数据。给定数据分布$\mathbf{x}_0 \sim q(\mathbf{x}_0)$，去噪扩散概率模型通过逐步扰动输入数据，将训练数据映射为噪声。具体而言，该过程由一个简单的随机过程实现：从一个数据样本开始，迭代生成噪声更大的样本$\mathbf{x}_T$，并使用 $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$，其中采用简单的高斯扩散核：

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{6}$$

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{7}$$

其中， T和 β_t 分别表示扩散步数和超参数。为简便起见，我们仅讨论以高斯噪声作为转移核的情况，在式 7 中记为 N。给定 $\alpha_t := 1 - \beta_t$和 $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$，即可按如下方式获得任意步骤 t 的加噪图像：

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{8}$$

在反向去噪过程中，去噪扩散概率模型通过逆转加噪过程来学习恢复数据，也就是说，它通过迭代去噪来撤销正向扩散。这个过程代表数据合成，而去噪扩散概率模型经过训练后，能够将随机噪声转化为真实数据，从而生成数据。它也被正式定义为一种随机过程，从 $p_\theta(T)$ 开始对输入数据进行迭代去噪，并生成能够遵循真实数据分布 $q(x_0)$ 的 $p_\theta(x_0)$。因此，该模型的优化目标如下：

$$E_{t\sim\mathcal{U}(1,T),\mathbf{x}_0\sim q(\mathbf{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\lambda(t)\,\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \tag{9}$$

去噪扩散概率模型的正向和反向过程通常都需要数千个步骤，以逐步注入噪声，并在生成过程中进行去噪。

4 人工智能生成内容任务：文本生成

自然语言处理研究自然语言的两项基本任务：理解与生成。这两项任务并非完全独立，因为生成恰当的文本往往取决于对某些文本输入的理解。例如，语言模型通常会将一个文本序列转换为另一个序列，这构成了文本生成的核心任务，包括机器翻译、文本摘要和对话系统。除此之外，文本生成沿着两个方向发展：可控性和多模态。第一个方向旨在使生成的内容

4.1 文本到文本

4.1.1 聊天机器人。 对话系统（聊天机器人）的主要任务是促进人类与机器之间更顺畅的交流 [85, 299]。根据应用中是否指定任务，对话系统可分为两类：(1) 面向任务的对话系统（TOD） [323, 502, 533] 和 (2) 开放域对话系统（OOD） [4, 532, 541]。具体而言，面向任务的对话系统专注于完成任务并解决特定问题（例如餐厅预订和票务预订） [533]。与此同时，开放域对话系统通常由数据驱动，旨在不受任务或领域限制地与人类聊天 [353, 533]。

面向任务的系统。 面向任务的对话系统可分为模块化系统和端到端系统。模块化方法包括四个主要部分：自然语言理解（NLU） [395, 409], 对话状态跟踪（DST） [382, 462], 对话策略学习（DPL） [169, 483], 和自然语言生成（NLG） [25, 99]。NLU 将用户输入编码为语义槽位后，DST 和 DPL 会决定下一步操作，再由 NLG 将其转换为自然语言，作为最终响应。这四个模块旨在以可控的方式生成响应，并且可以单独优化。然而，有些模块可能不可微，而且改进单个模块未必能提升整个系统 [533]。为了解决这些问题，端到端方法要么通过使每个模块都可微来实现端到端训练流程 [139, 162], ，要么在系统中采用单个端到端模块 [498, 531]。模块化系统和端到端系统仍面临一些挑战，包括如何提高 DST 的跟踪效率 [208, 312] ，以及如何在有限数据条件下提升端到端系统的响应质量 [145, 148, 282]。

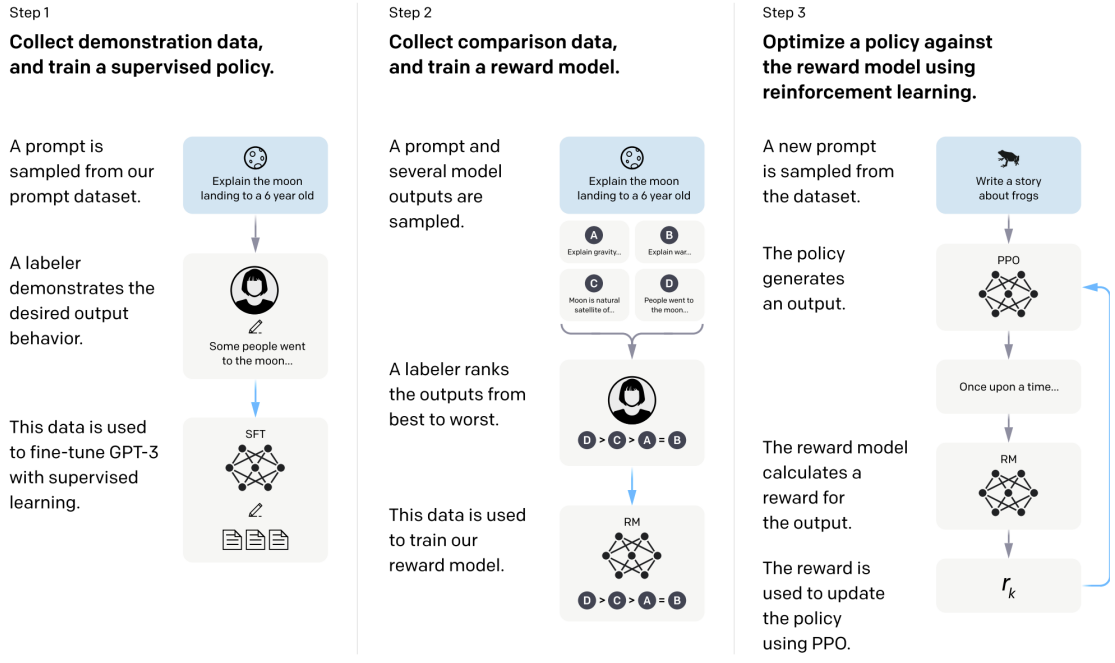


图 12. 展示 OpenAI 训练 ChatGPT 的三个步骤的示意图（图片引自 [311]）。

开放域系统。开放域系统旨在与用户交谈，不受任务和领域限制 [353,533], ，可分为三种类型：基于检索的系统、生成式系统和集成系统 [533]。具体而言，基于检索的系统总是从响应语料库中查找已有响应，而生成式系统能够生成训练集中可能不存在的响应。集成系统结合基于检索的方法和生成式方法，具体做法是选择最佳响应，或使用生成式方法改进基于检索的模型 [378, 533, 546]。以往的研究从多个方面改进开放域系统，包括对话上下文建模 [105, 181,250, 282], 、提升响应连贯性 [9, 117, 251, 483]和多样性 [31, 211, 335, 408]。最近，ChatGPT（见图 12）取得了前所未有的成功，也属于开放域对话系统的范畴。除了回答各种问题，ChatGPT 还可用于论文写作、代码调试、表格生成等。

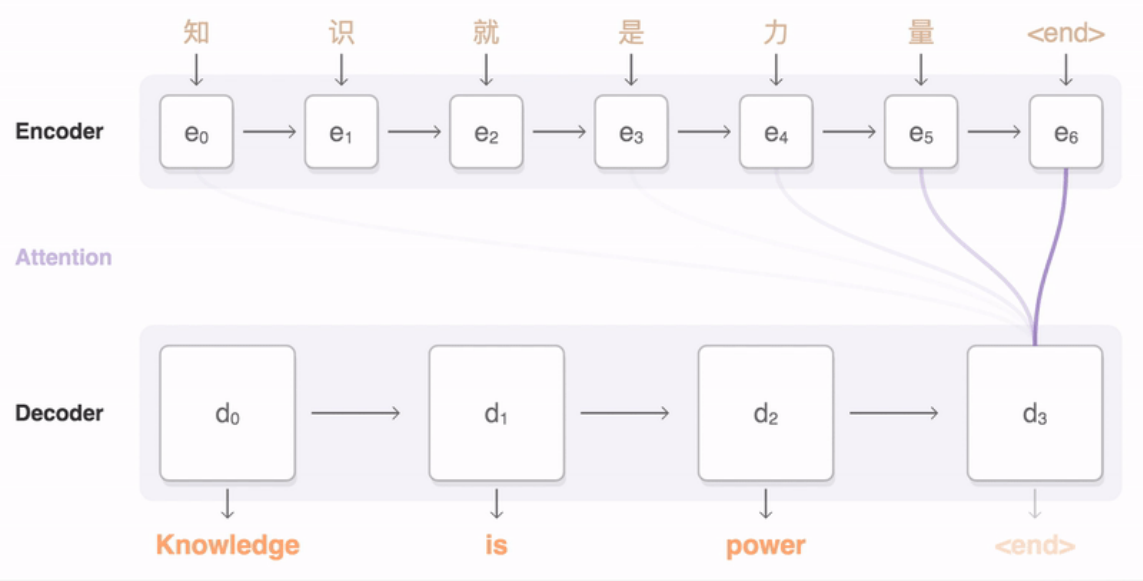


图 13. 机器翻译示例（图片取自 [39]）。

4.1.2 机器翻译。顾名思义，机器翻译会自动将文本从一种语言翻译成另一种语言 [171, 497]（见图 13）。随着深度学习取代基于规则 [108] 和统计 [212, 213] 的方法，神经机器翻译（NMT）只需极少的语言学专业知识 [399, 451] ，并凭借更强的句子长距离依赖捕捉能力，成为主流方法 [62]。神经机器学习的成功主要归功于语言模型 [34], ，这类模型会预测某个词在给定前面词语的条件下出现的概率。Seq2seq [413] 是率先将编码器-解码器循环神经网络结构 [191] 应用于机器翻译的工作。当句子变长时，Seq2seq [413] 的演奏会下降。因此， [24] 提出了一种注意力机制，通过额外的词对齐来帮助翻译长句。随着注意力机制日益受到重视，2006年，谷歌神经机器翻译系统相较于谷歌基于短语的生产系统，将人类的翻译工作量减少了约 60%，弥合了人类翻译与机器翻译之间的差距 [475]。研究人员也尝试将基于卷积神经网络的架构用于神经机器翻译 [190, 192], ，但未能达到

与注意力机制增强的循环神经网络相当的演奏水平 [24]。卷积序列到序列模型 [120] 使卷积神经网络能够兼容注意力机制，表明卷积神经网络能够达到与循环神经网络相当甚至更好的演奏水平。然而，这一改进后来被另一种名为 Transformer 的架构超越 [443]。神经机器翻译以循环神经网络或 Transformer 为架构时，通常采用自回归生成模型；在推理过程中，贪心搜索只会考虑概率最高的词来预测下一个词。

神经机器翻译的一个趋势是在低资源设置下实现令人满意的演奏，此时模型使用有限的双语语料库进行训练 [458]。缓解数据稀缺的一种方法是利用辅助语言，例如与其他语言对进行多语言训练 [187, 383, 547]，或以英语为中间枢轴语言进行枢轴翻译 [58, 350]。另一种常见方法是利用预训练语言模型，例如 BERT [87] 或 GPT [338]。例如，研究表明 [359]，使用 BERT [87] 或 RoBERTa [259] 初始化模型权重，能显著提升英德翻译的演奏。GPT系列模型 [40, 338, 339] 无需微调也展现出竞争力。最近，ChatGPT 在机器翻译方面展现了实力，其表现可与商业产品（例如 Google翻译）相媲美 [182]。

4.2 多模态文本生成

4.2.1 图像到文本。图像到文本也称为图像描述，指用自然语言描述给定图像的内容（见图 14）。该领域的一项开创性工作是神经图像描述（NIC） [447],，它采用卷积神经网络作为编码器，提取输入图像的高层表示，然后将这些表示输入循环神经网络解码器，以生成图像描述。这种两步式编码器-解码器架构已广泛应用于后续的图像描述研究中，我们分别将这两个阶段称为视觉编码 [407] 和语言解码。下面，我们将首先回顾图像描述这两个阶段的发展历程和近期趋势。

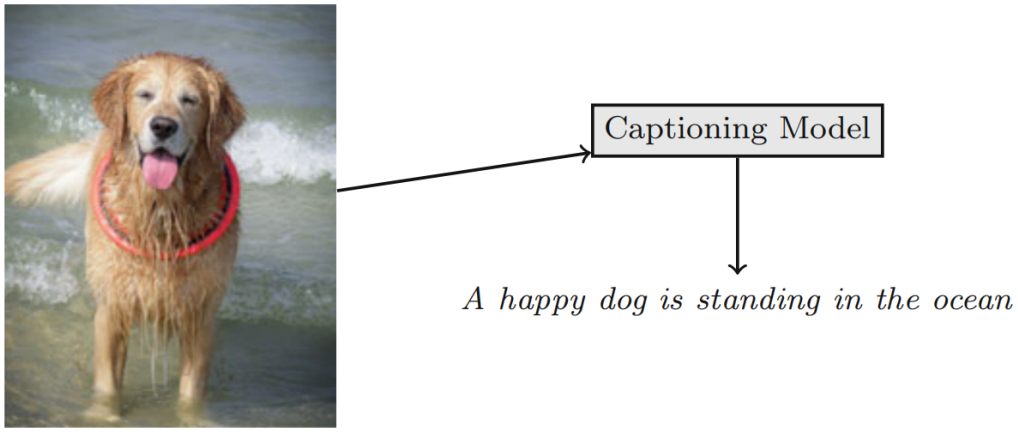


图 14. 图像描述示例（图源自 [109]）。

视觉编码。 提取图像的有效表示是视觉编码模块的主要任务。从采用 GoogleNet [417] 提取输入图像全局特征的 NIC [447] 开始，许多研究采用各种卷积神经网络骨干网络作为编码器，包括 [195] 中的 AlexNet [216] 和 [92, 272]中的 VGG网络 [393]。然而，语言模型很难仅凭全局视觉特征生成细粒度图像描述。后续研究引入了注意力机制。稿件已提交至美国计算机协会

用于提取细粒度视觉特征，包括针对不同网格上的 CNN特征 [56, 264, 463, 484]或不同视觉区域 [16, 200, 518]的注意力。另一类研究 [500, 536] 采用图神经网络对不同区域之间的语义和空间关系进行编码。然而，由人类定义的图结构可能会限制各元素之间的交互 [407], ，而自注意力方法 [231, 501, 530]（包括视觉Transformer [256]）能够连接所有元素，从而缓解这一问题。

语言解码。在图像描述任务中，语言解码器通过预测给定词序列的概率来生成图像描述 [407]。受自然语言处理领域突破的启发，语言解码器的骨干网络从循环神经网络 [200, 264, 447, 456] 演进至 Transformer [132, 149, 231], ，实现了显著的演奏提升。除了视觉编码器—语言解码器架构之外，还有一类研究采用类似 BERT 的架构，在单个模型的早期阶段融合图像和图像描述 [244, 526, 542]。例如， [542]采用单个编码器来学习图像和文本的共享空间，先在大型图文语料库上进行预训练，再针对图像描述任务进行微调。

4.2.2 语音转文本。语音转文本生成，也称为自动语音识别（ASR），是将口语，特别是语音信号，转换为相应文本的过程 [173, 347] （见图 15）。自动语音识别有许多潜在应用，例如语音拨号、计算机辅助语言学习、图像描述生成，以及 Alexa 和 Siri 等虚拟助手。自 20 世纪 50 年代以来，自动语音识别一直是一个令人振奋的研究领域 [194, 270, 345]，其技术也从隐马尔可夫模型（HMM） [188, 225] 发展到基于深度神经网络的系统 [75, 127, 152, 297, 473]。

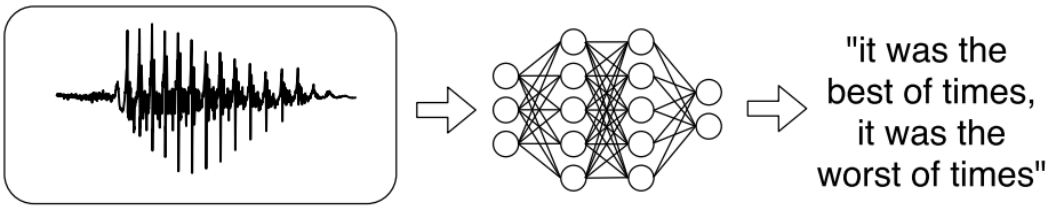


图 15. 语音识别示例（图片来源于 [46]）。

各类研究主题与挑战。以往的研究从多个方面改进了自动语音识别系统。许多研究探讨了语音信号的不同特征提取方法 [270], ，包括时域特征（例如离散小波变换 [287, 419]）和频谱特征，例如最常用的梅尔频率倒谱系数（MFCC） [61, 69, 429]。另一个研究方向是改进系统流程 [355] ，从多模型 [268] 转向端到端模型 [161, 233, 234, 296, 453]。具体而言，多模型系统 [268, 270] 首先学习声学模型（例如将特征映射到音素的音素分类器），然后学习用于生成词语输出的语言模型 [355]。另一方面，端到端模型则直接根据输入音频预测转写文本 [161, 233, 234, 296, 453]。尽管端到端模型在多种语言和方言中都取得了令人瞩目的演奏，但仍面临许多挑战。首先，将其应用于低资源语音任务仍然困难，因为获取大量标注训练数据既昂贵又耗时 [104, 355]。其次，这些系统可能难以处理包含特定未登录词的语音；它们在训练数据上表现良好，却未必能很好地泛化到新的或未见过的数据 [104, 334]。此外，训练数据中的偏见也会影响监督式自动语音识别系统的演奏，导致其在某些人群或说话风格上的准确率较低 [35]。



图 16. 图像复原示例（图片取自 [452]）。

低资源语音任务。研究人员致力于开发新技术，以克服自动语音识别系统面临的挑战。其中，我们主要讨论缺乏受损语音数据的低资源语音问题 [355]。一类研究 [321, 346] 采用多任务学习，为不同任务优化共享编码器。与此同时，近年来，自监督自动语音识别系统已成为一个活跃的研究领域，无需依赖大量带标签样本。具体而言，自监督自动语音识别系统首先使用海量无标签语音数据对模型进行预训练，然后使用较小规模的带标签数据集进行微调，以提高自动语音识别系统的效率。这种方法可用于低资源语言，处理不同的说话风格或噪声条件，以及转写多种语言 [23, 71, 255, 492]。

5 人工智能生成内容任务：图像生成

与文本生成类似，图像合成任务也可以根据输入控制方式划分为不同类别。由于输出是图像，一种直接的控制方式就是使用图像。图像类型控制催生了许多任务，例如超分辨率、去模糊、剪辑、翻译等。图像类型控制的一个局限是缺乏灵活性。相比之下，文本引导控制使人类能够随心所欲地生成任意内容、任意风格的图像。文本生成图像属于跨模态生成，因为输入文本与输出图像属于不同模态。

5.1 图像到图像

5.1.1 图像复原。图像复原旨在解决一种典型的逆问题，即从相应的退化图像中恢复干净图像，如图 16 所示。由于从退化图像到干净图像存在无限种可能的映射，这类逆问题具有不适定性，因此并不容易解决。图像退化有两个来源：原始图像中的信息缺失，以及向干净图像中添加了不需要的内容。前一种退化包括以低分辨率拍摄照片而丢失部分细节信息、裁剪某个区域，以及将彩色图像转换为灰度图像。相应地，复原任务依次包括图像超分辨率、图像修复和图像着色。另一类复原任务旨在去除不需要的扰动，例如去噪、去雨、去雾、去模糊等。早期的复原技术主要采用数学和统计建模来消除图像退化，包括用于去噪的空间滤波器 [123, 392, 529],提交给美国计算机协会的稿件

用于去模糊的核估计 [485, 489]。近年来，基于深度学习的方法 [42, 59, 93, 177, 248, 252, 481, 486]凭借其灵活性以及优于传统方法的视觉质量，已成为图像复原任务中的主流方法。卷积神经网络广泛用作图像复原的基础模块 [94, 411, 442, 459], ，与此同时，近期研究探索了更强大的Transformer架构，并在图像超分辨率 [247], 、图像着色 [218], 和图像修复 [240]等多种任务中取得了令人瞩目的演奏表现。还有一些研究将卷积神经网络和Transformer的优势结合起来 [103, 534, 535]。

用于图像复原的生成式方法。典型的图像复原模型通过重建损失学习源图像（退化图像）与目标图像（清晰图像）之间的映射。根据任务的不同，可以通过对清晰图像施加各种扰动来生成训练数据对，包括分辨率下采样和灰度变换。为了保留更多高频细节并生成更写实的图像，生成模型被广泛用于图像复原，例如用于超分辨率的生成对抗网络 [223, 460, 528] 和图像修复 [42, 252, 298]。然而，基于生成对抗网络的模型通常面临训练过程复杂和模式崩溃的问题。这些缺点以及扩散模型的巨大普及，促使近期大量研究采用扩散模型进行图像复原任务 [199, 232, 265, 349, 367, 369]。生成对抗网络和扩散模型等生成式方法也可以根据单张退化图像生成多个不同的清晰图像。

从单任务到多任务。 现有的大多数复原方法都会针对不同形式的图像退化分别训练模型。这限制了它们在实际场景中的有效性，因为图像往往会受到多种退化的共同影响。为解决这一问题，一些研究 [6, 207, 391, 540] 引入了多失真数据集，将各种形式的退化及其不同强度结合起来。一些研究 [207, 258, 505, 509]提出复原模型，由不同的子网络分别处理不同的退化。另一些研究 [228, 242,391, 410, 540] 则依靠注意力模块或引导子网络，在应对不同退化时辅助复原网络，使单个网络能够处理多种退化。

5.1.2 图像编辑。 与旨在提升图像质量的图像复原不同，图像编辑是指修改图像以满足特定需求，例如风格迁移（见图 17）。从技术上说，一些图像复原任务（如图像着色）也可以被视为图像编辑，因为添加颜色本身可能就是所需的效果。现代相机通常具备一些基本的编辑功能，例如锐度调整 [524], 自动裁剪 [525], 红眼消除 [396],等。然而，在人工智能生成内容领域，我们更关注能够以各种方式改变图像语义的高级图像编辑任务，例如改变内容、风格、对象属性等。

一类图像编辑方法旨在修改图像中主体对象（如人脸）的属性（如年龄）。典型应用是人脸属性编辑，可改变发型、年龄甚至性别。基于预训练 CNN 编码器，一系列开创性工作采用基于优化的方法 [236, 436], ，但由于其迭代特性，耗时较长。另一类工作则采用基于学习的方法，直接生成图像，发展趋势从单一属性 [237, 385] 扩展到多个属性 [146, 209, 478]。上述大多数方法的缺点是依赖属性的标注标签，因此，人们引入了无监督学习，以解耦不同属性 [60, 386]。

另一类图像编辑方法通过组合两幅图像来改变语义。例如，图像变形 [185]对两幅图像的内容进行插值，而风格迁移 [119] 则生成一幅新图像，其中内容来自一幅图像，风格来自另一幅图像。图像变形的一种简单方法是在像素空间中进行插值，但这会造成明显的伪影。相比之下，在潜在空间中进行插值可以考虑视角变化，并生成平滑的图像。这两幅图像的潜在空间可通过 GAN 反演方法 [477]获得。许多研究 [1, 490, 544, 545] 探索了预训练 GAN 的潜在空间在图像变形中的应用。对于风格



图 17. 风格迁移作为一种图像编辑形式的示例（图片来源于 [118]）。

迁移中，一种特定的基于风格的生成对抗网络变体 StyleGAN [197]颇受欢迎。从较早的层到较后的层，StyleGAN 控制的属性由较粗粒度（如结构）逐渐过渡到较细粒度（如纹理）。因此，可以通过混合内容图像较早层的潜在表示和风格图像较后层的潜在表示来使用 StyleGAN 进行风格迁移 [1, 131, 444, 467]。

与复原任务相比，各种编辑任务能够实现更加灵活的图像生成。然而，其多样性仍然有限，通过允许使用其他文本作为输入，这一问题得到了缓解。近年来，基于扩散模型的图像编辑受到广泛关注，并取得了令人瞩目的成果 [48, 150, 206, 450]。DiffusionCLIP [206] 是一项开创性工作，它对预训练扩散模型进行微调，使目标图像与文本对齐。相比之下，LDEdit [48] 基于潜在扩散模型 [357]，无需进行微调。一类工作探讨了图像编辑中的掩码问题，包括如何让人工设计的掩码区域与背景无缝衔接 [48, 150, 206, 450]。投稿至美国计算机协会的稿件

另一方面，DiffEdit [72]提出自动预测用于指示待编辑部分的掩码。还有一些工作基于扩散模型和文本引导来编辑三维物体 [47, 205, 230]。

5.2 多模态图像生成

5.2.1 文本生成图像。文本生成图像（T2I）任务旨在根据文本描述生成图像（见图??。），其发展可以追溯到根据标签或属性生成图像的时期 [405, 495]。AlignDRAW [271] 是利用自然语言生成图像的开创性工作；令人印象深刻的是，AlignDRAW [271] 能够根据新颖的文本描述生成图像，例如“一个停止标志正在蓝天下飞翔”。近年来，文本生成图像领域的进展可分为三类：基于生成对抗网络的方法、自回归方法和基于扩散的方法。

基于生成对抗网络的方法。AlignDRAW [271] 的局限在于生成的图像不够逼真，且需要额外的生成对抗网络进行后处理。基于深度卷积生成对抗网络（DC-GAN） [337], [348] 的方法，是首个从字符级到像素级的端到端可微架构。为生成高分辨率图像并稳定训练过程，StackGAN [522]和 StackGAN++ [523]提出了一种多阶段机制，由多个生成器生成不同尺度的图像，并以低分辨率图像为条件生成高分辨率图像。此外，AttnGAN [488] 和 Controlgan [229] 采用注意力网络，根据相关词语对各个子区域进行细粒度控制。

自回归方法。 受自回归 Transformer [443], 成功的启发，一类研究通过将图像映射为词元序列，以自回归方式生成图像，其中 DALL-E [343] 是开创性工作。具体而言，DALL-E [343] 首先利用预训练的离散变分自编码器（dVAE）将图像转换为图像词元，然后训练自回归 Transformer，以学习文本与图像词元的联合分布。一项同期工作 CogView [88] 独立提出了与 DALL-E [343] 相同的理念，但在模糊的 MS COCO 数据集上取得了优于 DALL-E [343] 的 FID [151] 。CogView2 [89] 通过遮蔽不同的词元，将 CogView [88] 扩展到图像描述等多种任务。Parti [504] 则通过将模型规模扩大到 200 亿参数，进一步提升了图像质量。

基于扩散的方法。基于扩散模型的方法近期取得了前所未有的成功并受到广泛关注，可根据其直接作用于像素空间 [300, 368] 还是潜在空间 [342, 357]进行分类。GLIDE [300] 通过将类别条件扩散模型扩展到文本条件设置，超越了 DALL-E；Imagen [368] 则利用预训练的大语言模型（如 T5）捕捉文本语义，进一步提升了图像质量。为降低扩散模型在像素空间中的资源消耗，Stable Diffusion [357] 首先将高分辨率图像压缩到低维潜在空间，再在潜在空间中训练扩散模型。这种方法也称为潜在扩散模型（LDM） [357]。与仅基于图像学习潜在空间的 Stable Diffusion [357] 不同，DALL-E2 [342] 应用扩散模型学习先验，使其对齐 CLIP 的图像空间与文本空间。其他工作还从多个方面改进模型，包括引入空间控制 [20, 449] 和参考图像 [37, 387]。

5.2.2 会说话的人脸。从输出的角度来看，会说话的人脸[537] 任务会生成一系列图像帧，因此从技术上来说，这些帧构成一个视频（见图19）。与通用视频生成（见第6.1节）不同，会说话的人脸需要一张人脸图像作为身份参考，并根据语音输入对其进行编辑。从这个意义上说，会说话的人脸与图像编辑的关系更为密切。此外，会说话的人脸将一个语音片段转换为对应的人脸图像，这与语音识别将一个语音片段转换为对应词语文本的方式类似。鉴于语音识别被视为一种

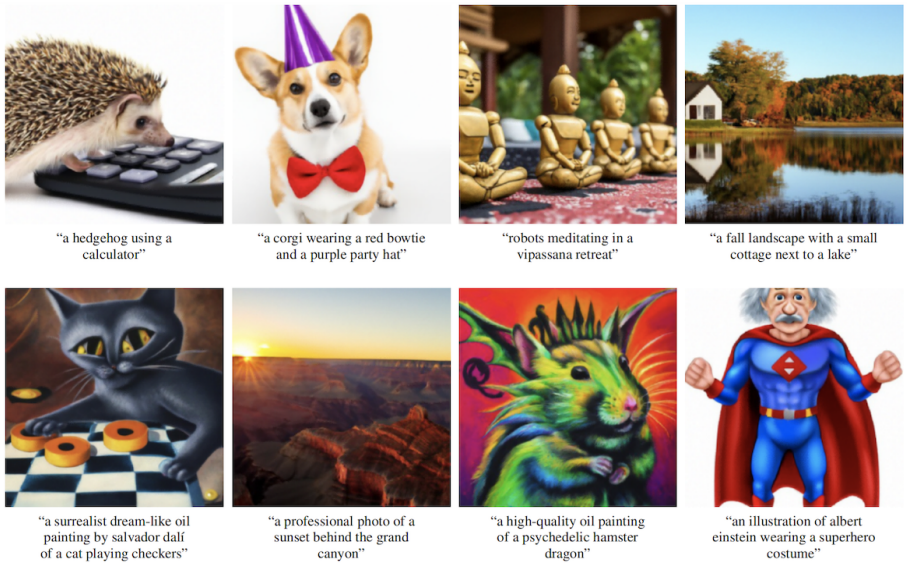


图 18. 文本生成图像示例（图片来自 [300]）。

多模态生成文本任务，本文将会说话的人脸视为一项多模态图像生成任务。在深度学习模型的推动下，语音到头部视频合成模型备受关注，可分为基于二维的方法和基于三维的方法。

基于二维的方法主要依赖关键点、语义图或类似表征来合成说话人脸视频。关键点被用作从低维音频到高维视频的中间层，同时还使用两个解码器将语音与说话人身份解耦，从而生成不受说话人身份影响的视频 [66]，这也是首项使用深度生成模型创建说话人脸的工作。此外，图像到图像转换生成 [178] 也可用于唇部合成，而将独立的视听表征与神经网络相结合，也可用于优化合成效果 [404, 539]

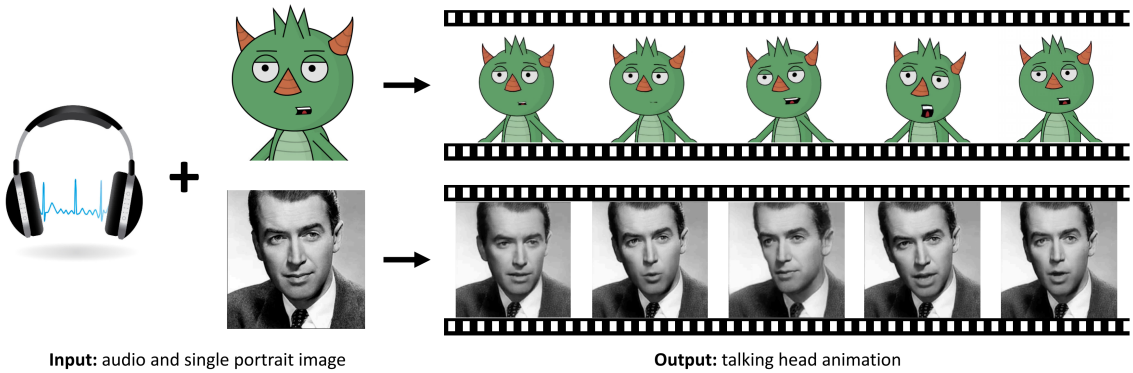


图 19. 会说话的人脸示例（图片来自 [51]）。

另一类工作基于构建三维模型，并通过渲染技术控制运动过程 [219, 414], ，但其缺点是构建成本高。随后，研究者建立了许多基于三维形变模型参数的生成式说话人脸模型 [74, 111, 196, 423] ，使用混合变形 [74], 、FLAME [239], 和三维网格 [352], 等模型，并以音频作为模型输入来生成内容。目前，大多数方法都是直接从训练视频中重建得到的。神经辐射场使用多层感知机来模拟隐式表示，这种表示可以存储三维空间坐标和外观信息，适用于高分辨率场景 [238, 286, 294]。此外，研究者还提出了用于合成不受限制的说话人脸视频的流程和端到端框架 [215, 328], ，以任意未经识别的视频和任意语音作为输入。

6 人工智能生成内容任务：超越文本与图像

6.1 视频

与图像生成相比，视频生成的进展大幅滞后，这主要是因为高维视频数据的建模更加复杂。视频生成不仅涉及像素的生成，还需要确保不同帧之间的语义连贯性。视频生成研究可分为无引导生成和有引导生成（例如，以文本、图像、视频和动作类别为引导）；其中，文本引导的视频生成（见图??）影响力较大，因此受到最多关注。

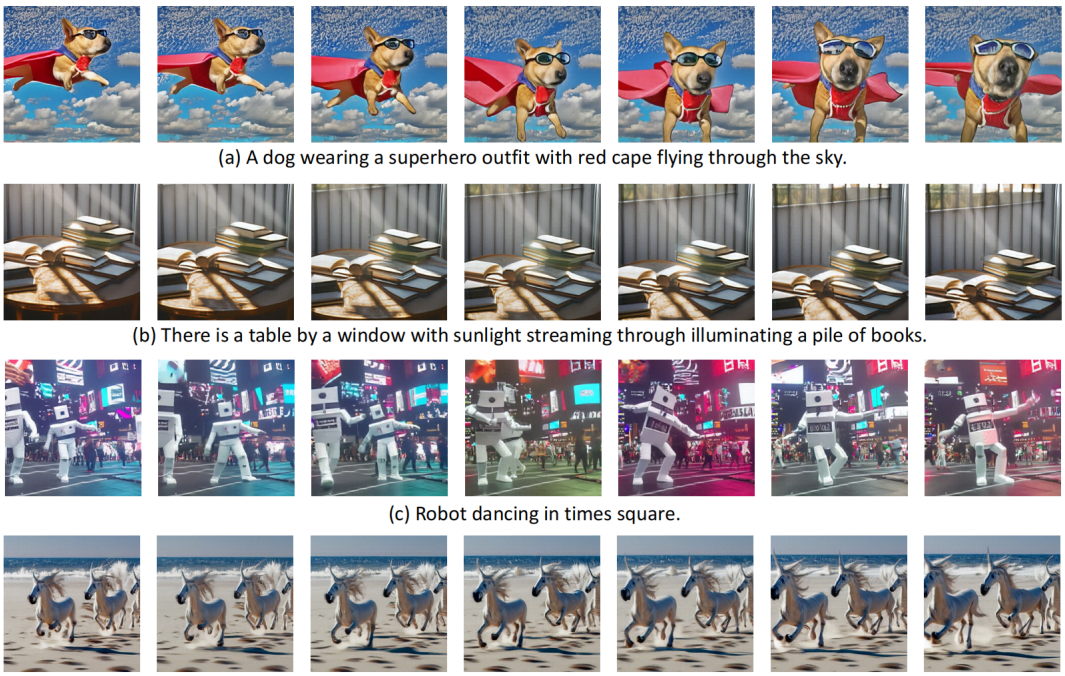


图 20. 文本引导的视频生成示例（图片引自 [394]）。

无引导视频生成。 早期将图像生成从单帧扩展到多帧的研究，仅限于生成海浪等单调但规律的内容。生成的动态纹理 [96, 466] 通常呈现出 Manuscript submittedtoACM

随时间变化的空间重复图案。随着生成模型的发展，许多研究 [2, 68, 305, 370, 433, 448, 512] 将探索范围从简单的动态纹理扩展到真正的视频生成。尽管如此，由于可用的数据集分辨率较低，这些研究的成功仅限于为简单场景生成短视频。近期的研究 [67, 157, 371, 424] 进一步提升了视频质量，其中 [157] 被视为扩散模型领域的开创性工作。

文本引导的视频生成。与能够创作出近乎照片级真实感图像的文生图模型相比，文本引导的视频生成更具挑战性。早期基于变分自编码器或生成对抗网络的研究 [136, 246, 260, 276, 290, 313] 专注于在简单场景中生成视频，例如数字弹跳和人类行走。鉴于向量量化变分自编码器模型在文本引导图像生成方面取得了巨大成功，一些研究 [160, 472] 将其扩展到文本引导的视频生成，从而生成更加逼真的视频场景。为实现高质量视频，[157] 首次将扩散模型应用于文本引导的视频生成，刷新了评估基准。此后，Meta 和谷歌分别提出了基于扩散模型的 Make-a-Video [394] 和 Imagen Video [155] 。具体而言，Make-a-Video 将基于扩散模型的文本引导图像生成模型扩展到视频生成，加快了生成速度，并消除了训练时对配对文本-视频数据的需求。然而，Make-a-Video 需要大规模文本-视频数据集进行微调，因此会消耗大量计算资源。最新的 Tune-a-Video [474] 提出了单次视频生成，由文本引导和图像输入驱动，并使用单个文本-视频对训练开放领域生成器。

6.2 三维生成

深度生成模型在二维图像上的巨大成功，促使研究人员开始探索三维数据生成；这实际上是对真实物理世界进行建模。与格式单一的二维数据不同，三维物体可以通过深度图像、体素网格[476],点云[330, 331],网格[140]和神经场[283],来表示，而每种表示方式都有其优缺点。

根据输入和引导类型的不同，三维物体可以由文本、图像和三维数据生成。尽管已有多种方法 [112, 175, 262] 探索了由语义标签或语言描述引导的形状编辑，但由于缺乏三维数据和合适的架构，三维生成仍然面临挑战。基于扩散模型，DreamFusion [326] 提出利用预训练文生二维模型来解决这些问题。另一类研究则从单视图图像 [33, 122, 243, 432, 457, 510] 或多视图图像 [63, 167, 454, 480],重建三维物体，这类方法称为图像到三维。多视图三维重建的一个新方向是神经辐射场（NeRF） [286] ，它用于对三维信息进行隐式表示。三维到三维任务包括从部分三维数据 [455]中进行补全以及进行变换[26], ，其中三维物体检索是具有代表性的变换任务。

6.3 语音

语音合成是语音处理领域的重要研究方向，旨在使机器能够将文本转化为自然、易懂的语音。传统的语音合成方法包括发音器官合成、 [217, 380], 共振峰合成 [12,377], 拼接合成 [293, 306],以及统计参数语音合成（统计参数语音合成） [198, 292]。这些方法已得到广泛研究和应用，例如，共振峰合成仍用于开源软件 NVDA（Windows 平台领先的免费屏幕阅读器之一）。然而，这些合成语音与人声仍可区分，而且语音合成中的瑕疵会降低可懂度。

早期工作 [102, 333, 514–516] 由三个模块组成：文本分析、声学模型和声码器。WaveNet [440]是语音合成领域的一项突破，能够根据语言学特征生成原始波形。为提高稿件质量，提交至美国计算机协会

语音质量和声音多样性，生成模型被引入语音合成领域，例如生成对抗网络 [124]。与生成对抗网络相比，扩散模型不需要判别器，因此训练更加稳定、简单。因此，语音合成领域的研究采用扩散模型，这已成为一种新兴趋势。一类工作 [57, 168, 220, 479]专注于高效语音合成，通过加速推理等方式缩短生成时间，例如在训练中结合调度网络和评分网络，以及联合训练生成对抗网络。另一类研究 [52, 289, 361, 390] 专注于端到端模型，这类模型无需任何中间表示，便可直接根据文本生成波形。完全端到端的模型不仅简化了训练和推理过程，也降低了对人工标注的需求。基于扩散模型的语音合成研究并不局限于上述两个方向，还包括语音增强和引导式语音合成等。

6.4 图

图在世界中无处不在，有助于可视化和定义各个领域中物体之间的关系，范围涵盖社交网络到化合物。图生成旨在从与现有图相似的训练分布中创建新图，因此备受关注。

传统的图生成方法 [11, 224, 464] 会创建具有特定特征的新图，这些特征与真实图的手工设计统计特征相关。此举简化了流程，但无法捕捉复杂场景中的关系结构。随着深度学习算法取得成功，研究人员开始将其应用于图生成。与传统方法不同，这些算法可以直接使用真实数据进行训练，并自动提取特征。其中，基于自回归模型的工作 [76, 249, 503] 以逐步方式依次创建图结构，这种方法具有更好的可扩展性，但无法对置换不变性进行建模，且计算开销很大。与此同时，单步模型 [254, 269, 269]（如变分自编码器和流模型）为了确保似然计算可处理，无法准确地对结构信息建模。尽管基于生成对抗网络的图生成方法 [80, 183, 278]通过使用判别器绕开了基于似然的优化，但训练过程并不稳定。

近期，开发用于图结构数据的扩散模型受到越来越多的关注。EDP-GNN [302]率先展示了扩散模型在图生成中的能力，旨在解决非不变属性问题。此后，一方面，基于扩散模型的研究 [138, 166, 186, 266, 445] 专注于真实图生成，即生成与给定图集相似的图；另一方面， [14, 388, 482, 513]则专注于目标导向图生成，即生成能够优化给定对象的图，例如用于分子和材料生成。

6.5 其他

此外，还有一些有趣的任务涉及生成不同模态的内容，例如音乐生成 [179] 和唇读 [106]。典型的音乐生成系统可按三种表示层级（自上而下）进行分类，分别生成乐谱、演奏和音频 [179]。随着深度学习的发展，音乐生成领域引入了各种方法以提升音乐质量，例如 [170] 中的 MusicVAE [354], 、MuseGAN [95] 和 Transformer。音乐生成推动并加速了计算机辅助作曲软件的发展，包括谷歌的 Magenta 项目和索尼计算机科学研究所的 Flow Machine 项目。唇读任务将唇部运动的视觉输入转换为解码后的语音 [106],；得益于语料库和架构的改进，该领域也取得了显著进展。

7 行业应用

毫无疑问，自 2022年以来，人工智能生成内容已在社交媒体上风靡一时。例如，用户积极分享使用 ChatGPT 进行互动对话，或使用 Stable Diffusion 根据文本提示生成图像的体验。然而，如果人工智能生成内容无法用于行业中的实际应用以展现其价值，这股热潮预计将逐渐消退。因此，我们将探讨人工智能生成内容可能对各行业产生的影响。

7.1 教育

人工智能生成内容正通过协助教学与学习，改变教育范式。生成式人工智能在教学中蕴含变革潜力，应用范围涵盖课程材料生成以及评估与评价 [324,517]。与此同时，生成模型的应用已开始影响学生的学习方式 [27, 420]。

生成式人工智能技术能够帮助教育工作者开展个性化辅导 [517], 、设计课程材料 [324], 以及进行评估与评价 [27, 517]。面向幼儿、采用 ChatGPT 等生成式技术的独特外语教学产品，能够吸引孩子的注意力、激发他们的学习动机，并营造有趣的学习环境。高等教育需要接纳人工智能在其中的应用，从而为学生创造更具吸引力、更有效且更高效的学习体验 [517]。生成式人工智能课程材料生成的一项主要优势是，它能够自动完成课程材料的创建与更新，节省教师的时间和精力。此外，ChatGPT 可以显著减轻法学院教师的工作负担，从而腾出时间提高学术生产力，或培养更复杂的教学技能 [324]。ChatGPT 在促进教学方面的益处包括但不限于推动个性化互动学习。然而，ChatGPT 也可能存在一些局限，例如生成错误信息、加剧训练数据中已有的偏见以及引发隐私问题 [27]。总的来说，应对这些挑战需要政策制定者和教育工作者共同努力，为生成式人工智能工具的适当使用提供建议或指导。

此外，生成式人工智能技术可以帮助学生撰写论文 [420], 居家测试或测验 [420], 理解某些理论和概念，以及处理学术问题中的不同语言论文和文章 [27, 517]。聊天机器人能够全天候为学生提供支持，让他们在需要时获得所需帮助。借助纠正语法、提出改进建议和找出薄弱环节的能力，ChatGPT 等聊天机器人可以即时反馈学生的写作，帮助他们从错误中学习，并随着时间推移提升写作技能。这不仅能节省学生的时间，还能帮助他们成为更优秀的写作者 [493]。一项由在线课程提供商开展的调查显示，89%的学生使用 ChatGPT 完成家庭作业，其中50%用它写论文，48%用它进行居家测试或测验 [420]。此外，生成式人工智能可以根据学生的个人需求（例如学习风格和进度）定制课程材料，从而有望提升学生参与度和学习成果。ChatGPT 还能帮助学生理解某些理论、概念以及不同语言的文章，让他们更有效地完成学习任务 [27, 517]。生成式人工智能课程材料生成也伴随着挑战和顾虑，包括生成材料的质量，以及用于训练人工智能的数据可能存在偏见。因此，在教育情境中使用生成的课程材料之前，必须对其进行仔细评估和验证 [79]。

借助上述应用场景，人工智能生成内容有望通过提升教育内容的质量和可及性、提高学生参与度和留存率，并为学习者提供个性化支持，彻底变革教育。随着人工智能的不断进步，人工智能生成内容有望成为教育行业不可或缺的一部分，为学生带来更具吸引力、可及性更高且更加个性化的学习体验。

7.2 游戏与元宇宙

在游戏和元宇宙中，大多数用户可能不会认同千篇一律的内容，因为个性化才能带来最佳体验。尽管游戏和元宇宙为用户提供了虚拟世界，但其中的内容体现了用户的角色和个性。生成式人工智能让这一切成为可能，不仅能让用户自定义自己的虚拟化身，还能提供丰富多样的场景和故事情节，让体验更加身临其境 [53, 325, 344, 344]。

由 GPT-3 模型驱动的 AI Dungeon 允许用户生成由文本引导的开放式故事，生成式人工智能会根据用户的不同操作生成新的事件，从而带来独一无二且出乎意料的游戏体验 [422]。Horizon Worlds 是最受欢迎的游戏之一，玩家可以在其中漫游于与内容消费和创作相关的虚拟世界。在 Horizon 中，用户可以更好地掌控如何定制在线体验，以满足个人需求。具体而言，用户可以使用包含预设物体和虚拟化身属性的小工具，设计独特的虚拟化身和场景 [284]。此外，视觉小说游戏 Traveler 专注于生成绚丽的场景，为用户带来视觉冲击，让玩家踏上穿越多元世界的旅程。玩家探索 Traveler 时，将置身于壮丽的视觉景象和沉浸式声音景观之中。每个场景都独一无二，内容可以从幽暗的森林延伸至繁华的城市，全部由生成式人工智能精心打造 [113]。

尽管“元宇宙”一词近来已成为热门话题，但在现实世界中，游戏打造的虚拟空间或许能成为通往元宇宙的入口 [301, 344]。沙盒游戏平台 Roblox 首次在招股说明书中纳入“元宇宙”概念，使其市值大幅飙升；在这个平台上，玩家可以创造超乎想象的专属世界 [301]。虚拟演唱会歌手拥有更丰富多样的音乐风格和才华。观众可以选择自己喜欢的风格，甚至挑选偶像，从而获得更加多元、个性化的演唱会体验。知名美国说唱歌手兼制作人 Travis Scott 曾在《堡垒之夜》游戏中举办一场具有历史意义的演唱会，他的虚拟化身带领玩家体验了从水下到外太空的不同场景 [465]。加利福尼亚大学伯克利分校在热门电脑游戏《Minecraft》中举行了毕业典礼。在《Minecraft》游戏中，学生和校友运用生成式人工智能技术建造了校园的复制版，让数千名毕业生能够使用自己的虚拟化身，从世界各地参加这场活动 [358]。总的来说，人工智能在游戏和元宇宙的发展演变中发挥了重要作用；随着技术进步、普及程度不断提高，其应用也在持续增长。

7.3 媒体

随着生成式人工智能技术的普及发展，其在媒体和广告领域发挥着日益重要的作用。人工智能生成内容不仅促进了媒体多样性，为受众带来更好的体验，还使媒体从业者能够更高效地完成工作 [44, 84, 107, 222, 288, 439, 527]。

由人工智能生成内容赋能的媒体能够提供更加多样化的内容和报道方式，改变媒体的生产方式和组织结构 [204, 273, 527]。人工智能生成内容可应用于媒体领域的各种场景，例如写作机器人、新闻主播和图像描述生成。传统上，媒体机构依赖专业新闻记者撰写新闻文章和报道，这需要投入大量精力和时间，因此文章数量有限。此外，新闻的时效性至关重要，新闻可能在一小时后就失去关注度。生成式人工智能可以利用文本生成技术，大幅助力新闻业提升效率和响应能力 [288]。美联社应用这些技术，每年生成约 40000 篇报道，其公司财报文章数量从 1200 篇增加到 14800 [130]。《洛杉矶时报》新闻部的机器人记者 Quakebot 只需三分钟

在洛杉矶地震后完成一篇相关报道 [309]。国际金融媒体公司彭博新闻于2018年推出了Buttetin，旨在提供具有个人叙事风格的内容，其一句话摘要由聊天机器人生成 [470]。

生成式人工智能与媒体的深度融合催生了人工智能新闻主播 [461, 491?]。人工智能新闻主播与真人主播相结合，使信息传播方式更加多样化。人工智能新闻主播可以根据文本播报新闻，其外貌和表情都模仿真人主播。中国国家通讯社新华社和中国搜索引擎搜狗开发了具有不同形象和语言能力的人工智能新闻主播。其中最令人印象深刻的是三维人工智能新闻主播邢晓薇，其播报形式可以从不同角度进行全方位呈现，显著增强了立体感和层次感 [279]。此外，韩国有线频道MBN创建了人工智能新闻主播AIKim，能够快速应对各种突发事件，甚至全天候进行报道 [397]。为了帮助听力障碍人士获取更多有关国际体育赛事和北京冬奥会的信息，科技巨头腾讯开发了人工智能驱动的手语服务提供商灵语，其中使用了包括三维数字人建模、机器翻译、图像生成和语音转文本在内的AIGC任务 [511]。此外，专注于提供数字信息的中国商业公司咪咕可以为北京冬奥会的直播提供智能字幕功能。因此，听力障碍人士能够观看体育赛事的直播，从而获得更加沉浸式的体验。

7.4 广告

各种人工智能应用改变了广告行业，为广告商提供了强大的工具，帮助他们创作出创新且引人入胜的内容，以更深层次与消费者建立联系 [84, 133, 439]。在人工智能于广告领域的各种应用中，人工智能生成内容尤其具有影响力，因为它使广告商能够创作出个性化且富有吸引力的内容，与每位消费者产生共鸣。创意广告系统（CAS）遵循利用人工智能生成和测试广告创意的原则，帮助有抱负的创作者和成熟的创作者认识到，创造力并非精英专属，而是一个可以借助数据和计算辅助完成的系统化过程 [439]。程序化广告的实施并未充分利用自生成技术，导致不同消费者接触到相同的内容。幸运的是，个性化广告文案智能生成系统（SGS-PAC）能够自动定制广告内容，以满足每位消费者的需求 [84]。广告海报是用于推广产品的一种常见信息展示形式。另一种智能系统 Vinci 支持自动生成广告海报 [133]。用户输入指定的产品图像和标语后，Vinci 会利用深度生成模型生成精美的海报。

此外，Brandmark.io 是一款基于人工智能生成内容的工具，可自动为企业生成标志。该工具会根据用户的偏好和要求生成多种标志方案。广告商可以购买并使用该工具为其企业创建的标志，使其成为一种简单且经济高效的标志设计解决方案 [193]。通过生成对抗网络强制输出包含特定关键词，这种方法可自动生成可能吸引潜在买家的商品列表 [275]。它能提升用户在点对点交易市场上的营销效果。此外，技术创新为广告行业提供了数字化和自动化工具，也让广告商能够自动制作“合成广告”。据 [381]，报道，人工智能生成内容通过帮助广告商大规模创作高度个性化且引人入胜的内容，同时节省时间和资源，改变了广告行业。我们预计，人工智能生成内容在广告领域还将出现更多创新且影响深远的应用。

7.5 电影

技术如今几乎影响着电影创作的每一个环节，这一点很有意思。研究促成了计算机环境的发展，助力剪辑、标注、视频检索等诸多工作 [78,125, 295, 317–319, 366, 415, 428, 434]。首先，人工智能编剧软件对电影制作流程产生了重大影响。人工智能通过整合视觉特效 (VFX) [291], 、改进音效，并引入新的观影平台，创造了全新的电影体验。4K、IMAX 和 3D电影以及动画都深受其影响。

剧本奠定了一部电影的票房表现。人工智能生成设备会存储并处理海量数据，以创作“理想”的剧本 [15]。人工智能软件还被用于将旧电影剧本改写成润色后的版本，再由导演和编剧进行分析和改进。它的用途不止于创作和分析现有剧本。Jasper AI [362] 和 Scalenut [329]就是人工智能编剧的两个例子。

电影会运用视觉特效（VFX）来增强观众吸引力。视觉特效将原始图像与实拍视频结合，营造引人入胜、逼真且贴合情境的画面，其中可能包括数字环境、数字减龄等效果。《本杰明·巴顿奇事》 [416] 的视觉特效团队在一间明亮的房间里架设了两组摄像机，并利用 MOVA Contour现实捕捉技术 [82] 构建了一个三维数据库，记录主角的面部表情。随后，视觉特效团队为主角在不同年龄阶段制作了高分辨率三维模型，最后运用人工智能处理从三维数据库中提取的数据，使头部模型呈现衰老效果。最终效果十分逼真，赢得了电影界的广泛赞誉。在《银翼杀手》 2049 [274] 和《双子杀手》 [360], 等电影中，视觉特效团队借助动作捕捉技术 [5] 追踪并记录了主角的面部数据。随后，他们在计算机中重建三维面部数据，并进一步润饰，以实现数字减龄。尽管这种方法需要昂贵的技术和大量资金，但其精度极高，适应性也很强。

人工智能通过让已故演员重现于电影中，拓展了娱乐的边界。深度伪造技术利用计算机视觉和人工智能，制作出令人惊叹、栩栩如生的真实人物或虚构角色伪造视频。《速度与激情》 7 [469] 利用视觉特效和旧影像，在保罗·沃克于2013年去世后让他重现银幕，并将他的面容移植到他的兄弟身上。

除了视觉特效，字幕也在观众体验中发挥着重要作用。为了帮助有听力障碍的观众，自动生成的听障字幕（也称为 SDH） [8], 会包含语音转写文本、说话人变化和背景噪声的文字描述。这些优势显著提升了电影的票房收入。自然语言处理（NLP）是一种专注于解读口语的人工智能，为电影提供多语言字幕。Rev [351], 是一款云端程序，深受电影爱好者使用。人工智能也彻底改变了无声电影中的语音预测任务。人工智能生成的语音合成可以为无声电影配上旁白，也能为电影制作多种语言的配音。利用大量人类音频样本训练的深度学习系统可以生成自然的旁白配音。线性预测编码（LPC） [101] 和梅尔频谱图 [100] 等特征通过转换生成高质量、清晰易懂的音频。近期， [327]提出了一种用于视频到音频合成的 Tacotron2 模型 [384]变体。 [494] 提出了一种高效的随机模型，可为特定的无声视频生成无穷无尽的高质量音频模式，从而有效涵盖语音预测问题的多模态性。除了视觉和音效，我们还可以使用 Colourlab.Ai [70] 进行调色、使用 Descript [86] 进行视频编辑，还有许多其他工具也在电影行业持续掀起热潮。

7.6 音乐

人工智能生成内容也进入了音乐产业，并取得了显著进展 [499]。人工智能不仅能够发现庞大数据集中人类难以察觉的图案和趋势，还能为业余音乐家提供前沿技术，助力他们提升创作过程，这无疑是一个绝佳的机会。许多专家、研究人员、音乐家和唱片公司都在探索将人工智能技术融入音乐这一新趋势 [280]。许多人利用人工智能生成内容创作全新的音乐，也有一些软件能够按照不同作曲家的风格编辑作品。

音乐产业预计将在这一领域投入巨额资金，无论是利用人工智能作曲，还是帮助音乐人。谷歌的 Magenta 项目 [172], [10]便是一个出色的人工智能旋律生成器案例。IBM 的 Watson Beat 则是另一个例子。它利用人工智能和机器学习来创作原创歌曲 [49], [110]。2016年，DeepMind 的研究人员成功制作了文本转语音（TTS）录音，以及听起来像音乐的录音 [421], [496]。人工智能也广泛用于数字音频的处理和改进。LANDR 是一项卓越的人工智能创意工具服务，能让音乐人将自己的音乐发布到 Spotify 和 Apple Music 等多个流媒体服务平台。作词人经常会遇到一个被称为“创作瓶颈 [379]”的重大难题。不过，多亏了人工智能，这如今已不再是问题。如今，许多音乐人都借助人工智能为歌曲创作新歌词 [277]。OpenAI 这家人工智能技术公司开发了 GPT-2 [7], [538], ，一种文本生成工具。这款出色的文本生成器不仅能生成真实可信的新闻，还能为披头士乐队的歌曲以及其他所有流派的音乐创作歌词。然而，人工智能并非只能生成文本；它还能够创作原创配乐和旋律。索尼 CSL 流动机器能为艺术家提供帮助，让他们根据自己的想法创作原创音乐 [356]。AIVA 是最知名的原创音乐创作人工智能工具之一 [189]。用户先选择一种预设风格，再调整调性、乐器编制、拍号等各种参数，从而创作出独特的曲目。AIVA 在读取巴赫、莫扎特等著名音乐家的数百部作品时，能够将其中复杂的听觉信息分解为独立特征 [98]。随后，这些特征可以重新加以诠释，并用于创作全新的音乐作品。除上述工具外，还有许多其他应用对音乐产业产生了重大影响，例如基于 iOS 的工具 Amadeus Code [38], 、云端平台 Amper [147], 、Ecrett Music [36], 等。

7.7 绘画

人工智能生成内容通过提供自动绘画工具、鼓励创意实验等多种方式，正在革新绘画产业。人工智能程序可以分析图片，生成可用于创作艺术作品的配色方案、图案和纹理。利用这些算法生成的自动绘图工具能够将这些图案和纹理应用于作品中，创作出独特而精巧的艺术作品 [548]。人工智能还可以分析个人的偏好、兴趣和风格，创作定制化的艺术作品。赋予艺术家根据个人偏好和兴趣创作艺术作品的能力，可以提升作品的吸引力和价值。

MidJourney 创作的艺术作品《太空歌剧剧院》在科罗拉多州博览会艺术竞赛中荣获第一名 [81], ，展现了人工智能绘画工具创作优秀艺术作品的能力。Midjourney 是一款功能全面的优秀 AI图像生成器，许多艺术家都用它来激发灵感。生成式人工智能创作的各种艺术形式，例如抽象画生成 [235], 、中国山水画 [543], 和中国水墨画 [64], ，无疑推动了绘画的发展。此外，人工智能生成内容还能协助保护与修复 [506], [164]。人工智能算法能够分析并修复受损的

艺术作品。这些算法通过检测并清除灰尘、划痕和其他瑕疵，让文物修复师更轻松地将艺术作品恢复至原始状态。

对于不熟悉绘画或动画的非专业人士而言，人工智能生成内容也很有帮助，因为它能让他们创作出高质量的视觉特效。通过为扩散模型添加额外的约束，ControlNet [341] 可以提高生成图像的多样性。它可以描述生成的图像，并结合边界图、深度信息、霍夫线图、法线图和姿态估计等其他约束。人工智能生成内容也开启了协作艺术创作的新时代 [41, 50]。人工智能算法可以创作由多位艺术家共同参与的协作绘画作品。这些算法能够分析每位艺术家的风格，并生成融合所有艺术家作品元素的统一风格。

7.8 代码开发

生成式人工智能能够为代码开发领域作出贡献 [134, 135, 176, 412], ，人工智能生成内容无需手动编码即可生成代码。 [412] 的研究探讨了代码生成式人工智能的可解释性要求，并展示了以人为中心的方法如何推动新领域中可解释人工智能（XAI）技术的发展。为了提高测试效率并增加测试覆盖率，自动生成高质量的测试用例尤为重要 [134, 135]。为了优化数据工程的效率，可以设计一种基于神经网络、用于数据集增强的新型软件工程方法 [176]。人工智能生成代码的一项热门应用是 GitHub 与 OpenAI 联合开发的人工智能工具 GitHub Copilot。用户可以使用软件开发工具，通过 GitHub Copilot 自动补全代码 [91]。此外，人工智能生成技术还可以协助进行代码重构，在不改变原有功能的前提下改进现有代码。这可以缩短开发者重构代码所需的时间，并提高代码质量。DeepCode 是一款热门的代码重构工具 [174], ，它是一款由人工智能支持的代码审查工具，可以检查代码并提供改进建议。此外，人工智能生成内容也会对电子商务和金融行业产生影响 [163]。亚马逊、京东等电子商务平台可以利用人工智能客服为客户提供购物指南服务，从而为企业节省成本。金融公司可以使用虚拟投资顾问，为客户提供证券开户、金融投资及其他相关服务方面的建议。

7.9 手机应用与功能

许多人工智能生成内容应用以趣味性移动应用的形式涌现，通常用于图像和视频编辑。Photoshop 一直是常见的图像编辑工具，但手动操作耗时，而且可能产生不自然或不真实的效果。此外，视频编辑需要分析每个视频片段，并根据音频和视觉内容作出编辑决策。由于视频是一种基于时间的双轨媒介，需要仔细考量每一帧，因此这一过程十分耗时。幸运的是，一些研究 [17, 32, 398] 探索了人工智能生成内容背后的人工智能技术在图像或视频编辑中的应用，使人脸替换和数字化身等人工智能生成内容应用成为可能。

一些基于人脸替换的热门应用正在互联网上广受欢迎。这项技术利用先进的人工智能技术分析人脸，并在几秒钟内将用户的面孔替换成他们喜爱的名人或其他任何人，因此与传统的 PS 技术相比，使用起来更加轻松快捷。VanceAI、Voila AI Artist 和 FaceAPP 都是其中的佼佼者。FaceApp 被公认为最佳面部照片编辑应用，屡获殊荣，下载量已超过 5 亿次，且仍在持续增长 [372]。另一种热门应用是变声技术。这项技术可以调整人声的音高、音色、语速和其他特征，从而

改变人声的音质。MagicMic [471] 和 Voicemod [202] 是两款热门应用，可用于实时变声和音效板操作。人们可以用它们改变自己的声音，以创作有趣的内容、进行直播或满足其他用途，从而增进交流的乐趣。

此外，另一项技术趋势是将个人转变为虚拟角色，从而提升娱乐性。虚拟角色是虚拟世界中人物的数字化身，既可以是现实人物的部分复制，也可以是完全数字化的版本。苹果的第一代“数字化身”技术 Animoji 主要侧重于生成预设的卡通和动物角色，不支持自定义生成 [426]。以 iPhone 的 Memoji 和小米的 Mimoji 为代表的第二代“数字化身”技术开始支持个性化虚拟化身定制，提供多种选项，包括发型、眼睛、鼻子、服装等 [406]。这项升级让用户能够创建不仅可以追踪面部动作，而且看起来也像自己的虚拟化身。此外，用户还可以将创建的虚拟化身作为评论发布在微信或 Facebook 聊天中，从而在社交媒体上以更个性化的方式表达自己。此后，数字化身技术已成为各类智能手机厂商推出的智能手机标准功能之一。

7.10 其他领域

除了上述领域，人工智能生成内容有望在更多领域得到应用。例如，新型药物的设计与开发过程复杂、成本高且耗时。平均而言，一种新药需要花费约 30 亿美元、历时 10 多年才能获准上市 [201]。因此，人们希望利用人工智能生成内容加速药物发现进程并降低成本。2018年，DeepMind 创建了 AlphaFold [364], ，它能够准确预测蛋白质的结构，被视为药物发现和基础生物学研究领域的里程碑。其升级版本 AlphaFold2 于2020年发布，准确度高于前代。Justas Dauparas 设计的 ProteinMPNN[77], 能够针对特定任务设计蛋白质序列，在短短几秒内快速生成全新的蛋白质。除了直接利用生成的内容，人工智能生成内容还可以帮助各领域的从业者提高效率。例如，在医疗咨询中，患者可以依靠聊天机器人获取基本的医疗建议，只有遇到较严重的情况时才求助医生。在制造设计中，可以将人工智能生成内容与广泛使用的计算机辅助设计系统相结合，减少重复性工作，让设计师专注于更有意义的部分。

8 挑战与展望

8.1 挑战

尽管人工智能生成内容在各个领域生成逼真且多样化的输出方面取得了显著成功，但在实际应用中仍面临诸多挑战。除了需要大量的训练数据和计算资源外，我们还列出以下几项最重要的挑战。

(1) **缺乏可解释性**。 尽管人工智能生成内容模型能够生成令人印象深刻的结果，但人们仍难以理解模型是如何得出这些结果的。当模型生成不理想的结果时，这尤其令人担忧。缺乏可解释性使得输出结果难以控制。(2) **伦理与法律问题**。 人工智能生成内容模型容易受到数据偏见的影响。例如，主要使用英文文本训练的语言模型可能会偏向西方文化。侵犯版权和侵犯隐私是不可忽视的潜在法律问题。此外，人工智能生成内容模型也可能被恶意使用。例如，学生可以利用这些工具在论文作业中作弊，因此需要人工智能内容检测器。提交给美国计算机协会的论文稿

人工智能生成内容模型还可用于传播政治竞选活动中的误导性内容。

(3) 特定领域的技术挑战。 目前及在不久的将来，不同领域都需要各自独特的人工智能生成内容模型。每个领域仍面临其特有的挑战。例如，广受欢迎的文生图人工智能生成内容工具 Stable Diffusion 偶尔会生成与用户期望相去甚远的结果，例如将人类画成动物、把一个人画成两个人等。另一方面，聊天机器人也偶尔会出现事实性错误。

8.2 展望

尽管生成式人工智能的人气前所未有，但它仍处于早期阶段。下面，我们将介绍人工智能生成内容在不久的将来可能如何发展。

(1) 更灵活的控制。 人工智能生成内容任务的一大趋势是实现更灵活的控制。以图像生成为例，早期基于生成对抗网络的模型能够生成高质量图像，但可控性较差。近期在大量文本图像数据上训练的扩散模型能够通过文本指令进行控制，从而生成更符合用户需求的图像。尽管如此，当前的文生图模型仍需要更精细的控制，以便更灵活地生成图像

(2) 从预训练到微调。 目前，ChatGPT 等人工智能生成内容模型的发展重点在于预训练阶段。相关技术已相对成熟；然而，如何针对下游任务微调这些基础模型，仍是一个尚未得到充分探索的领域。与从头开始训练模型不同，微调的目标是在基础模型原有的通用能力与其在新任务上的适应性能之间取得平衡。

(3) 从大型科技公司到初创公司。 目前，人工智能生成内容技术主要由谷歌和 Meta 等大型科技公司推动发展。在大型科技公司的支持下，一些初创公司也已崭露头角，展现出巨大的潜力，例如 OpenAI（由微软支持）和 DeepMind（由谷歌支持）。随着关注重点从核心技术开发转向应用，由于需求不断增长，预计将有更多初创公司涌现。

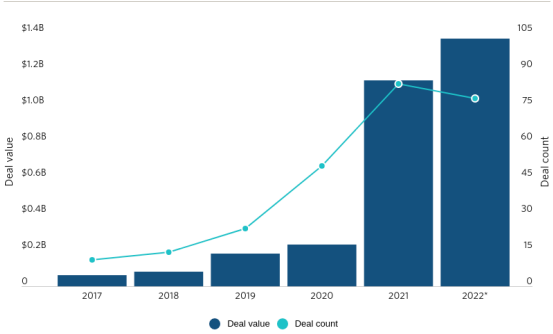


图 21. 过去 6 年生成式人工智能融资的交易金额和交易数量（图表来源：PitchBook）。

讨论：投资、泡沫与就业机会。 从技术角度看，过去几年里，人工智能生成内容取得了显著进展，这一点毋庸置疑。变革性技术出现时，市场往往会 Manuscript submittedtoACM

对其潜在应用和未来增长抱有过度乐观的态度，生成式人工智能也不例外。根据 PitchBook（见图 21），过去两年，风险投资（VC）对生成式人工智能的投资大幅增加。一些批评者担心，生成式人工智能可能成为下一个泡沫。他们主要担忧的是，大多数人工智能生成内容工具主要是供人娱乐，而非具有实用价值。例如，文生图模型玩起来很有趣，但它们如何创造收入仍不明朗。生成式人工智能可能会如何演进，很难预测。不过，本文作者认为，考虑到生成式人工智能仍是一个相对新兴、发展迅速且拥有众多潜在应用的领域，它不太可能成为下一个泡沫。关于生成式人工智能是否会取代人类、导致大量就业机会流失，也存在激烈争论。另一方面，生成式人工智能也能为具备人工智能研究和实施技能的人创造新的就业机会。受益于人工智能生成内容强大能力的行业也可能蓬勃发展，从而创造更多就业机会。

参考文献

[1] Rameen Abdal、Yipeng Qin 和 Peter Wonka。2019年。Image2stylegan：如何将图像嵌入 StyleGAN 潜在空间？载于《IEEE/CVF国际计算机视觉会议论文集》。4432–4441。[2] Dinesh Acharya、Zhiwu Huang、Danda Pani Paudel 和 Luc Van Gool。2018年。通过渐进式增长的切片 Wasserstein 生成对抗网络实现高分辨率视频生成。arXiv 预印本 arXiv:1810.02419（2018年）。[3] Johannes Ackermann 和 Minjun Li。2022年。通过多阶段混合扩散进行高分辨率图像编辑。arXiv 预印本 arXiv:2210.12965（2022年）。[4] Daniel Adiwardana、Minh-Thang Luong、David R So、Jamie Hall、Noah Fiedel、Romal Thoppilan、Zi Yang、Apoorv Kulshreshtha、Gaurav Nemade、Yifeng Lu 等。2020年。迈向类人开放域聊天机器人。arXiv 预印本 arXiv:2001.09977（2020年）。[5] Adobe。2022年。什么是动作捕捉？它是如何运作的？ https://www.adobe.com/uk/creativecloud/animation/discover/motion-capture.html（2022年）。[6] Namhyuk Ahn、Byungkon Kang 和 Kyung-Ah Sohn。2017年。使用卷积神经网络进行图像失真检测。载于《2017年第4届IAPR亚洲模式识别会议（ACPR）》。电气电子工程师学会，220–225。[7] Open AI。2019年。GPT-2 博客。https://openai.com/blog/tags/gpt-2/（2019年）。[8] ai media。2017年。聋人或听障人士字幕（SDH）——字幕、隐藏式字幕和 SDH。https://blog.ai-media.tv/blog/what-is-sdh（2017年）。[9] Reina Akama、Sho Yokoi、Jun Suzuki 和 Kentaro Inui。2020年。通过连通性和内容相关性筛选含噪对话语料库。arXiv 预印本 arXiv:2004.14008（2020年）。[10] Mincer Alaeddine 和 Anthony Tannoury。2021年。音乐创作中的人工智能。载于《人工智能应用与创新：第17届 IFIP WG 12.5 国际会议，AIAI 2021，希腊克里特岛赫索尼索斯，2021年6月25日至27日，论文集17》。施普林格，387–397。[11] Réka Albert 和 Albert-László Barabási。2002年。复杂网络的统计力学。《现代物理评论》74卷1期（2002年），47。[12] Jonathan Allen、Sharon Hunnicutt、Rolf Carlson 和 Bjorn Granstrom。1979年。MITalk-79：1979年 MIT 文本转语音系统。《美国声学学会杂志》65卷，S1期（1979年），S130–S130。[13] Jaan Altosaar。2016年。教程——什么是变分自编码器？ https://doi.org/10.5281/zenodo.4462916[14] Namrata Anand 和 Tudor Achim。2022年。使用等变去噪扩散概率模型生成蛋白质结构与序列。arXiv 预印本 arXiv:2205.15019（2022年）。[15] Nantheera Anantrasirichai 和 David Bull。2022年。创意产业中的人工智能：综述。《人工智能评论》（2022年），1–68。[16] Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、Stephen Gould 和 Lei Zhang。2018年。用于图像描述和视觉问答的自底向上与自顶向下注意力。载于《IEEE计算机视觉与模式识别会议论文集》。6077–6086。[17] Dawit Mureja Argaw、Fabian Caba Heilbron、Joon-Young Lee、Markus Woodson 和 In So Kweon。2022年。视频编辑剖析：一个面向人工智能辅助视频编辑的数据集和基准套件。载于《计算机视觉——ECCV 2022：第17届欧洲计算机视觉会议，以色列特拉维夫，2022年10月23日至27日，论文集，第VIII部分》。施普林格，201–218。[18] Martin Arjovsky、Amar Shah 和 Yoshua Bengio。2016年。酉演化循环神经网络。载于《国际机器学习会议》。PMLR，1120–1128。[19] Omri Avrahami、Ohad Fried 和 Dani Lischinski。2022年。混合潜在扩散。arXiv 预印本 arXiv:2206.02779（2022年）。[20] Omri Avrahami、Thomas Hayes、Oran Gafni、Sonal Gupta、Yaniv Taigman、Devi Parikh、Dani Lischinski、Ohad Fried 和 Xi Yin。2022年。SpaText：用于可控图像生成的空间文本表示。arXiv 预印本 arXiv:2211.14305（2022年）。[21] Omri Avrahami、Dani Lischinski 和 Ohad Fried。2022年。用于自然图像文本驱动编辑的混合扩散。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。18208–18218。[22] Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E. Hinton。2016年。层归一化。arXiv 预印本 arXiv:1607.06450（2016年）。

[23] Alexei Baevski、Michael Auli 和 Abdelrahman Mohamed。2019年。自监督预训练对语音识别的有效性。arXiv 预印本 arXiv:1911.03912（2019年）。[24] Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。2014年。通过联合学习对齐与翻译实现神经机器翻译。arXiv 预印本 arXiv:1409.0473（2014年）。[25] Ashutosh Baheti、Alan Ritter 和 Kevin Small。2020年。用于对话式问答的流畅响应生成。arXiv 预印本 arXiv:2005.10464（2020年）。[26] Song Bai、Xiang Bai、Wenyu Liu 和 Fabio Roli。2015年。用于三维模型检索的神经形状编码。《模式识别快报》65卷（2015年），15–21。[27] David Baidoo-Anu 和 Leticia Owusu Ansah。2023年。生成式人工智能（AI）时代的教育：了解 ChatGPT 在促进教学与学习方面的潜在益处。可在 SSRN 4337484 获取（2023年）。[28] Dana H Ballard。1987年。神经网络中的模块化学习。载于《AAAI》，第647卷。279–284。[29] Tadas Baltrušaitis、Chaitanya Ahuja 和 Louis-Philippe Morency。2018年。多模态机器学习：综述与分类。《IEEE模式分析与机器智能汇刊》41卷2期（2018年），423–443。[30] Hangbo Bao、Li Dong 和 Furu Wei。2022年。BEiT：图像Transformer的 BERT 预训练。国际学习表征会议（2022年）。[31] Siqi Bao、Huang He、Fan Wang、Rongzhong Lian 和 Hua Wu。2019年。增进彼此了解：通过综合评估演进对话策略。arXiv 预印本 arXiv:1906.00549（2019年）。[32] Omer Bar-Tal、Dolev Ofri-Amar、Rafail Fridman、Yoni Kasten 和 Tali Dekel。2022年。Text2live：文本驱动的分层图像与视频编辑。载于《计算机视觉——ECCV 2022：第17届欧洲计算机视觉会议，以色列特拉维夫，2022年10月23日至27日，论文集，第XV部分》。施普林格，707–723。[33] Jan Bednarik、Pascal Fua 和 Mathieu Salzmann。2018年。从单一视图中学习重建无纹理可变形表面。载于《2018年国际三维视觉会议（3DV）》。电气电子工程师学会，606–615。[34] Yoshua Bengio、Réjean Ducharme 和 Pascal Vincent。2000年。一种神经概率语言模型。《神经信息处理系统进展》第13卷（2000年）。[35] Mohamed Benzeghiba、Renato De Mori、Olivier Deroo、Stephane Dupont、Teodora Erbes、Denis Jouvet、Luciano Fissore、Pietro Laface、Alfred Mertins、Christophe Ris 等。2007年。自动语音识别与语音变异性：综述。《语音通信》49卷10–11期（2007年），763–786。[36] Sean Berry。2019年。Ecrett Music 使用人工智能为你的视频生成免版税音乐。https://www.videomaker.com/news/ecrett-music-uses-ai-to- generate-royalty-free-music-for-your-videos/（2019年）。[37] Andreas Blattmann、Robin Rombach、Kaan Oktay 和 Björn Ommer。2022年。检索增强扩散模型。arXiv 预印本 arXiv:2204.11824（2022年）。[38] Amadeus Code Blog。2019年。人工智能驱动的歌曲创作助手。https://blog.amadeuscode.com/（2019年）。[39] Denny Britz、Anna Goldie、Thang Luong 和 Quoc Le。2017年。大规模探索神经机器翻译架构。ArXiv 电子预印本（2017年3月）。arXiv:1703.03906 [cs.CL][40] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。《神经信息处理系统进展》（2020年）。[41] Melissa G Bublitz、Tracy Rank-Christman、Luca Cian、Xavier Cortada、Adriana Madzharov、Vanessa M Patrick、Laura A Peracchio、Maura L Scott、Aparna Sundar、Ngoc To 等。2019年。协作艺术：社区中的变革力量。《消费者研究协会学报》4卷4期（2019年），313–331。[42] Weiwei Cai 和 Zhanguo Wei。2020年。PiiGAN：用于多样化图像修复的生成对抗网络。《IEEE Access》8卷（2020年），48451–48463。[43] Guendalina Caldarini、Sardar Jaf 和 Kenneth McGarry。2022年。聊天机器人最新进展的文献综述。《信息》13卷1期（2022年），41。[44] Colin Campbell、Kirk Plangger、Sean Sands 和 Jan Kietzmann。2022年。为深度伪造和人工智能生成广告时代做好准备：理解操纵性广告反应的框架。《广告学报》51卷1期（2022年），22–38。[45] Hanqun Cao、Cheng Tan、Zhangyang Gao、Guangyong Chen、Pheng-Ann Heng 和 Stan Z Li。2022年。生成式扩散模型综述。arXiv 预印本 arXiv:2209.02646（2022年）。[46] Nicholas Carlini 和 David Wagner。2018年。音频对抗样本：针对语音转文本的定向攻击。载于《2018年IEEE安全与隐私研讨会（SPW）》。电气电子工程师学会，1–7。[47] Eric R Chan、Connor Z Lin、Matthew A Chan、Koki Nagano、Boxiao Pan、Shalini De Mello、Orazio Gallo、Leonidas J Guibas、Jonathan Tremblay、Sameh Khamis 等。2022年。高效的几何感知三维生成对抗网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。16123–16133。[48] Paramanand Chandramouli 和 Kanchana Vaishnavi Gandikota。2022年。LDEdit：借助潜在扩散模型实现通用文本引导图像操控。arXiv 预印本 arXiv:2210.02249（2022年）。[49] Anna Chaney。2018年。The Watson Beat：利用机器学习激发音乐创造力。https://medium.com/@anna_seg/the-watson-beat- d7497406a202（2018年）。[50] Rong Chang、Xinmiao Song 和 Huiwen Liu。2022年。山水画与风景画之间：连接中国与西方绘画的人工智能美学研究。载于《HCI International 2022 海报集：第24届人机交互国际会议，HCII 2022，线上活动，2022年6月26日至7月1日，论文集，第III部分》。施普林格，179–185。[51] Lele Chen、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu。2019年。采用动态逐像素损失的分层跨模态说话人脸生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。7832–7841。稿件已提交至美国计算机协会

[52] Nanxin Chen、Yu Zhang、Heiga Zen、Ron J Weiss、Mohammad Norouzi、Najim Dehak 和 William Chan。2021年。WaveGrad 2：用于文本转语音合成的迭代细化。arXiv预印本 arXiv:2106.09660（2021年）。[53] Shu-Ching Chen。2022年。面向元宇宙的多媒体研究。IEEE MultiMedia 29, 1（2022年），125–127。[54] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。2020年。用于视觉表征对比学习的简单框架。载于 ICML。[55] Xinlei Chen 和 Kaiming He。2021年。探索简单的孪生表征学习。载于 CVPR。[56] Xinpeng Chen、Lin Ma、Wenhao Jiang、Jian Yao 和 Wei Liu。2018年。通过用现在重构过去来正则化用于图像描述生成的循环神经网络。载于 IEEE计算机视觉与模式识别会议论文集。7995–8003。[57] Zehua Chen、Xu Tan、Ke Wang、Shifeng Pan、Danilo Mandic、Lei He 和 Sheng Zhao。2022年。Infergrad：在训练中考虑推理，以改进用于声码器的扩散模型。载于 ICASSP 2022-2022 IEEE声学、语音与信号处理国际会议（ICASSP）。电气电子工程师学会，8432–8436。[58] Yong Cheng。2019年。基于枢轴语言的神经机器翻译联合训练。载于《神经机器翻译的联合训练》。施普林格，41–54。[59] Zezhou Cheng、Qingxiong Yang 和 Bin Sheng。2015年。深度着色。载于 IEEE国际计算机视觉会议论文集。415–423。[60] Anton Cherepkov、Andrey Voynov 和 Artem Babenko。2021年。探索用于语义图像编辑的生成对抗网络参数空间。载于 IEEE/CVF计算机视觉与模式识别会议论文集。3671–3680。[61] Chung-Cheng Chiu、Tara N Sainath、Yonghui Wu、Rohit Prabhavalkar、Patrick Nguyen、Zhifeng Chen、Anjuli Kannan、Ron J Weiss、Kanishka Rao、Ekaterina Gonina 等。2018年。基于序列到序列模型的先进语音识别。载于 2018年 IEEE声学、语音与信号处理国际会议（ICASSP）。电气电子工程师学会，4774–4778。[62] Kyunghyun Cho、Bart Van Merriënboer、Dzmitry Bahdanau 和 Yoshua Bengio。2014 年。神经机器翻译的特性：编码器 - 解码器方法。arXiv 预印本 arXiv:1409.1259 （ 2014 年）。[63] Christopher B Choy、Danfei Xu、JunYoung Gwak、Kevin Chen 和 Silvio Savarese。2016年。3D-R2N2：用于单视图和多视图三维物体重建的统一方法。载于欧洲计算机视觉会议。施普林格，628–644。[64] Chieh-Yu Chung 和 Szu-Hao Huang。2022年。利用边界增强生成对抗网络，将中国水墨画交互式转换为写实图像。Multimedia Tools and Applications（2022年），1–34。[65] Junyoung Chung、Caglar Gulcehre、KyungHyun Cho 和 Yoshua Bengio。2014年。门控循环神经网络在序列建模中的实证评估。arXiv预印本 arXiv:1412.3555（2014年）。[66] Joon Son Chung、Amir Jamaludin 和 Andrew Zisserman。2017年。你是这么说的吗？arXiv预印本 arXiv:1705.02966（2017年）。[67] Aidan Clark、Jeff Donahue 和 Karen Simonyan。2019年。在复杂数据集上进行对抗式视频生成。arXiv预印本 arXiv:1907.06571（2019年）。[68] Aidan Clark、Jeff Donahue 和 Karen Simonyan。2019年。在复杂数据集上高效生成视频。（2019年）。[69] Ronan Collobert、Christian Puhrsch 和 Gabriel Synnaeve。2016年。Wav2letter：一种基于卷积网络的端到端语音识别系统。arXiv预印本 arXiv:1609.03193（2016年）。[70] @ColourlabAI。2023年。为那些“这是我做的”时刻注入动力。https://colourlab.ai/（2023年）。[71] Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli。2020年。面向语音识别的无监督跨语言表征学习。arXiv预印本 arXiv:2006.13979（2020年）。[72] Guillaume Couairon、Jakob Verbeek、Holger Schwenk 和 Matthieu Cord。2022年。DiffEdit：利用掩码引导进行基于扩散的语义图像编辑。arXiv预印本 arXiv:2210.11427（2022年）。[73] Florinel-Alin Croitoru、Vlad Hondru、Radu Tudor Ionescu 和 Mubarak Shah。2022年。视觉领域的扩散模型：综述。arXiv预印本 arXiv:2209.04747（2022年）。[74] Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael J Black。2019年。三维说话风格的捕捉、学习与合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10101–10111。[75] George E Dahl、Dong Yu、Li Deng 和 Alex Acero。2011年。用于大词汇量语音识别的上下文相关预训练深度神经网络。IEEE音频、语音与语言处理汇刊 20, 1（2011年），30–42。[76] Hanjun Dai、Azade Nazi、Yujia Li、Bo Dai 和 Dale Schuurmans。2020年。面向稀疏图的可扩展深度生成建模。载于国际机器学习会议。PMLR，2302–2312。[77] Justas Dauparas、Ivan Anishchenko、Nathaniel Bennett、Hua Bai、Robert J Ragotte、Lukas F Milles、Basile IM Wicky、Alexis Courbet、Rob J de Haas、Neville Bethel 等。2022年。利用 ProteinMPNN 进行基于稳健深度学习的蛋白质序列设计。《科学》378, 6615（2022年），49–56。[78] Marc Davis。1994年。媒体流：用于检索和再利用的视频表示。载于第二届 ACM国际多媒体会议论文集。478–479。[79] Marcelo Zerwes Dawn Gilmore、Anitra Nottingham。2023年。ChatGPT 与学习设计：它能为在线内容创作提供哪些机会？https://www.csmonitor.com/Technology/2023/0217/Tremendous-potential-Why-some-disability-advocates-laud-ChatGPT（2023年）。[80] Nicola De Cao 和 Thomas Kipf。2018年。MolGAN：用于小型分子图的隐式生成模型。arXiv预印本 arXiv:1805.11973（2018年）。[81] Paul DelSignore。2022年。人工智能艺术赢得比赛并引发争议。https://medium.com/mlearning-ai/ai-art-wins-fine-arts-competition-and- sparks-controversy-882f9b4df98c（2022年）。[82] Rick DeMott。2006年。Mova Contour 将“动作捕捉”转变为“现实捕捉”。https://www.awn.com/news/mova-contour-moves-motion-capture- reality-capture（2006年）。稿件已提交至美国计算机协会

[83] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。2009年。ImageNet：大规模层次化图像数据库。载于 CVPR。[84] Shasha Deng、Chee-Wee Tan、Weijun Wang 和 Yu Pan。2019年。个性化广告文案智能生成系统及其在广告实践与研究中的应用。《广告学报》48, 4（2019年），356–365。[85] Jan Deriu、Alvaro Rodrigo、Arantxa Otegi、Guillermo Echegoyen、Sophie Rosset、Eneko Agirre 和 Mark Cieliebak。2021年。对话系统评估方法综述。《人工智能评论》54, 1（2021年），755–810。[86] @descript。2022年。一种制作视频和播客的新方法。一种好方法。https://www.descript.com/（2022年）。[87] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：面向语言理解的深度双向变换器预训练。arXiv预印本 arXiv:1810.04805（2018年）。[88] Ming Ding、Zhuoyi Yang、Wenyi Hong、Wendi Zheng、Chang Zhou、Da Yin、Junyang Lin、Xu Zou、Zhou Shao、Hongxia Yang 等。2021年。CogView：通过 Transformer 掌握文本到图像生成。神经信息处理系统进展 34（2021年），19822–19835。[89] Ming Ding、Wendi Zheng、Wenyi Hong 和唐杰。2022年。CogView2：通过层次化变换器实现更快速、更优质的文本到图像生成。arXiv预印本 arXiv:2204.14217（2022年）。[90] Laurent Dinh、David Krueger 和 Yoshua Bengio。2015年。NICE：非线性独立分量估计。国际学习表征会议2015年研讨会论文集（2015年）。[91] Thomas Dohmke。2022年。GitHub Copilot 已面向所有开发者全面开放。https://github.blog/2022-06-21-github-copilot-is-generally-available- to-all-developers/（2022年）。[92] Jeffrey Donahue、Lisa Anne Hendricks、Sergio Guadarrama、Marcus Rohrbach、Subhashini Venugopalan、Kate Saenko 和 Trevor Darrell。2015年。用于视觉识别与描述的长时循环卷积网络。载于 IEEE计算机视觉与模式识别会议论文集。2625–2634。[93] Chao Dong、Chen Change Loy、Kaiming He 和 Xiaoou Tang。2014年。学习用于图像超分辨率的深度卷积网络。载于欧洲计算机视觉会议。施普林格，184–199。[94] Chao Dong、Chen Change Loy、Kaiming He 和 Xiaoou Tang。2015年。利用深度卷积网络实现图像超分辨率。IEEE模式分析与机器智能汇刊 38, 2（2015年），295–307。[95] Hao-Wen Dong、Wen-Yi Hsiao、Li-Chia Yang 和 Yi-Hsuan Yang。2018年。MuseGAN：用于符号音乐生成和伴奏的多轨序列生成对抗网络。载于 AAAI人工智能会议论文集，第32卷。[96] Gianfranco Doretto、Alessandro Chiuso、Ying Nian Wu 和 Stefano Soatto。2003年。动态纹理。《国际计算机视觉杂志》51, 2（2003年），91–109。[97] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2020年。一张图像胜过16×16个词：用于大规模图像识别的 Transformer。arXiv预印本 arXiv:2010.11929（2020年）。[98] Eric Drott。2021年。音乐人工智能产业中的版权、补偿与共享资源。《创意产业期刊》14, 2（2021年），190–207。[99] Henry Elder、Alexander O'Connor 和 Jennifer Foster。2020年。如何让面向任务的对话中的神经自然语言生成像模板一样可靠。载于2020年自然语言处理实证方法会议（EMNLP）论文集。2877–2888。[100] Ariel Ephrat、Tavi Halperin 和 Shmuel Peleg。2017年。改进从无声视频中重建语音。载于 IEEE国际计算机视觉会议研讨会论文集。455–462。[101] Ariel Ephrat 和 Shmuel Peleg。2017年。Vid2Speech：从无声视频中重建语音。载于2017年 IEEE声学、语音与信号处理国际会议（ICASSP）。电气电子工程师学会，5095–5099。[102] Yuchen Fan、Yao Qian、Feng-Long Xie 和 Frank K Soong。2014年。基于双向长短期记忆网络的文本转语音合成。载于国际语音通信协会第十五届年会。[103] Jinsheng Fang、Hanjiang Lin、Xinyu Chen 和 Kun Zeng。2022年。用于轻量级图像超分辨率的卷积神经网络与 Transformer 混合网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集。1103–1112。[104] Jean Louis K E Fendji、Diane CM Tala、Blaise O Yenke 和 Marcellin Atemkeng。2022年。使用有限词汇量的自动语音识别：综述。《应用人工智能》36, 1（2022年），2095039。[105] Shaoxiong Feng、Xuancheng Ren、Hongshen Chen、Bin Sun、Kan Li 和 Xu Sun。2020年。通过模仿隐含场景来正则化对话生成。arXiv预印本 arXiv:2010.01893（2020年）。[106] Souheil Fenghour、Daqing Chen、Kun Guo、Bo Li 和 Perry Xiao。2021年。基于深度学习的自动唇读：综述。IEEE Access（2021年）。[107] Feyyaz Fırat。2019年。机器人新闻。《国际新闻学百科全书》（2019年），1–5。[108] Mikel L Forcada、Mireia Ginestí-Rosell、Jacob Nordfalk、Jim O'Regan、Sergio Ortiz-Rojas、Juan Antonio Pérez-Ortiz、Felipe Sánchez-Martínez、Gema Ramírez-Sánchez 和 Francis M Tyers。2011年。Apertium：用于基于规则的机器翻译的自由开源平台。《机器翻译》25, 2（2011年），127–144。[109] Danny Francis 和 Benoit Huet。2021年。利用深度架构生成图像和视频描述。《多面向深度学习：模型与数据》（2021年），151–174。[110] Emma Frid、Celso Gomes 和 Zeyu Jin。2020年。通过示例进行音乐创作。载于2020年计算机系统中的人为因素 CHI会议论文集。1–13。[111] Ohad Fried、Ayush Tewari、Michael Zollhöfer、Adam Finkelstein、Eli Shechtman、Dan B Goldman、Kyle Genova、Zeyu Jin、Christian Theobalt 和 Maneesh Agrawala。2019年。对人物讲话视频进行基于文本的编辑。《ACM图形学汇刊》（TOG）38, 4（2019年），1–14。

[112] Rao Fu、詹晓、陈一文、Daniel Ritchie 和 Srinath Sridhar。2022年。Shapecrafter：一种递归式文本条件三维形状生成模型。arXiv 预印本 arXiv:2207.09446（2022年）。[113] Yocat Games。2023年。Traveler-The AI Story。https://skidrowcracked.com/traveler-the-ai-story/（2023年）。[114] Zhe Gan、Ricardo Henao、David Carlson 和 Lawrence Carin。2015年。通过数据增强学习深度 sigmoid 信念网络。载于《人工智能与统计学》。PMLR，268–276。[115] Zhe Gan、李林杰、Chunyuan Li、Lijuan Wang、刘子成、Jianfeng Gao 等。2022年。视觉语言预训练：基础、最新进展与未来趋势。《计算机图形学与视觉基础与趋势》14卷3–4期（2022年），163–352。[116] Ruiqi Gao、卢阳、周俊培、朱松纯和 Ying Nian Wu。2018年。通过多网格建模与采样学习生成式卷积网络。9155–9164。[117] Xiang Gao、Yizhe Zhang、Michel Galley、Chris Brockett 和 Bill Dolan。2020年。利用大规模人类反馈数据训练对话回复排序模型。arXiv 预印本 arXiv:2009.06978（2020年）。[118] Leon A Gatys、Alexander S Ecker 和 Matthias Bethge。2015年。艺术风格神经算法。arXiv 预印本 arXiv:1508.06576（2015年）。[119] Leon A Gatys、Alexander S Ecker 和 Matthias Bethge。2016年。使用卷积神经网络进行图像风格迁移。载于计算机视觉与模式识别会议。[120] Jonas Gehring、Michael Auli、David Grangier、Denis Yarats 和 Yann N Dauphin。2017年。卷积序列到序列学习。载于国际机器学习会议。PMLR，1243–1252。[121] Spyros Gidaris、Praveer Singh 和 Nikos Komodakis。2018年。通过预测图像旋转进行无监督表征学习。国际学习表征会议（2018年）。[122] Vladislav Golyanik、Soshi Shimada、Kiran Varanasi 和 Didier Stricker。2018年。Hdm-net：利用学习型变形模型进行单目非刚性三维重建。载于国际虚拟现实与增强现实会议。施普林格，51–72。[123] RC Gonzalez。2006年。Woods。RE，（2002）《数字图像处理》。[124] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、徐冰、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。2014年。生成对抗网络。载于 NeurIPS。[125] Andrew S Gordon 和 Eric A Domeshek。1995年。用于视频检索的概念索引。载于 IJCAI 智能多媒体信息检索研讨会工作论文集，蒙特利尔。23–38。[126] Alex Graves。2013年。使用循环神经网络生成序列。arXiv 预印本 arXiv:1308.0850（2013年）。[127] Alex Graves、Abdel-rahman Mohamed 和 Geoffrey Hinton。2013年。使用深度循环神经网络进行语音识别。载于2013年 IEEE 国际声学、语音与信号处理会议。电气电子工程师学会，6645–6649。[128] Ulf Grenander 和 Michael I Miller。1994年。复杂系统中的知识表示。《英国皇家统计学会会刊：B辑》（方法论）56卷4期（1994年），549–581。[129] Jean-Bastien Grill、Florian Strub、Florent Altché、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Doersch、Bernardo Avila Pires、郭昭涵、Mohammad Gheshlaghi Azar 等。2020年。自举式潜变量学习：一种自监督学习新方法。《神经信息处理系统进展》（2020年）。[130] Barbara Gruber。2022年。《事实、虚假信息与数据：人工智能如何影响新闻业》。https://www.goethe.de/prj/k40/en/lan/aij.html（2022年）。[131] 管善言、戴莹、倪冰冰、朱飞达、Feiyue Huang 和杨小康。2020年。通过协同学习加速 StyleGAN 嵌入。arXiv 预印本 arXiv:2007.01758（2020年）。[132] 郭龙腾、刘静、朱欣欣、姚鹏、卢世晨和卢汉清。2020年。用于图像描述的归一化几何感知自注意力网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10327–10336。[133] 郭淑楠、金卓宸、孙福玲、李静雯、李昭睿、施洋和曹楠。2021年。Vinci：用于生成广告海报的智能平面设计系统。载于2021年计算机系统中的人为因素会议论文集。1–17。[134] 郭秀静。2021年。探索利用生成对抗网络实现自动化软件测试。载于2021年第51届 IEEE/IFIP 国际可靠系统与网络会议增刊（DSN-S）。IEEE，21–22。[135] 郭秀静、Hiroyuki Okamura 和 Tadashi Dohi。2022年。利用生成对抗网络自动生成软件测试数据。IEEE Access 10（2022年），20690–20700。[136] Tanmay Gupta、Dustin Schwenk、Ali Farhadi、Derek Hoiem 和 Aniruddha Kembhavi。2018年。想象一下！从剧本到构图再到视频。载于欧洲计算机视觉会议论文集。598–613。[137] Michael Gutmann 和 Aapo Hyvärinen。2010年。噪声对比估计：一种用于非归一化统计模型的新估计原理。载于人工智能与统计学会议。[138] Kilian Konstantin Haefeli、Karolis Martinkus、Nathanaël Perraudin 和 Roger Wattenhofer。2022年。离散状态空间有益于图扩散模型。arXiv 预印本 arXiv:2210.01549（2022年）。[139] Donghoon Ham、Jeong-Gwan Lee、Youngsoo Jang 和 Kee-Eung Kim。2020年。使用 GPT-2 的目标导向对话系统端到端神经网络流水线。载于计算语言学协会第58届年会论文集。583–592。[140] Rana Hanocka、Amir Hertz、Noa Fish、Raja Giryes、Shachar Fleishman 和 Daniel Cohen-Or。2019年。MeshCNN：一种以边为基础的网络。《ACM图形学汇刊》（TOG）38卷4期（2019年），1–12。[141] Kaiming He、Xinlei Chen、谢赛宁、李扬昊、Piotr Dollár 和 Ross Girshick。2022年。掩码自编码器是可扩展的视觉学习器。载于计算机视觉与模式识别会议。[142] Kaiming He、范浩奇、吴育昕、谢赛宁和 Ross Girshick。2020年。用于无监督视觉表征学习的动量对比。载于计算机视觉与模式识别会议。提交给美国计算机协会的稿件

[143] Kaiming He、张祥雨、任少卿和 Jian Sun。2016年。用于图像识别的深度残差学习。载于计算机视觉与模式识别会议。[144] 何瑞飞、孙书扬、余鑫、薛楚辉、张文清、Philip Torr、Song Bai 和 Xiaojuan Qi。2022年。生成模型生成的合成数据是否已适用于图像识别? arXiv 预印本 arXiv:2210.07574（2022年）。[145] 何万伟、杨敏、Rui Yan、李成明、沈颖和徐瑞峰。2020年。通过对抗训练融合两位教师的知识，以构建面向任务的对话系统。载于2020年自然语言处理实证方法会议（EMNLP）论文集。3498–3507。[146] 何振良、左旺孟、阚美娜、山世光和陈熙霖。2019年。AttGAN：只改变所需内容的人脸属性编辑。《IEEE图像处理汇刊》28卷11期（2019年），5464–5478。[147] Brian Heater。2022年。Amper 正提供即插即用的解决方案，助力制造业数字化。https://techcrunch.com/2022/04/22/amper-is-providing-a- plug-and-play-solution-to-digitize-manufacturing/（2022年）。[148] Matthew Henderson、Ivan Vulić、Daniela Gerz、Iñigo Casanueva、Paweł Budzianowski、Sam Coope、Georgios Spithourakis、Tsung-Hsien Wen、Nikola Mrkšić 和 Pei-Hao Su。2019年。为任务型对话系统训练神经响应选择模型。arXiv 预印本 arXiv:1906.01543（2019年）。[149] Simao Herdade、Armin Kappeler、Kofi Boakye 和 Joao Soares。2019年。图像描述：将物体转化为文字。《神经信息处理系统进展》32（2019年）。[150] Amir Hertz、Ron Mokady、Jay Tenenbaum、Kfir Aberman、Yael Pritch 和 Daniel Cohen-Or。2022年。利用交叉注意力控制实现提示词到提示词图像编辑。arXiv 预印本 arXiv:2208.01626（2022年）。[151] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。2017年。通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。《神经信息处理系统进展》30（2017年）。[152] Geoffrey Hinton、Li Deng、Dong Yu、George E Dahl、Abdel-rahman Mohamed、Navdeep Jaitly、Andrew Senior、Vincent Vanhoucke、Patrick Nguyen、Tara N Sainath 等。2012年。用于语音识别声学建模的深度神经网络：四个研究团队的共同观点。《IEEE信号处理杂志》29卷6期（2012年），82–97。[153] Geoffrey E Hinton。2002年。通过最小化对比散度训练专家乘积模型。《神经计算》14卷8期（2002年），1771–1800。[154] Geoffrey E Hinton 和 Richard Zemel。1993年。自编码器、最小描述长度与亥姆霍兹自由能。NeurIPS（1993年）。[155] Jonathan Ho、William Chan、Chitwan Saharia、Jay Whang、Ruiqi Gao、Alexey Gritsenko、Diederik P Kingma、Ben Poole、Mohammad Norouzi、David J Fleet 等。2022年。Imagen Video：利用扩散模型生成高清视频。arXiv 预印本 arXiv:2210.02303（2022年）。[156] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020年。去噪扩散概率模型。《神经信息处理系统进展》33（2020年），6840–6851。[157] Jonathan Ho、Tim Salimans、Alexey Gritsenko、William Chan、Mohammad Norouzi 和 David J Fleet。2022年。视频扩散模型。arXiv 预印本 arXiv:2204.03458（2022年）。[158] Sepp Hochreiter 和 Jürgen Schmidhuber。1997年。长短期记忆网络。《神经计算》9卷8期（1997年），1735–1780。[159] Margaret Holland。2016年。YouTube 如何发展成为成功的用户生成内容平台。《埃隆大学传播学本科生研究期刊》7卷1期（2016年）。[160] Wenyi Hong、Ming Ding、Wendi Zheng、刘兴翰和唐杰。2022年。CogVideo：利用 Transformer 进行大规模文本生成视频预训练。arXiv 预印本 arXiv:2205.15868（2022年）。[161] Takaaki Hori、Jaejin Cho 和 Shinji Watanabe。2018年。使用基于词的循环神经网络语言模型进行端到端语音识别。载于2018年 IEEE口语语言技术研讨会（SLT）。电气电子工程师学会，389–396。[162] Ehsan Hosseini-Asl、Bryan McCann、Chien-Sheng Wu、Semih Yavuz 和 Richard Socher。2020年。一种简单的面向任务的对话语言模型。《神经信息处理系统进展》33（2020年），20179–20191。[163] Stephanie Houde、Vera Liao、Jacquelyn Martino、Michael Muller、DavidPiorkowski、John Richards、Justin Weisz 和张云峰。2020年。生成式人工智能在商业中的（不当）用例。arXiv 预印本 arXiv:2003.07679（2020年）。[164] 胡振江。2022年。人工智能技术辅助环境保护对中国画完整性的影响分析。《环境与公共卫生杂志》2022年（2022年）。[165] 黄高、刘壮、Laurens Van Der Maaten 和 Kilian Q. Weinberger。2017年。密集连接卷积网络。载于计算机视觉与模式识别会议。[166] 黄瀚、孙磊磊、杜博文、傅彦杰和吕伟锋。2022年。GraphGDP：用于排列不变图生成的生成式扩散过程。arXiv 预印本 arXiv:2212.01842（2022年）。[167] Po-Han Huang、Kevin Matzen、Johannes Kopf、Narendra Ahuja 和 Jia-Bin Huang。2018年。DeepMVS：学习多视图立体视觉。载于 IEEE计算机视觉与模式识别会议论文集。2821–2830。[168] 黄荣杰、赵洲、刘华岱、刘景林、崔晨烨和任毅。2022年。ProDiff：用于高质量文本转语音的渐进式快速扩散模型。arXiv 预印本 arXiv:2207.06389（2022年）。[169] 黄心婷、齐建中、孙宇和 Rui Zhang。2020年。通过随机奖励估计进行半监督对话策略学习。arXiv 预印本 arXiv:2005.04379（2020年）。[170] 黄郁翔和 Yi-Hsuan Yang。2020年。流行音乐 Transformer：基于节拍的富有表现力的流行钢琴作品建模与生成。载于第28届 ACM国际多媒体会议论文集。1180–1188。[171] William John Hutchins。1986年。机器翻译：过去、现在与未来。Ellis Horwood，奇切斯特。[172] Matthew Hutson。2017年。谷歌如何利用人工智能创作音乐。（2017年）。[173] Nitin Indurkhya 和 Fred J Damerau。2010年。《自然语言处理手册》。Chapman and Hall/CRC。

[174]Prathamesh Ingle. 2023年. 可帮助程序员生成代码的顶尖人工智能（AI）工具。
https://www.marktechpost.com/2023/01/01/top-artificial-intelligence-ai-tools-that-can-generate-code-to-help-programmers/（
2023年）。[175] Tansin Jahan、Yanran Guan 和 Oliver van Kaick. 2021年. 面向形状生成的语义引导潜在空间探索。载于《计算机图形论坛》第40卷。Wiley Online Library，115–126。[176] Benjamin Jahić、Nicolas Guelfi 和 Benoit Ries. 2019年. 使用生成对抗网络进行数据集增强的软件工程。载于2019年IEEE第十届国际软件工程与服务科学会议（ICSESS）。电气电子工程学会，59–66。[177] Viren Jain 和 Sebastian Seung. 2008年. 使用卷积网络对自然图像进行去噪。神经信息处理系统进展21（2008年）。[178] Amir Jamaludin、Joon Son Chung 和 Andrew Zisserman. 2019年. 你是这么说的吗？：从音频合成说话人脸。《国际计算机视觉杂志》127（2019年），1767–1779。[179] Shulei Ji、Jing Luo 和 Xinyu Yang. 2020年. 深度音乐生成综合综述：多层级表示、算法、评估与未来方向。arXiv预印本 arXiv:2011.06801（2020年）。[180] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig. 2021年. 利用噪声文本监督扩大视觉及视觉语言表示学习的规模。载于ICML。[181] Qi Jia、Yizhu Liu、Siyu Ren、Kenny Q Zhu 和 Haifeng Tang. 2020年. 使用对话依存关系进行多轮响应选择。arXiv预印本 arXiv:2010.01502（2020年）。[182] Wenxiang Jiao、Wenxuan Wang、Jen-tse Huang、Xing Wang 和 Zhaopeng Tu. 2023年. ChatGPT是优秀的翻译器吗？一项初步研究。arXiv预印本 arXiv:2301.08745（2023年）。[183] Wengong Jin、Kevin Yang、Regina Barzilay 和 Tommi Jaakkola. 2018年. 学习用于分子优化的多模态图到图转换。arXiv预印本 arXiv:1812.01070（2018年）。[184] Li Jing、Caglar Gulcehre、John Peurifoy、Yichen Shen、Max Tegmark、Marin Soljacic 和 Yoshua Bengio. 2019年. 门控正交循环单元：关于遗忘的学习。《神经计算》31卷4期（2019年），765–783。[185] Yongcheng Jing、Yezhou Yang、Zunlei Feng、Jingwen Ye、Yizhou Yu 和 Mingli Song. 2019年. 神经风格迁移：综述。《IEEE可视化与计算机图形学汇刊》26卷11期（2019年），3365–3385。[186] Jaehyeong Jo、Seul Lee 和 Sung Ju Hwang. 2022年. 通过随机微分方程组进行基于分数的图生成建模。arXiv预印本 arXiv:2202.02514（2022年）。[187] Melvin Johnson、Mike Schuster、Quoc V Le、Maxim Krikun、Yonghui Wu、Zhifeng Chen、Nikhil Thorat、Fernanda Viégas、Martin Wattenberg、Greg Corrado 等. 2017年. 谷歌的多语言神经机器翻译系统：实现零样本翻译。《计算语言学协会汇刊》5（2017年），339–351。[188] Biing Hwang Juang 和 Laurence R Rabiner. 1991年. 用于语音识别的隐马尔可夫模型。《技术计量学》33卷3期（1991年），251–272。[189] Romain Juillet. 2021年. 人工智能如何改变音乐产业。https://www.bocasay.com/ai-transforming-music-industry/（
2021年）。[190] Łukasz Kaiser 和 Samy Bengio. 2016年. 主动记忆能取代注意力吗？神经信息处理系统进展29（2016年）。[191] Nal Kalchbrenner 和 Phil Blunsom. 2013年. 循环连续翻译模型。载于《2013年自然语言处理实证方法会议论文集》，1700–1709。[192] Nal Kalchbrenner、Lasse Espeholt、Karen Simonyan、Aaron van den Oord、Alex Graves 和 Koray Kavukcuoglu. 2016年. 线性时间内的神经机器翻译。arXiv预印本 arXiv:1610.10099（2016年）。[193] Shitanshu kapadia. 2023年. 面向企业的7种人工智能营销工具。https://www.entrepreneur.com/science-technology/how-will-chatgpt-change-education- and-teaching/445018（2023年）。[194] S Karpagavalli 和 Edy Chandra. 2016年. 自动语音识别架构与方法综述。《国际信号处理、图像处理与模式识别期刊》9卷4期（2016年），393–404。[195] Andrej Karpathy 和 Li Fei-Fei. 2015年. 用于生成图像描述的深度视觉语义对齐。载于IEEE计算机视觉与模式识别会议论文集，3128–3137。[196] Tero Karras、Timo Aila、Samuli Laine、Antti Herva 和 Jaakko Lehtinen. 2017年. 通过联合端到端学习姿态与情感，实现音频驱动的人脸动画。《ACM图形学汇刊》（TOG）36卷4期（2017年），1–12。[197] Tero Karras、Samuli Laine 和 Timo Aila. 2019年. 用于生成对抗网络的基于风格的生成器架构，4401–4410。[198] Hideki Kawahara. 2006年. STRAIGHT：发掘声码器的另一面——语音的感知同构分解。《声学科学与技术》27卷6期（2006年），349–353。[199] Bahjat Kawar、Michael Elad、Stefano Ermon 和 Jiaming Song. 2022年. 去噪扩散复原模型。arXiv预印本 arXiv:2201.11793（2022年）。[200] Lei Ke、Wenjie Pei、Ruiyu Li、Xiaoyong Shen 和 Yu-Wing Tai. 2019年. 用于图像描述的反思解码网络。载于IEEE/CVF国际计算机视觉会议论文集，8888–8897。[201] Regina Kelder 和 Mary Parker. 2021年. 药物开发为何耗时如此之久?
https://www.criver.com/eureka/why-does-drug-development-take-so-long（2021年）。[202] Tilly Kenyon. 2022年. Voicemod：助创作者找到自己的声音。https://technologymagazine.com/ai-and-machine-learning/voicemod-allowing-creators-to-find-their-voice（2022年）。

[203] Angella J Kim 和 Kim KP Johnson. 2016年. 消费者运用社交媒体的力量：考察品牌相关用户生成内容对Facebook的影响。《人类行为中的计算机》58（2016年），98–108。[204] Daewon Kim 和 Seongcheol Kim. 2017年. 报业公司采用机器人新闻的决定因素。《技术预测与社会变迁》117（2017年），184–195。[205] Gwanghyun Kim 和 Se Young Chun. 2022年. DATID-3D：利用文本生成图像扩散技术实现多样性保留的领域自适应，用于三维生成模型。arXiv预印本 arXiv:2211.16374（2022年）。[206] Gwanghyun Kim 和 Jong Chul Ye. 2021年. Diffusionclip：利用扩散模型进行文本引导的图像编辑。（2021年）。[207] Sijin Kim、Namhyuk Ahn 和 Kyung-Ah Sohn. 2020年. 使用专家混合网络复原空间异质失真。载于亚洲计算机视觉会议论文集。[208] Sungdong Kim、Sohee Yang、Gyuwan Kim 和 Sang-Woo Lee. 2019年. 通过选择性覆盖记忆实现高效的对话状态跟踪。arXiv预印本 arXiv:1911.03906（2019年）。[209] Taeksoo Kim、Byoungjip Kim、Moonsu Cha 和 Jiwon Kim. 2017年. 利用可重构生成对抗网络进行无监督视觉属性迁移。arXiv预印本 arXiv:1707.09798（2017年）。[210] Diederik P Kingma 和 Max Welling. 2013年. 自动编码变分贝叶斯。arXiv预印本 arXiv:1312.6114（2013年）。[211] Wei-Jen Ko、Avik Ray、Yilin Shen 和 Hongxia Jin. 2020年. 从语义潜空间生成对话回复。arXiv预印本 arXiv:2010.01658 （ 2020 年）。[212] Philipp Koehn 、 Hieu Hoang 、 Alexandra Birch 、 Chris Callison-Burch 、 Marcello Federico、Nicola Bertoldi、Brooke Cowan、Wade Shen、Christine Moran、Richard Zens 等. 2007年. Moses：用于统计机器翻译的开源工具包。载于计算语言学协会第45届年会论文集（演示与海报环节配套卷），177–180。[213] Philipp Koehn、Franz J Och 和 Daniel Marcu. 2003年. 统计短语翻译。技术报告。南加州大学玛丽娜德尔雷伊信息科学研究所。[214] Marijn Koolen、Jaap Kamps、Gabriella Kazai 等. 2013年. 社交图书搜索：专业内容和用户生成内容对图书推荐的影响。载于DIR，38–39。[215] Prajwal KR、Rudrabha Mukhopadhyay、Jerin Philip、Abhishek Jha、Vinay Namboodiri 和 CV Jawahar. 2019年. 迈向自动面对面翻译。载于ACM第27届国际多媒体会议论文集，1428–1436。[216] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E. Hinton. 2012年. 使用深度卷积神经网络进行ImageNet分类。载于NeurIPS。[217] BJ Kröger. 1992年. 发音语音合成的最小规则。EUSIPCO92会议论文集（1）（1992年），331–334。[218] Manoj Kumar、Dirk Weissenborn 和 Nal Kalchbrenner. 2021年. 着色变换器。arXiv预印本 arXiv:2102.04432（2021年）。[219] Rithesh Kumar、Jose Sotelo、Kundan Kumar、Alexandre de Brébisson 和 Yoshua Bengio. 2017年. Obamanet：从文本生成照片级真实感唇形同步。arXiv预印本 arXiv:1801.01442（2017年）。[220] Max WY Lam、Jun Wang、Dan Su 和 Dong Yu. 2022年. BDDM：用于快速、高质量语音合成的双边去噪扩散模型。arXiv 预印本 arXiv:2203.13508 （ 2022 年）。[221] Hugo Larochelle 和 Iain Murray. 2011年. 神经自回归分布估计器。载于第十四届人工智能与统计学国际会议论文集。JMLR研讨会与会议论文集，29–37。[222] Noam Lemelshtrich Latar. 2018年. 机器人新闻：人类新闻业能否存续？世界科学出版社。[223] Christian Ledig、Lucas Theis、Ferenc Huszár、Jose Caballero、Andrew Cunningham、Alejandro Acosta、Andrew Aitken、Alykhan Tejani、Johannes Totz、Zehan Wang 等. 2017年. 使用生成对抗网络实现照片级真实感的单幅图像超分辨率。载于IEEE计算机视觉与模式识别会议论文集，4681–4690。[224] Jure Leskovec、Deepayan Chakrabarti、Jon Kleinberg、Christos Faloutsos 和 Zoubin Ghahramani. 2010年. 克罗内克图：一种网络建模方法。《机器学习研究期刊》11卷2期（2010年）。[225] Stephen E Levinson、Lawrence R Rabiner 和 M Mohan Sondhi. 1983年. 将马尔可夫过程概率函数理论应用于自动语音识别的介绍。《贝尔系统技术期刊》62卷4期（1983年），1035–1074。[226] Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Ves Stoyanov 和 Luke Zettlemoyer. 2019年. BART：面向自然语言生成、翻译与理解的去噪序列到序列预训练。arXiv预印本 arXiv:1910.13461（2019年）。[227] Colin Lewis-Beck 和 Michael Lewis-Beck. 2015年. 应用回归：入门。第22卷。Sage Publications。[228] Boyun Li、Xiao Liu、Peng Hu、Zhongqin Wu、Jiancheng Lv 和 Xi Peng. 2022年. 面向未知损坏的全能图像复原。载于IEEE/CVF计算机视觉与模式识别会议论文集，17452–17462。[229] Bowen Li、Xiaojuan Qi、Thomas Lukasiewicz 和 Philip Torr. 2019年. 可控文本生成图像。神经信息处理系统进展32（2019年）。[230] Gang Li、Heliang Zheng、Chaoyue Wang、Chang Li、Changwen Zheng 和 Dacheng Tao. 2022年. 3DDesigner：利用文本引导扩散模型实现照片级真实感的三维物体生成与编辑。arXiv预印本 arXiv:2211.14108（2022年）。[231] Guang Li、Linchao Zhu、Ping Liu 和 Yi Yang. 2019年. 用于图像描述的纠缠变换器。载于IEEE/CVF国际计算机视觉会议论文集，8928–8937。[232] Haoying Li、Yifan Yang、Meng Chang、Huajun Feng、Zhi hai Xu、Qi Li 和 Yue ting Chen. 2022年. SRDiff：基于扩散概率模型的单幅图像超分辨率。《神经计算》479（2022年），47–59。提交给ACM的手稿

[233] Jason Li、Vitaly Lavrukhin、Boris Ginsburg、Ryan Leary、Oleksii Kuchaiev、Jonathan M Cohen、Huyen Nguyen 和 Ravi Teja Gadde。2019年。Jasper：端到端卷积神经声学模型。arXiv预印本 arXiv:1904.03288（2019年）。[234] Jinyu Li、Yu Wu、Yashesh Gaur、Chengyi Wang、Rui Zhao 和 Shujie Liu。2020年。大规模语音识别中热门端到端模型的比较。arXiv预印本 arXiv:2005.14327（2020年）。[235] Mao Li、Jiancheng Lv、Jian Wang 和 Yongsheng Sang。2020年。一种基于深度生成模型的抽象画生成方法。《神经处理快报》52（2020年），949–960。[236] Mu Li、左旺孟和张 David。2016年。用于属性驱动且保持身份的人脸生成的卷积网络。arXiv预印本 arXiv:1608.06434（2016年）。[237] Mu Li、左旺孟和张 David。2016年。面部属性的深度身份感知迁移。arXiv预印本 arXiv:1610.05586（2016年）。[238] Ruilong Li、Matthew Tancik 和 Angjoo Kanazawa。2022年。NerfAcc：通用神经辐射场加速工具箱。arXiv预印本 arXiv:2210.04847（2022年）。[239] Tianye Li、Timo Bolkart、Michael J Black、Hao Li 和 Javier Romero。2017年。从四维扫描中学习面部形状和表情模型。《ACM图形学汇刊》36，6（2017年），194–1。[240] Wenbo Li、Zhe Lin、Kun Zhou、Lu Qi、Yi Wang 和 Jiaya Jia。2022年。MAT：用于大孔洞图像修复的掩码感知变换器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，10758–10768。[241] Wei Li、Xue Xu、Xinyan Xiao、Jiachen Liu、Hu Yang、Guohao Li、Zhanpeng Wang、Zhifan Feng、Qiaoqiao She、Yajuan Lyu 等。2022年。UPainting：通过跨模态引导实现统一的文本到图像扩散生成。arXiv预印本 arXiv:2210.16031（2022年）。[242] Xin Li、Xin Jin、Jianxin Lin、Sen Liu、Yaojun Wu、Tao Yu、Wei Zhou 和 Zhibo Chen。2020年。学习解耦特征表示以复原混合失真图像。载于《计算机视觉欧洲会议 ECCV 2020：第16届欧洲会议，英国格拉斯哥，2020年8月23日至28日，论文集，第XXIX部分》第16卷。施普林格，313–329。[243] Xi Li 和 Ping Kuang。2021年。3D-VRVT：利用视觉变换器从单张图像进行三维体素重建。载于《2021年国际文化导向科学与技术会议（ICCST）》。电气电子工程师学会，343–348。[244] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei 等。2020年。Oscar：面向视觉语言任务的对象语义对齐预训练。载于《计算机视觉欧洲会议 ECCV 2020：第16届欧洲会议，英国格拉斯哥，2020年8月23日至28日，论文集，第XXX部分》第16卷。施普林格，121–137。[245] Xiang Lisa Li、John Thickstun、Ishaan Gulrajani、Percy Liang 和 Tatsunori B Hashimoto。2022年。Diffusion-LM 改进可控文本生成。arXiv预印本 arXiv:2205.14217（2022年）。[246] Yitong Li、Martin Min、Dinghan Shen、David Carlson 和 Lawrence Carin。2018年。文本生成视频。载于《AAAI人工智能会议论文集》第32卷。[247] Jingyun Liang、Jiezhang Cao、Guolei Sun、Kai Zhang、Luc Van Gool 和 Radu Timofte。2021年。Swinir：使用 Swin Transformer 进行图像复原。载于《IEEE/CVF国际计算机视觉会议论文集》，1833–1844。[248] Jianglin Liang 和 Ruifang Liu。2015年。结合使用堆叠去噪自编码器和丢弃法，防止深度神经网络过拟合。载于《2015年第八届图像与信号处理国际大会（CISP）》。电气电子工程师学会，697–701。[249] Renjie Liao、Yujia Li、Yang Song、Shenlong Wang、Will Hamilton、David K Duvenaud、Raquel Urtasun 和 Richard Zemel。2019年。利用图循环注意力网络高效生成图。神经信息处理系统进展32（2019年）。[250] Zibo Lin、Deng Cai、Yan Wang、Xiaojiang Liu、Hai-Tao Zheng 和 Shuming Shi。2020年。世界并非非黑即白：利用灰度数据学习对对话响应进行排序。arXiv预印本 arXiv:2004.02421（2020年）。[251] Pierre Lison 和 Serge Bibauw。2017年。对话并非一概而论：神经对话模型的实例加权。arXiv预印本 arXiv:1704.08966（2017年）。[252] Hongyu Liu、Ziyu Wan、Wei Huang、Yibing Song、Xintong Han 和 Jing Liao。2021年。Pd-gan：用于图像修复的概率多样化生成对抗网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，9371–9381。[253] Ming Liu、Yuxiang Wei、Xiaohe Wu、左旺孟和 Lei Zhang。2022年。利用预训练生成对抗网络进行图像编辑与复原的综述。arXiv预印本 arXiv:2207.10309（2022年）。[254] Qi Liu、Miltiadis Allamanis、Marc Brockschmidt 和 Alexander Gaunt。2018年。用于分子设计的约束图变分自编码器。神经信息处理系统进展31（2018年）。[255] Shuo Liu、Adria Mallol-Ragolta、Emilia Parada-Cabeleiro、Kun Qian、Xin Jing、Alexander Kathan、Bin Hu 和 Bjoern W Schuller。2022年。音频自监督学习：综述。arXiv预印本 arXiv:2203.01205（2022年）。[256] Wei Liu、Sihan Chen、郭龙腾、朱欣欣和刘静。2021年。CPTR：用于图像描述的全变换器网络。arXiv预印本 arXiv:2101.10804（2021年）。[257] Xia Liu、Alvin C Burns 和 Yingjian Hou。2017年。对Twitter上品牌相关用户生成内容的研究。《广告学报》46，2（2017年），236–247。[258] Xing Liu、Masanori Suganuma、Xiyang Luo 和 Takayuki Okatani。2019年。通过重复使用多分支网络复原退化因素未知的图像。arXiv预印本 arXiv:1907.04508（2019年）。[259] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种稳健优化的BERT预训练方法。arXiv预印本 arXiv:1907.11692（2019年）。[260] Yue Liu、Xin Wang、Yitian Yuan 和 Wenwu Zhu。2019年。用于句子到视频生成的跨模态双向学习。载于《第27届ACM国际多媒体会议论文集》，1239–1247。M

[261] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。2021年。Swin Transformer：利用移位窗口的层次化视觉变换器。国际计算机视觉大会。[262] Zhengzhe Liu、Yi Wang、Xiaojuan Qi 和 Chi-Wing Fu。2022年。迈向隐式文本引导的三维形状生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，17896–17906。[263] Dengsheng Lu 和 Qihao Weng。2007年。图像分类方法及提升分类性能技术综述。《国际遥感杂志》28，5（2007年），823–870。[264] Jiasen Lu、Caiming Xiong、Devi Parikh 和 Richard Socher。2017年。掌握何时关注：利用视觉哨兵实现图像描述的自适应注意力。载于《IEEE计算机视觉与模式识别会议论文集》，375–383。[265] Andreas Lugmayr、Martin Danelljan、Andres Romero、Fisher Yu、Radu Timofte 和 Luc Van Gool。2022年。Repaint：使用去噪扩散概率模型进行图像修复，11461–11471。[266] Tianze Luo、Zhanfeng Mo 和 Sinno Jialin Pan。2022年。基于谱扩散的快速图生成模型。arXiv预印本 arXiv:2211.08892（2022年）。[267] Minh-Thang Luong、Hieu Pham 和 Christopher D Manning。2015年。基于注意力的神经机器翻译的有效方法。arXiv预印本 arXiv:1508.04025（2015年）。[268] Christoph Lüscher、Eugen Beck、Kazuki Irie、Markus Kitza、Wilfried Michel、Albert Zeyer、Ralf Schlüter 和 Hermann Ney。2019年。亚琛工业大学 LibriSpeech 自动语音识别系统：混合模型与注意力模型——不使用数据增强。arXiv预印本 arXiv:1905.03072（2019年）。[269] Kaushalya Madhawa、Katushiko Ishiguro、Kosuke Nakago 和 Motoki Abe。2019年。Graphnvp：用于生成分子图的可逆流模型。arXiv预印本 arXiv:1905.11600（2019年）。[270] Mishaim Malik、Muhammad Kamran Malik、Khawar Mehmood 和 Imran Makhdoom。2021年。自动语音识别：综述。Multimedia Tools and Applications 80，6（2021年），9411–9457。[271] Elman Mansimov、Emilio Parisotto、Jimmy Lei Ba 和 Ruslan Salakhutdinov。2016年。利用注意力从图像描述生成图像。国际学习表征会议（2016年）。[272] Junhua Mao、Wei Xu、Yi Yang、Jiang Wang、Zhiheng Huang 和 Alan Yuille。2014年。利用多模态循环神经网络进行深度图像描述（M-RNN）。arXiv预印本 arXiv:1412.6632（2014年）。[273] Francesco Marconi。2020年。《新闻制造者：人工智能与新闻业的未来》。哥伦比亚大学出版社。[274] Rick Marshall。2018年。《银翼杀手2049》令人惊叹的视觉特效幕后揭秘。https://www.digitaltrends.com/movies/blade-runner-2049-visual- effects-john-nelson/（2018年）。[275] Richard Diehl Martinez 和 John Kaleialoha Kamalu。2018年。将生成对抗网络用于营销：爱彼迎案例研究。arXiv预印本 arXiv:1806.11432（2018年）。[276] Tanya Marwah、Gaurav Mittal 和 Vineeth N Balasubramanian。2017年。利用图像描述进行注意力引导的语义视频生成。载于《IEEE国际计算机视觉会议论文集》，1426–1434。[277]Alexandra Kennedy Maya Ackerman。2022年。人工智能驱动的歌词平台 LyricStudio 创作歌曲突破100万首。prweb.com/releases/2022/7/prweb18785303.htm（2022年）。[278] Łukasz Maziarka、Agnieszka Pocha、Jan Kaczmarczyk、Krzysztof Rataj、Tomasz Danel 和 Michał Warchoł。2020年。Mol-CycleGAN：一种用于分子优化的生成模型。《化学信息学杂志》12，1（2020年），1–18。[279] Alex McFarland。2022年。中国国家通讯社推出人工智能主播。https://www.unite.ai/chinas-state-news-agency- introduces-new-artificial-intelligence-anchor/（2022年）。[280] Alex McFarland。2023年。八款优秀的人工智能音乐生成器。https://www.unite.ai/best-ai-music-generators/（2023年）。[281] Larry R Medsker 和 LC Jain。2001年。循环神经网络。《循环神经网络：设计与应用》5（2001年），64–67。[282] Shikib Mehri、Evgeniia Razumovskaia、Tiancheng Zhao 和 Maxine Eskenazi。2019年。对话上下文表示学习的预训练方法。arXiv预印本 arXiv:1906.00414（2019年）。[283] Lars Mescheder、Michael Oechsle、Michael Niemeyer、Sebastian Nowozin 和 Andreas Geiger。2019年。占用网络：在函数空间中学习三维重建。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，4460–4470。[284] Meta。2022年。什么是 Meta Horizon Worlds? https://www.marktechpost.com/2023/01/01/top-artificial-intelligence-ai-tools-that-can-generate-code-to- help-programmers/（2022年）。[285] Tomas Mikolov、Martin Karafiát、Lukas Burget、Jan Cernock`y 和 Sanjeev Khudanpur。2010年。基于循环神经网络的语言模型。载于 Interspeech，第2卷。幕张，1045–1048。[286] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。2021年。神经辐射场：用于视图合成的场景表示。《美国计算机协会通讯》65，1（2021年），99–106。[287] Diego H Milone 和 Leandro E Di Persia。2008年。将隐马尔可夫树作为观测分布来学习隐马尔可夫模型。《人工智能：伊比利亚美洲人工智能评论》12，37（2008年），7–13。[288] Andrey Miroshnichenko。2018年。人工智能绕过创造力：机器人会取代新闻记者吗？（答案是“会”）。《信息》9，7（2018年），183。[289] Gautam Mittal、Jesse Engel、Curtis Hawthorne 和 Ian Simon。2021年。利用扩散模型生成符号音乐。arXiv预印本 arXiv:2103.16091（2021年）。[290] Gaurav Mittal、Tanya Marwah 和 Vineeth N Balasubramanian。2017年。Sync-draw：利用深度循环注意力架构自动生成视频。载于《第25届ACM国际多媒体会议论文集》，1096–1104。[291] Irina Momot。2022年。电影制作过程中的人工智能：未来场景。（2022年）。

[292] Masanori Morise、Fumiya Yokomori 和 Kenji Ozawa。2016年。WORLD：一种基于声码器、适用于实时应用的高质量语音合成系统。《IEICE Transactions on Information and Systems》99，7（2016年），1877–1884。[293] Eric Moulines 和 Francis Charpentier。1990年。使用双音素进行文本转语音合成的基音同步波形处理技术。《Speech Communication》9，5-6（1990年），453–467。[294] Thomas Müller、Alex Evans、Christoph Schied 和 Alexander Keller。2022年。采用多分辨率哈希编码的即时神经图形基元。《ACM图形学汇刊》（ToG）41，4（2022年），1–15。[295] Frank-Michael Nack。1996年。AUTEUR：视频语义和主题表示在自动化电影剪辑中的应用。博士论文。Citeseer。[296] Tomohiro Nakatani。2019年。结合连接主义时序分类和语言模型集成，改进基于 Transformer 的端到端语音识别。载于 Interspeech 会议论文集。[297] Ali Bou Nassif、Ismail Shahin、Imtinan Attili、Mohammad Azzeh 和 Khaled Shaalan。2019年。使用深度神经网络进行语音识别：系统综述。《IEEE Access》7（2019年），19143–19165。[298] Kamyar Nazeri、Eric Ng、Tony Joseph、Faisal Z Qureshi 和 Mehran Ebrahimi。2019年。EdgeConnect：通过对抗性边缘学习进行生成式图像修复。arXiv 预印本 arXiv:1901.00212（2019年）。[299] Jinjie Ni、Tom Young、Vlad Pandelea、Fuzhao Xue 和 Erik Cambria。2022年。基于深度学习的对话系统最新进展：系统综述。《人工智能评论》（2022年），1–101。[300] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2022年。GLIDE：迈向基于文本引导扩散模型的照片级真实感图像生成与编辑。ICML（2022年）。[301] Huansheng Ning、Hang Wang、Yujia Lin、Wenxi Wang、Sahraoui Dhelim、Fadi Farha、Jianguo Ding 和 Mahmoud Daneshmand。2021年。元宇宙综述：最新进展、技术、应用与挑战。arXiv 预印本 arXiv:2111.09673（2021年）。[302] Chenhao Niu、Yang Song、Jiaming Song、Shengjia Zhao、Aditya Grover 和 Stefano Ermon。2020年。通过基于得分的生成建模实现置换不变的图生成。载于人工智能与统计学会议。PMLR，4474–4484。[303] Mehdi Noroozi 和 Paolo Favaro。2016年。通过解决拼图任务进行视觉表征的无监督学习。载于欧洲计算机视觉会议。[304]中国信息通信研究院。2022年。《人工智能生成内容（AIGC）白皮书》。http://www.caict.ac.cn/english/research/whitepapers/202211/t*20221111_411288*.html（2022年）。[305] Katsunori Ohnishi、Shohei Yamamoto、Yoshitaka Ushiku 和 Tatsuya Harada。2018年。从正交信息生成分层视频：光流与纹理。载于 AAAI人工智能会议论文集，第32卷。[306] Joseph Olive。1977年。基于二元单元的语音规则合成。载于 ICASSP'77。IEEE国际声学、语音与信号处理会议，第2卷。电气电子工程师学会，568–570。[307] OpenAI。2023年。GPT-4技术报告。arXiv 预印本 arXiv:2303.08774（2023年）。[308] Tim O'Reilly。2009年。什么是 Web 2.0。《O'Reilly Media, Inc.》。[309] Will Oremus。2014年。关于洛杉矶地震的首篇新闻报道由机器人撰写。https://slate.com/technology/2014/03/quakebot-los- angeles-times-robot-journalist-writes-article-on-la-earthquake.html（2014年）。[310] Achraf Oussidi 和 Azeddine Elhassouny。2018年。深度生成模型：综述。载于2018年智能系统与计算机视觉国际会议（ISCV）。电气电子工程师学会，1–8。[311] Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。2022年。通过人类反馈训练语言模型遵循指令。arXiv 预印本 arXiv:2203.02155（2022年）。[312] Yawen Ouyang、Moxin Chen、Xinyu Dai、Yinggong Zhao、Shujian Huang 和 Jiajun Chen。2020年。通过显式槽位连接建模进行对话状态跟踪。载于计算语言学协会第58届年会论文集。34–40。[313] Yingwei Pan、Zhaofan Qiu、Ting Yao、Houqiang Li 和 Tao Mei。2017年。言出图生：根据图像描述生成视频。载于第25届 ACM 国际多媒体会议论文集。1789–1798。[314] Tianyu Pang、Kun Xu、Chongxuan Li、Yang Song、Stefano Ermon 和 Jun Zhu。2020年。通过有限差分得分匹配高效学习生成模型。《神经信息处理系统进展》33（2020年），19175–19188。[315] Ankur P Parikh、Oscar Täckström、Dipanjan Das 和 Jakob Uszkoreit。2016年。一种用于自然语言推理的可分解注意力模型。arXiv 预印本 arXiv:1606.01933（2016年）。[316] Giorgio Parisi。1981年。关联函数与计算机模拟。《Nuclear Physics B》180，3（1981年），378–384。[317] A Parkes。1989年。场景与场景结构：用于浏览视盘图像状态的网络描述与自动传播。载于第12届 ACM 国际信息检索研究与发展年会论文集。229–238。[318] Alan Philip Parkes。1988年。一种用于视盘图像概念描述的人工智能方法。兰卡斯特大学（英国）。[319] Alan P Parkes。1989年。原型 CLORIS 系统：描述、检索和讨论视盘静态图像与序列。《Information Processing & Management》25，2（1989年），171–186。[320] Santiago Pascual、Gautam Bhattacharya、Chunghsin Yeh、Jordi Pons 和 Joan Serrà。2022年。基于得分扩散的全频带通用音频合成。arXiv 预印本 arXiv:2210.14661（2022年）。[321] Santiago Pascual、Mirco Ravanelli、Joan Serra、Antonio Bonafonte 和 Yoshua Bengio。2019年。从多个自监督任务中学习与问题无关的语音表征。arXiv 预印本 arXiv:1904.03416（2019年）。

[322] Steve Paulussen 和 Pieter Ugille。2008年。新闻编辑部中的用户生成内容：参与式新闻业面临的专业与组织约束。《Westminster Papers in Communication & Culture》5，2（2008年）。[323] 彭宝林、朱晨光、Chunyuan Li、Xiujun Li、李金超、Michael Zeng 和 Jianfeng Gao。2020年。面向任务型对话的少样本自然语言生成。arXiv 预印本 arXiv:2002.12328（2020年）。[324] Tammy Pettinato Oltz。2023年。ChatGPT，法学教授。《Professor of Law》（2023年2月4日）（2023年）。[325] Cale Plut 和 Philippe Pasquier。2020年。电子游戏中的生成式音乐：最新进展、挑战与前景。《Entertainment Computing》33（2020年），100337。[326] Ben Poole、Ajay Jain、Jonathan T Barron 和 Ben Mildenhall。2022年。DreamFusion：通过二维扩散实现文本生成三维。arXiv 预印本 arXiv:2209.14988（2022年）。[327] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。2020年。学习个体说话风格以实现准确的唇动到语音合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集。13796–13805。[328] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。2020年。野外语音到唇形生成，只需一位唇形同步专家。载于第28届 ACM 国际多媒体会议论文集。484–492。[329] Ash & Pri。2022年。Scalenut 评测：功能介绍及如何将其用作内容生成器。https://ashandpri.com/scalenutreview（2022年）。[330] Charles R Qi、Hao Su、Kaichun Mo 和 Leonidas J Guibas。2017年。PointNet：基于点集的深度学习，用于三维分类与分割。载于计算机视觉与模式识别会议。[331] Charles Ruizhongtai Qi、Li Yi、Hao Su 和 Leonidas J Guibas。2017年。PointNet++：在度量空间中基于点集进行深层次特征学习。《神经信息处理系统进展》30（2017年）。[332] Weizhen Qi、Yu Yan、Yeyun Gong、Dayiheng Liu、Nan Duan、Jiusheng Chen、Ruofei Zhang 和 Ming Zhou。2020年。ProphetNet：为序列到序列预训练预测未来 n 元语法。arXiv 预印本 arXiv:2001.04063（2020年）。[333] Yao Qian、Yuchen Fan、Wenping Hu 和 Frank K Soong。2014年。关于用于参数化文本转语音合成的深度神经网络（DNN）训练方面的研究。载于2014年 IEEE 声学、语音与信号处理国际会议（ICASSP）。电气电子工程师学会，3829–3833。[334] Long Qin。2013年。自动语音识别中的未登录词学习。博士论文。卡内基梅隆大学。[335] Lisong Qiu、Juntao Li、Wei Bi、Dongyan Zhao 和 Rui Yan。2019年。训练样本是否相关？学习利用多个参考答案生成对话回复。载于计算语言学协会第57届年会论文集。3826–3835。[336] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移视觉模型。载于 ICML。[337] Alec Radford、Luke Metz 和 Soumith Chintala。2015年。利用深度卷积生成对抗网络进行无监督表征学习。arXiv 预印本 arXiv:1511.06434（2015年）。[338] Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。2018年。通过生成式预训练改进语言理解。（2018年）。[339] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督多任务学习者。《OpenAI 博客》（2019年）。[340] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本转换器的迁移学习极限。《机器学习研究杂志》21，140（2020年），1–67。[341] Salvatore Raieli。2023年。ControlNet：掌控你的人工智能艺术生成。https://levelup.gitconnected.com/controlnet-control-your-ai-art-generation- 616c86c88964（2023年）。[342] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。使用 CLIP 潜变量进行分层文本条件图像生成。arXiv 预印本 arXiv:2204.06125（2022年）。[343] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。2021年。零样本文本生成图像。载于 ICML。[344] Jeremiah Ratican、James Hutson 和 Andrew Wright。2023年。一种拟议的元现实沉浸式开发流程：面向元宇宙的生成式人工智能模型与扩展现实（XR）内容。《智能学习系统与应用杂志》15（2023年）。[345] Pratiksha C Raut 和 Seema U Deoghare。2016年。自动语音识别及其应用。《国际工程与技术研究期刊》3，5（2016年），2368–2371。[346] Mirco Ravanelli、Jianyuan Zhong、Santiago Pascual、Pawel Swietojanski、Joao Monteiro、Jan Trmal 和 Yoshua Bengio。2020年。用于鲁棒语音识别的多任务自监督学习。载于 ICASSP 2020—2020年 IEEE国际声学、语音与信号处理会议（ICASSP）。电气电子工程师学会，6989–6993。[347] D Raj Reddy。1976年。机器语音识别：综述。《IEEE Proceedings》64，4（1976年），501–531。[348] Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran、Bernt Schiele 和 Honglak Lee。2016年。生成对抗式文本到图像合成。载于国际机器学习会议。PMLR，1060–1069。[349] Mengwei Ren、Mauricio Delbracio、Hossein Talebi、Guido Gerig 和 Peyman Milanfar。2022年。使用领域泛化扩散模型进行图像去模糊。arXiv 预印本 arXiv:2212.01789（2022年）。[350] Shuo Ren、Wenhu Chen、Shujie Liu、Mu Li、Ming Zhou 和 Shuai Ma。2018年。用于低资源语言翻译的三角架构。arXiv 预印本 arXiv:1805.04813（2018年）。[351] @rev。2023年。快速、准确的转录服务。https://www.rev.com/（2023年）。

[352] Alexander Richard、Michael Zollhöfer、Yandong Wen、Fernando De la Torre 和 Yaser Sheikh。2021年。Meshtalk：利用跨模态解耦实现语音驱动的三维人脸动画。载于 IEEE/CVF国际计算机视觉会议论文集。1173–1182。[353] Alan Ritter、Colin Cherry 和 Bill Dolan。2011年。社交媒体中的数据驱动回复生成。载于自然语言处理实证方法会议（EMNLP）。[354] Adam Roberts、Jesse Engel、Colin Raffel、Curtis Hawthorne 和 Douglas Eck。2018年。用于学习音乐长期结构的分层潜在向量模型。载于国际机器学习会议。PMLR，4364–4373。[355] Vincent Roger、Jérôme Farinas 和 Julien Pinquier。2022年。用于自动语音处理的深度神经网络：从大型语料库到有限数据的综述。EURASIP音频、语音与音乐处理期刊 2022，1（2022），1–15。[356] Ben Rogerson。2020年。Sony CSL推出 Flow Machines，一款AI辅助音乐制作插件。https://www.musicradar.com/news/sony-csl-launches-flow-machines-an-ai-assisted-music-production-plugin（2020年）。[357] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022年。利用潜在扩散模型进行高分辨率图像合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10684–10695。[358] Joe Rosato。2020年。数千名加州大学伯克利分校高年级学生将在Minecraft毕业典礼上毕业。https://www.nbcbayarea.com/news/coronavirus/thousands- of-uc-berkeley-seniors-to-graduate-in-minecraft-ceremony/2291312/（2020年）。[359] Sascha Rothe、Shashi Narayan 和 Aliaksei Severyn。2020年。利用预训练检查点完成序列生成任务。《计算语言学协会汇刊》8（2020），264–280。[360] JOSH ROTTENBERG。2019年。《双子杀手》视觉特效团队如何打造年轻版威尔·史密斯。https://www.latimes.com/entertainment-arts/movies/story/2019-08-27/gemini-man-visual-effects-young-will-smith（2019年）。[361] Simon Rouard 和 Gaëtan Hadjeres。2021年。CRASH：用于可控高分辨率鼓声合成的基于原始音频评分的生成建模。arXiv预印本 arXiv:2106.07431（2021年）。[362] Daniel Ruby。2023年。Jasper AI评测（2023）：使用18个月后的体验。https://www.demandsage.com/jasper-ai-review/（2023年）。[363] Riaan Rudman 和 Rikus Bruwer。2016年。定义Web 3.0：机遇与挑战。《电子图书馆》（2016年）。[364] Kiersten M Ruff 和 Rohit V Pappu。2021年。AlphaFold及其对内在无序蛋白质的影响。《分子生物学杂志》433，20（2021），167208。[365] David E Rumelhart、Geoffrey E Hinton 和 Ronald J Williams。1985年。通过误差传播学习内部表征。技术报告。加利福尼亚大学圣迭戈分校拉霍亚认知科学研究所。[366] Warren Sack 和 Marc Davis。1994年。IDIC：根据故事计划和内容标注组装视频序列。载于 ICMCS。30–36。[367] Chitwan Saharia、William Chan、Huiwen Chang、Chris Lee、Jonathan Ho、Tim Salimans、David Fleet 和 Mohammad Norouzi。2022年。Palette：图像到图像扩散模型。载于计算机图形学与交互技术特别兴趣小组会议论文集。1–10。[368] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。2022年。结合深度语言理解的照片级真实感文本到图像扩散模型。arXiv预印本 arXiv:2205.11487（2022年）。[369] Chitwan Saharia、Jonathan Ho、William Chan、Tim Salimans、David J Fleet 和 Mohammad Norouzi。2022年。通过迭代细化实现图像超分辨率。IEEE模式分析与机器智能汇刊（2022年）。[370] Masaki Saito、Eiichi Matsumoto 和 Shunta Saito。2017年。采用奇异值裁剪的时间生成对抗网络。载于 IEEE国际计算机视觉会议论文集。2830–2839。[371] Masaki Saito、Shunta Saito、Masanori Koyama 和 Sosuke Kobayashi。2020年。稀疏训练，密集生成：高分辨率时间生成对抗网络的高效内存无监督训练。《国际计算机视觉杂志》128，10（2020），2586–2606。[372] Rose Salia。2021年。十大最佳AI人脸应用评测。https://topten.ai/face-apps-review/（2021年）。[373] Tim Salimans、Andrej Karpathy、Xi Chen 和 Diederik P Kingma。2017年。Pixelcnn++：通过离散化逻辑斯蒂混合似然及其他改进提升PixelCNN。arXiv预印本 arXiv:1701.05517（2017年）。[374] Saeed Saremi、Arash Mehrjou、Bernhard Schölkopf 和 Aapo Hyvärinen。2018年。深度能量估计网络。arXiv预印本 arXiv:1805.08306（2018年）。[375] Jiirgen Schmidhuber。1990年。让世界可微：在动态强化学习和非平稳环境规划中使用自监督全循环神经网络。（1990年）。[376] Mike Schuster 和 Kuldip K Paliwal。1997年。双向循环神经网络。IEEE信号处理汇刊 45，11（1997），2673–2681。[377] P Seeviour、J Holmes 和 M Judd。1976年。为并联共振峰语音合成器自动生成控制信号。载于 ICASSP'76。IEEE国际声学、语音与信号处理会议，第1卷。电气电子工程师学会，690–693。[378] Iulian V Serban、Chinnadhurai Sankar、Mathieu Germain、Saizheng Zhang、Zhouhan Lin、Sandeep Subramanian、Taesup Kim、Michael Pieper、Sarath Chandar、Nan Rosemary Ke 等。2017年。深度强化学习聊天机器人。arXiv预印本 arXiv:1709.02349（2017年）。[379] Quin Sexton。2023年。电影配乐：电影音乐创作解析。（2023年）。[380] Christine H Shadle 和 Robert I Damper。2001年。发音合成的前景：立场论文。载于第四届 ISCA 语音合成教程与研究研讨会（ITRW）。[381] Neil Shah、Sarth Engineer、Nandish Bhagat、Hirwa Chauhan 和 Manan Shah。2020年。广告中机器学习与人工智能应用的研究趋势。《增强人类研究》5（2020），1–15。

[382] Yong Shan、Zekang Li、Jinchao Zhang、Fandong Meng、Yang Feng、Cheng Niu 和 Jie Zhou。2020年。用于对话状态跟踪的自适应目标函数上下文层次注意力网络。载于计算语言学协会第58届年会论文集。6322–6333。[383] Itamar Shatz。2017年。第二语言习得过程中的母语影响：大规模学习者语料库分析。载于太平洋第二语言研究论坛论文集（PacSLRF 2016）。175–188。[384] Jonathan Shen、Ruoming Pang、RonJ Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、YuxuanWang、Rj Skerrv-Ryan 等。2018年。通过以梅尔频谱图预测为条件的WaveNet实现自然语音合成。载于2018年 IEEE声学、语音与信号处理国际会议（ICASSP）。电气电子工程师学会，4779–4783。[385] Wei Shen 和 Rujie Liu。2017年。学习用于人脸属性操控的残差图像。载于IEEE计算机视觉与模式识别会议论文集。4030–4038。[386] Yujun Shen 和 Bolei Zhou。2021年。生成对抗网络中潜在语义的闭式分解。载于 IEEE/CVF计算机视觉与模式识别会议论文集。1532–1540。[387] Shelly Sheynin、Oron Ashual、Adam Polyak、Uriel Singer、Oran Gafni、Eliya Nachmani 和 Yaniv Taigman。2022年。KNN-Diffusion：通过大规模检索实现图像生成。arXiv预印本 arXiv:2204.02849（2022年）。[388] Chence Shi、Shitong Luo、Minkai Xu 和 Jian Tang。2021年。学习用于分子构象生成的梯度场。9558–9568。[389] Jiaxin Shi、Shengyang Sun 和 Jun Zhu。2018年。用于隐式分布梯度估计的谱方法。载于国际机器学习会议。PMLR，4644–4653。[390] Ziqiang Shi 和 Shoule Wu。2022年。ITÔN：利用伊藤随机微分方程实现端到端音频生成。《数字信号处理》132（2022），103781。[391] Wooksu Shin、Namhyuk Ahn、Jeong-Hyeon Moon 和 Kyung-Ah Sohn。2022年。利用失真信息进行多重退化图像复原。载于 IEEE/CVF计算机视觉与模式识别会议论文集。537–546。[392] Suman Shrestha。2014年。使用新型自适应中值滤波器进行图像去噪。arXiv预印本 arXiv:1410.2175（2014年）。[393] Karen Simonyan 和 Andrew Zisserman。2015年。用于大规模图像识别的超深卷积网络。载于国际学习表征会议。[394] Uriel Singer、Adam Polyak、Thomas Hayes、Xi Yin、Jie An、Songyang Zhang、Qiyuan Hu、Harry Yang、Oron Ashual、Oran Gafni 等。2022年。Make-a-Video：无需文本-视频数据的文本生成视频。arXiv预印本 arXiv:2209.14792（2022年）。[395] Karan Singla、Zhuohao Chen、David Atkins 和 Shrikanth Narayanan。2020年。迈向端到端学习：从心理治疗对话中的口语话语预测行为编码。载于计算语言学协会第58届年会论文集。3797–3803。[396] Bogdan Smolka、K Czubin、Jon Yngve Hardeberg、Kostas N Plataniotis、Marek Szczepanski 和 Konrad Wojciechowski。2003年。迈向自动去除红眼效应。《模式识别快报》24，11（2003），1767–1785。[397] Yoon So-Yeon。2020年。MBN推出韩国首位人工智能新闻主播。https://koreajoongangdaily.joins.com/2020/11/10/entertainment/television/MBN-AI-artificial-intelligence/20201110153900457.html（2020年）。[398] Than Htut Soe。2021年。视频编辑自动化：视频编辑中的辅助工作流程。载于 AutomationXP@ CHI。[399] Fei Song 和 W Bruce Croft。1999年。用于信息检索的通用语言模型。载于第八届信息与知识管理国际会议论文集。316–321。[400] Kaitao Song、Xu Tan、Tao Qin、Jianfeng Lu 和 Tie-Yan Liu。2019年。MASS：用于语言生成的掩码序列到序列预训练。arXiv预印本 arXiv:1905.02450（2019年）。[401] Yang Song 和 Stefano Ermon。2019年。通过估计数据分布的梯度进行生成建模，第32卷。[402] Yang Song、Sahaj Garg、Jiaxin Shi 和 Stefano Ermon。2020年。切片分数匹配：一种可扩展的密度与分数估计方法。载于人工智能中的不确定性。PMLR，574–584。[403] Yang Song 和 Diederik P Kingma。2021年。如何训练基于能量的模型。arXiv预印本 arXiv:2101.03288（2021年）。[404] Yang Song、Jingwen Zhu、Dawei Li、Xiaolong Wang 和 Hairong Qi。2018年。通过条件循环对抗网络生成说话人脸。arXiv预印本 arXiv:1804.04786（2018年）。[405] Nitish Srivastava 和 Russ R Salakhutdinov。2012年。利用深度玻尔兹曼机进行多模态学习。神经信息处理系统进展25（2012年）。[406] Nick Statt。2019年。什么是Memoji? 如何创建一个长得像你的Animoji。https://www.theverge.com/2019/7/2/20679241/xiaomi-mimoji-apple-memoji-clone-copying-smartphone-ar-avatar（2019年）。[407] Matteo Stefanini、Marcella Cornia、Lorenzo Baraldi、Silvia Cascianelli、Giuseppe Fiameni 和 Rita Cucchiara。2021年。从展示到讲述：图像描述综述。arXiv预印本 arXiv:2107.06912（2021年）。[408] Hui Su、Xiaoyu Shen、Sanqiang Zhao、Xiao Zhou、Pengwei Hu、Randy Zhong、Cheng Niu 和 Jie Zhou。2020年。利用非对话文本丰富对话生成。arXiv预印本 arXiv:2005.04346（2020年）。[409] Shang-Yu Su、Chao-Wei Huang 和 Yun-Nung Chen。2019年。用于自然语言理解与生成的双重监督学习。arXiv预印本 arXiv:1905.06196（2019年）。[410] Masanori Suganuma、Xing Liu 和 Takayuki Okatani。2019年。在未知组合失真情况下，采用基于注意力的自适应操作选择进行图像复原。载于 IEEE/CVF计算机视觉与模式识别会议论文集。9039–9048。[411] Jian Sun、Wenfei Cao、Zongben Xu 和 Jean Ponce。2015年。学习用于去除非均匀运动模糊的卷积神经网络。载于 IEEE计算机视觉与模式识别会议论文集。769–777。

[412] Jiao Sun、Q Vera Liao、Michael Muller、Mayank Agarwal、Stephanie Houde、Kartik Talamadupula 和 Justin D Weisz。2022年。通过基于情景的设计，探究生成式人工智能代码的可解释性。载于第27届智能用户界面国际会议。212–228。[413] Ilya Sutskever、Oriol Vinyals 和 Quoc V Le。2014年。利用神经网络进行序列到序列学习。载于《神经信息处理系统进展》第27卷（2014年）。[414] Supasorn Suwajanakorn、Steven M Seitz 和 Ira Kemelmacher-Shlizerman。2017年。合成奥巴马：从音频中学习唇形同步。《ACM图形学汇刊》（ToG）36卷第4期（2017年），1–13。[415] Deborah Swanberg、Chiao Fe Shu 和 Ramesh Jain。1993年。基于内容检索的多媒体信息系统架构。载于《数字音视频网络与操作系统支持：第三届国际研讨会，拉霍亚，加利福尼亚州，美国，1992年11月12日至13日，论文集》第3卷。施普林格，387–392。[416] Laura Sydell。2009年。《本杰明·巴顿的奇事》中的奇异面孔是如何打造的。https://www.npr.org/2009/02/17/100668766/building-the-curious-faces-of-benjamin-button（2009年）。[417] Christian Szegedy、Wei Liu、Yangqing Jia、Pierre Sermanet、Scott Reed、Dragomir Anguelov、Dumitru Erhan、Vincent Vanhoucke 和 Andrew Rabinovich。2015年。通过卷积网络不断深入。载于计算机视觉与模式识别会议。[418] Mingxing Tan 和 Quoc Le。2019年。EfficientNet：重新思考卷积神经网络的模型缩放。载于国际机器学习会议。PMLR，6105–6114。[419] Xian Tang。2009年。用于自动语音识别的混合隐马尔可夫模型与人工神经网络。载于2009年亚太电路、通信与系统会议。电气电子工程师学会，682–685。[420] Victor Tangermann。2023年。研究称，89%的大学生承认用ChatGPT完成家庭作业。https://futurism.com/the-byte/students-admit-chatgpt-homework（2023年）。[421] DeepMind研究团队。2017年。DeepMind 2016年的工作：回顾。https://www.deepmind.com/blog/deepminds-work-in-2016-a-round-up（2017年）。[422] Latitude团队。2020年。AI Dungeon：Dragon模型升级。https://aidungeon.medium.com/ai-dungeon-dragon-model-upgrade-7e8ea579abfe（2020年）。[423] Justus Thies、Mohamed Elgharib、Ayush Tewari、Christian Theobalt 和 Matthias Nießner。2020年。神经语音操控：音频驱动的人脸重演。载于《计算机视觉欧洲会议 ECCV 2020：第16届欧洲计算机视觉会议，格拉斯哥，英国，2020年8月23日至28日，论文集》第XVI部分第16卷。施普林格，716–731。[424] Yu Tian、Jian Ren、Menglei Chai、Kyle Olszewski、Xi Peng、Dimitris N Metaxas 和 Sergey Tulyakov。2021年。高分辨率视频合成所需的图像生成器。arXiv预印本 arXiv:2104.15069（2021年）。[425] Tijmen Tieleman。2008年。利用似然梯度近似训练受限玻尔兹曼机。载于第25届国际机器学习会议论文集。1064–1071。[426] Maggie Tillman。2022年。什么是Memoji？如何创建一个长得像你的Animoji。https://www.pocket-lint.com/phones/news/apple/144743-what-are-memoji-how-to-create-an-animoji-that-looks-like-you/（2022年）。[427] Artem Timoshenko 和 John R Hauser。2019年。从用户生成内容中识别客户需求。《营销科学》38卷第1期（2019年），1–20。[428] Yoshinobu Tonomura、Akihito Akutsu、Yukinobu Taniguchi 和 Gen Suzuki。1994年。结构化视频计算。《IEEE MultiMedia》1卷03期（1994年），34–43。[429] László Tóth。2011年。一种用于改进音素识别的分层上下文相关神经网络架构。载于2011年IEEE声学、语音与信号处理国际会议（ICASSP）。电气电子工程师学会，5040–5043。[430] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Hervé Jégou。2020年。训练数据高效图像Transformer与基于注意力的蒸馏。arXiv预印本 arXiv:2012.12877（2020年）。[431] Hugo Touvron、Matthieu Cord、Alexandre Sablayrolles、Gabriel Synnaeve 和 Hervé Jégou。2021年。通过图像Transformer不断深入。载于IEEE/CVF国际计算机视觉会议论文集。32–42。[432] Aggeliki Tsoli、Antonis Argyros等。2019年。根据单张RGB图像，对无纹理可变形三维表面进行基于块的重建。载于IEEE/CVF国际计算机视觉会议研讨会论文集。0–0。[433] Sergey Tulyakov、Ming-Yu Liu、Xiaodong Yang 和 Jan Kautz。2018年。Mocogan：分解运动与内容以生成视频。载于IEEE计算机视觉与模式识别会议论文集。1526–1535。[434] Hirotada Ueda、Takafumi Miyatake、Shigeo Sumino 和 Akio Nagasaka。1993年。用于视频剪辑的自动结构可视化。载于INTERACT'93和CHI'93计算系统中的人为因素会议论文集。137–141。[435] Anwaar Ulhaq、Naveed Akhtar 和 Ganna Pogrebna。2022年。视觉高效扩散模型：综述。arXiv预印本 arXiv:2210.09292（2022年）。[436] Paul Upchurch、Jacob Gardner、Geoff Pleiss、Robert Pless、Noah Snavely、Kavita Bala 和 Kilian Weinberger。2017年。用于改变图像内容的深度特征插值。载于IEEE计算机视觉与模式识别会议论文集。7064–7073。[437] Benigno Uria、Marc-Alexandre Côté、Karol Gregor、Iain Murray 和 Hugo Larochelle。2016年。神经自回归分布估计。《机器学习研究期刊》17卷第1期（2016年），7184–7220。[438] Benigno Uria、Iain Murray 和 Hugo Larochelle。2013年。RNADE：实值神经自回归密度估计器。载于《神经信息处理系统进展》第26卷（2013年）。[439] Demetrios Vakratsas 和 Xin Wang。2020年。广告创意中的人工智能。《广告学报》50卷第1期（2020年），39–51。[440] Aäron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew W. Senior 和 Koray Kavukcuoglu。2016年。WaveNet：一种原始音频生成模型。载于第9届ISCA语音合成研讨会。已提交至美国计算机协会的手稿

[441] Aaron Van den Oord、Nal Kalchbrenner、Lasse Espeholt、Oriol Vinyals、Alex Graves等。2016年。使用PixelCNN解码器进行条件图像生成。载于《神经信息处理系统进展》第29卷（2016年）。[442] Domonkos Varga 和 Tamás Szirányi。2016年。基于卷积神经网络的全自动图像着色。载于2016年第23届国际模式识别会议（ICPR）。电气电子工程师学会，3691–3696。[443] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。Attention is All You Need。载于NeurIPS。[444] Yuri Viazovetskyi、Vladimir Ivashkin 和 Evgeny Kashin。2020年。用于前馈图像操控的StyleGAN2蒸馏。载于欧洲计算机视觉会议。施普林格，170–186。[445] Clement Vignac、Igor Krawczuk、Antoine Siraudin、Bohan Wang、Volkan Cevher 和 Pascal Frossard。2022年。DiGress：用于图生成的离散去噪扩散。arXiv预印本 arXiv:2209.14734（2022年）。[446] Pascal Vincent。2011年。分数匹配与去噪自编码器之间的联系。《神经计算》23卷第7期（2011年），1661–1674。[447] Oriol Vinyals、Alexander Toshev、Samy Bengio 和 Dumitru Erhan。2015年。看图说话：神经图像描述生成器。载于IEEE计算机视觉与模式识别会议论文集。3156–3164。[448] Carl Vondrick、Hamed Pirsiavash 和 Antonio Torralba。2016年。利用场景动态生成视频。载于《神经信息处理系统进展》第29卷（2016年）。[449] Andrey Voynov、Kfir Aberman 和 Daniel Cohen-Or。2022年。草图引导的文本到图像扩散模型。arXiv预印本 arXiv:2211.13752（2022年）。[450] Bram Wallace、Akash Gokul 和 Nikhil Naik。2022年。EDICT：通过耦合变换实现精确扩散反演。arXiv预印本 arXiv:2211.12446（2022年）。[451] Hanna M Wallach。2006年。主题建模：超越词袋。载于第23届国际机器学习会议论文集。977–984。[452] Ziyu Wan、Bo Zhang、Dongdong Chen、Pan Zhang、Dong Chen、Fang Wen 和 Jing Liao。2022年。通过深度潜在空间转换修复老照片。《IEEE模式分析与机器智能汇刊》（2022年）。[453] Chengyi Wang、Yu Wu、Yujiao Du、Jinyu Li、Shujie Liu、Liang Lu、Shuo Ren、Guoli Ye、Sheng Zhao 和 Ming Zhou。2019年。用于基于Transformer的端到端语音识别的语义掩码。arXiv预印本 arXiv:1912.03010（2019年）。[454] Dan Wang、Xinrui Cui、Xun Chen、Zhengxia Zou、Tianyang Shi、Septimiu Salcudean、Z Jane Wang 和 Rabab Ward。2021年。利用Transformer进行多视图三维重建。载于IEEE/CVF国际计算机视觉会议论文集。5722–5731。[455] Jun Wang、Ying Cui、Dongyan Guo、Junxia Li、Qingshan Liu 和 Chunhua Shen。2022年。PointAttN：点云补全只需注意力。arXiv预印本 arXiv:2203.08485（2022年）。[456] Li Wang、Zechen Bai、Yonghua Zhang 和 Hongtao Lu。2020年。看、回忆、讲述：利用回忆机制进行图像描述。载于AAAI人工智能会议论文集，第34卷。12176–12183。[457] Nanyang Wang、Yinda Zhang、Zhuwen Li、Yanwei Fu、Wei Liu 和 Yu-Gang Jiang。2018年。Pixel2mesh：从单张RGB图像生成三维网格模型。载于欧洲计算机视觉会议（ECCV）论文集。52–67。[458] Rui Wang、Xu Tan、Renqian Luo、Tao Qin 和 Tie-Yan Liu。2021年。低资源神经机器翻译综述。arXiv预印本 arXiv:2107.04239（2021年）。[459] Xuejiao Wang、Qiuyan Tao、Lianghao Wang、Dongxiao Li 和 Ming Zhang。2015年。用于自然图像去噪的深度卷积架构。载于2015年国际无线通信与信号处理会议（WCSP）。电气电子工程师学会，1–4。[460] Xintao Wang、Ke Yu、Shixiang Wu、Jinjin Gu、Yihao Liu、Chao Dong、Yu Qiao 和 Chen Change Loy。2018年。ESRGAN：增强型超分辨率生成对抗网络。载于欧洲计算机视觉会议（ECCV）研讨会论文集。0–0。[461] Yuting Wang。2021年。人工智能在中国新闻媒体中的应用。载于2021年第二届国际人工智能与信息系统会议。1–4。[462] Yexiang Wang、Yi Guo 和 Siqi Zhu。2020年。用于多领域对话状态跟踪的值归一化槽位注意力。载于2020年自然语言处理实证方法会议（EMNLP）论文集。3019–3028。[463] Yufei Wang、Zhe Lin、Xiaohui Shen、Scott Cohen 和 Garrison W Cottrell。2017年。骨架关键：通过骨架-属性分解生成图像描述。载于IEEE计算机视觉与模式识别会议论文集。7272–7281。[464] Duncan J Watts 和 Steven H Strogatz。1998年。“小世界”网络的集体动态。《自然》393卷第6684期（1998年），440–442。[465] Andrew Webster。2020年。Travis Scott的首场《堡垒之夜》演唱会如梦似幻、精彩绝伦。https://www.theverge.com/2020/4/23/21233637/travis-scott-fortnite-concert-astronomical-live-report（2020年）。[466] Li-Yi Wei 和 Marc Levoy。2000年。利用树结构向量量化快速合成纹理。载于第27届计算机图形学与交互技术年度会议论文集。479–488。[467] Tianyi Wei、Dongdong Chen、Wenbo Zhou、Jing Liao、Weiming Zhang、Lu Yuan、Gang Hua 和 Nenghai Yu。2022年。E2Style：提升StyleGAN反演的效率与效果。《IEEE图像处理汇刊》31卷（2022年），3267–3280。[468] Max Welling 和 Geoffrey E Hinton。2002年。一种用于平均场玻尔兹曼机的新学习算法。载于国际人工神经网络会议。施普林格，351–357。[469] 维基百科。2013年。《速度与激情7》（2015年）。https://en.wikipedia.org/wiki/Furious_7（2013年）。[470] Max Willens。2018年。彭博媒体有一个机器人在撰写故事摘要。https://digiday.com/media/bloomberg-media-robot-writing-story-summaries/（2018年）。

[471] Karen William。2021。《Voicemod 的 6 款最佳替代变声器》 [Windows/Mac/Android/iOS]。https://filme.imyfone.com/audio-edit/voicemod- alternative/（2021）。[472] Chenfei Wu、Lun Huang、Qianxi Zhang、Binyang Li、Lei Ji、Fan Yang、Guillermo Sapiro 和 Nan Duan。2021。Godiva：根据自然语言描述生成开放域视频。arXiv 预印本 arXiv:2104.14806（2021）。[473] Jibin Wu、Emre Yılmaz、Malu Zhang、Haizhou Li 和 Kay Chen Tan。2020。用于大词汇量自动语音识别的深度脉冲神经网络。《神经科学前沿》14（2020），199。[474] Jay Zhangjie Wu、Yixiao Ge、Xintao Wang、Weixian Lei、Yuchao Gu、Wynne Hsu、Ying Shan、Xiaohu Qie 和 Mike Zheng Shou。2022。Tune-A-Video：为文本生成视频对图像扩散模型进行单次调优。arXiv 预印本 arXiv:2212.11565（2022）。[475] Yonghui Wu、Mike Schuster、Zhifeng Chen、Quoc V Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、Yuan Cao、Qin Gao、Klaus Macherey 等。2016。谷歌神经机器翻译系统：弥合人工翻译与机器翻译之间的差距。arXiv 预印本 arXiv:1609.08144（2016）。[476] Zhirong Wu、Shuran Song、Aditya Khosla、Fisher Yu、Linguang Zhang、Xiaoou Tang 和 Jianxiong Xiao。2015。三维 ShapeNets：一种用于体积形状的深度表征。载于《IEEE计算机视觉与模式识别会议论文集》。1912–1920。[477] Weihao Xia、Yulun Zhang、Yujiu Yang、Jing-Hao Xue、Bolei Zhou 和 Ming-Hsuan Yang。2022。GAN反演综述。《IEEE模式分析与机器智能汇刊》（2022）。[478] Taihong Xiao、Jiapeng Hong 和 Jinwen Ma。2017。DNA-GAN：从多属性图像中学习解耦表示。arXiv 预印本 arXiv:1711.05415（2017）。[479] Zhisheng Xiao、Karsten Kreis 和 Arash Vahdat。2021。通过去噪扩散生成对抗网络应对生成式学习三难问题。arXiv 预印本 arXiv:2112.07804（2021）。[480] Haozhe Xie、Hongxun Yao、Xiaoshuai Sun、Shangchen Zhou 和 Shengping Zhang。2019。Pix2Vox：从单视图和多视图图像进行上下文感知的三维重建。载于《IEEE/CVF国际计算机视觉会议论文集》。2690–2698。[481] Junyuan Xie、Linli Xu 和 Enhong Chen。2012。使用深度神经网络进行图像去噪与图像修复。载于《神经信息处理系统进展》25（2012）。[482] Tian Xie、Xiang Fu、Octavian-Eugen Ganea、Regina Barzilay 和 Tommi Jaakkola。2021。用于周期性材料生成的晶体扩散变分自编码器。arXiv 预印本 arXiv:2110.06197（2021）。[483] Jun Xu、Haifeng Wang、Zheng-Yu Niu、Hua Wu、Wanxiang Che 和 Ting Liu。2020。用于开放域对话生成的对话图谱驱动策略学习。载于《第58届计算语言学协会年会论文集》。1835–1845。[484] Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel 和 Yoshua Bengio。2015。展示、关注与讲述：结合视觉注意力的神经图像描述生成。载于《国际机器学习会议》。PMLR，2048–2057。[485] Li Xu 和 Jiaya Jia。2010。用于稳健运动去模糊的两阶段核估计。载于《欧洲计算机视觉会议》。施普林格，157–170。[486] Li Xu、Jimmy S Ren、Ce Liu 和 Jiaya Jia。2014。用于图像反卷积的深度卷积神经网络。载于《神经信息处理系统进展》27（2014）。[487] Peng Xu、Xiatian Zhu 和 David A Clifton。2022。基于Transformer的多模态学习：综述。arXiv 预印本 arXiv:2206.06488（2022）。[488] Tao Xu、Pengchuan Zhang、Qiuyuan Huang、Han Zhang、Zhe Gan、Xiaolei Huang 和 Xiaodong He。2018。AttnGAN：利用注意力生成对抗网络进行细粒度文本到图像生成。载于《IEEE计算机视觉与模式识别会议论文集》。1316–1324。[489] Xianqiu Xu、Hongqing Liu、Yong Li 和 Yi Zhou。2016。利用RGB通道中的模糊核估计进行图像去模糊。载于《2016 IEEE国际数字信号处理会议（DSP）》。电气电子工程师学会，681–684。[490] Yangyang Xu、Yong Du、Wenpeng Xiao、Xuemiao Xu 和 Shengfeng He。2021。从连续性到可编辑性：利用连续图像反演生成对抗网络。载于《IEEE/CVF国际计算机视觉会议论文集》。13910–13918。[491] Ke Xue、Yifei Li 和 Hanqing Jin。2022。你如何看待人工智能？人工智能新闻主播形象对观看意愿的影响研究。《行为科学》12，11（2022），465。[492] Hemant Yadav 和 Sunayana Sitaram。2022。自动语音识别多语言模型综述。arXiv 预印本 arXiv:2202.12576（2022）。[493] Karan Yadav。2023。教育聊天机器人如何帮助学生和教师。http://www.eyeshenzhen.com/content/2022-02/08/content_*24921512*.htm（2023）。[494] Ravindra Yadav、Ashish Sardana、Vinay P Namboodiri 和 Rajesh M Hegde。2021。使用变分自编码器对无声视频进行语音预测。载于《ICASSP 2021—2021 IEEE声学、语音与信号处理国际会议（ICASSP）》。电气电子工程师学会，7048–7052。[495] Xinchen Yan、Jimei Yang、Kihyuk Sohn 和 Honglak Lee。2016。Attribute2Image：基于视觉属性的条件图像生成。载于《欧洲计算机视觉会议》。施普林格，776–791。[496] Li-Chia Yang、Szu-Yu Chou 和 Yi-Hsuan Yang。2017。MidiNet：用于符号域音乐生成的卷积生成对抗网络。arXiv 预印本 arXiv:1703.10847（2017）。[497] Shuoheng Yang、Yuxin Wang 和 Xiaowen Chu。2020。神经机器翻译深度学习技术综述。arXiv 预印本 arXiv:2002.07526（2020）。[498] Shiquan Yang、Rui Zhang 和 Sarah Erfani。2020。GraphDialog：将图知识整合到端到端任务型对话系统中。arXiv 预印本 arXiv:2010.01447（2020）。稿件已提交至美国计算机协会

[499] Tiancheng Yang 和 Shah Nazir。2022。人工智能赋能的音乐分类及其对游戏影响的全面概述。《Soft Computing》26，16（2022），7679–7693。[500] Xu Yang、Kaihua Tang、Hanwang Zhang 和 Jianfei Cai。2019。用于图像描述的场景图自动编码。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。10685–10694。[501] Xu Yang、Hanwang Zhang 和 Jianfei Cai。2019。学习为图像描述组合神经模块。载于《IEEE/CVF国际计算机视觉会议论文集》。4250–4260。[502] Yunyi Yang、Yunhao Li 和 Xiaojun Quan。2021。Ubar：迈向完全端到端的面向任务对话系统。载于《AAAI人工智能会议论文集》第35卷。14230–14238。[503] Jiaxuan You、Rex Ying、Xiang Ren、William Hamilton 和 Jure Leskovec。2018。GraphRNN：利用深度自回归模型生成逼真图。载于《国际机器学习会议》。PMLR，5708–5717。[504] Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan 等。2022。扩展自回归模型，以生成内容丰富的文本生成图像结果。arXiv 预印本 arXiv:2206.10789（2022）。[505] Ke Yu、Chao Dong、Liang Lin 和 Chen Change Loy。2018。通过深度强化学习构建图像复原工具链。载于《IEEE计算机视觉与模式识别会议论文集》。2443–2452。[506] Tianxiu Yu、Cong Lin、Shijie Zhang、Chunxue Wang、Xiaohong Ding、Huili An、Xiaoxiang Liu、Ting Qu、Liang Wan、Shaodi You 等。2022。人工智能助力敦煌文化遗产保护：项目与数据集。《国际计算机视觉杂志》130，11（2022），2646–2673。[507] Lu Yuan、Dongdong Chen、Yi-Ling Chen、Noel Codella、Xiyang Dai、Jianfeng Gao、Houdong Hu、Xuedong Huang、Boxin Li、Chunyuan Li 等。2021。Florence：一种用于计算机视觉的新型基础模型。arXiv 预印本 arXiv:2111.11432（2021）。[508] Li Yuan、Yunpeng Chen、Tao Wang、Weihao Yu、Yujun Shi、Francis EH Tay、Jiashi Feng 和 Shuicheng Yan。2021。Tokens-to-Token ViT：在ImageNet上从头训练视觉变换器。arXiv 预印本 arXiv:2101.11986（2021）。[509] Yuan Yuan、Siyuan Liu、Jiawei Zhang、Yongbing Zhang、Chao Dong 和 Liang Lin。2018。使用循环中循环生成对抗网络进行无监督图像超分辨率。载于《IEEE计算机视觉与模式识别会议研讨会论文集》。701–710。[510] Yi Yuan、Jilin Tang 和 Zhengxia Zou。2021。VANet：一种利用视图注意力引导的网络，用于从单视图和多视图图像进行三维重建。载于《2021 IEEE多媒体与博览会国际会议（ICME）》。电气电子工程师学会，1–6。[511] Xia Yuanjie。2022。人工智能手语主播亮相奥运会。http://www.eyeshenzhen.com/content/2022-02/08/content_*24921512*.htm（2022）。[512] Vladyslav Yushchenko、Nikita Araslanov 和 Stefan Roth。2019。用于视频生成的马尔可夫决策过程。载于《IEEE/CVF国际计算机视觉会议研讨会论文集》。0–0。[513] Shehtab Zaman、Ethan Ferguson、Cecile Pereira、Denis Akhiyarov、Mauricio Araya-Polo 和 Kenneth Chiu。2022。ParticleGrid：利用材料的三维表示实现深度学习。arXiv 预印本 arXiv:2211.08506（2022）。[514] Heiga Ze、Andrew Senior 和 Mike Schuster。2013。使用深度神经网络进行统计参数语音合成。载于《2013 IEEE声学、语音与信号处理国际会议》。电气电子工程师学会，7962–7966。[515] Heiga Zen。2015。统计参数语音合成中的声学建模——从隐马尔可夫模型到长短期记忆循环神经网络。（2015）。[516] Heiga Zen 和 Haşim Sak。2015。用于低延迟语音合成的带循环输出层的单向长短期记忆循环神经网络。载于《2015 IEEE声学、语音与信号处理国际会议（ICASSP）》。电气电子工程师学会，4470–4474。[517] Aeron Zentner。2022。应用创新：高等教育中的人工智能。SSRN 4314180（2022）。[518] Zheng-Jun Zha、Daqing Liu、Hanwang Zhang、Yongdong Zhang 和 Feng Wu。2019。用于细粒度图像描述的上下文感知视觉策略网络。《IEEE模式分析与机器智能汇刊》44，2（2019），710–722。[519] Chaoning Zhang、Chenshuang Zhang、Junha Song、John Seon Keun Yi、Kang Zhang 和 In So Kweon。2022。视觉及其他领域中用于自监督学习的掩码自编码器综述。arXiv 预印本 arXiv:2208.00173（2022）。[520] Chaoning Zhang、Kang Zhang、Trung X. Pham、Changdong Yoo 和 In-So Kweon。2022。双温度机制无需大量负样本即可助力对比学习：理解并简化MoCo。载于CVPR。[521] Chaoning Zhang、Kang Zhang、Chenshuang Zhang、Trung X Pham、Chang D Yoo 和 In So Kweon。2022。SimSiam如何在没有负样本的情况下避免崩溃？基于自监督对比学习的统一理解。载于国际学习表征会议。[522] Han Zhang、Tao Xu、Hongsheng Li、Shaoting Zhang、Xiaogang Wang、Xiaolei Huang 和 Dimitris N Metaxas。2017。StackGAN：利用堆叠生成对抗网络合成照片级真实感图像。载于《IEEE国际计算机视觉会议论文集》。5907–5915。[523] Han Zhang、Tao Xu、Hongsheng Li、Shaoting Zhang、Xiaogang Wang、Xiaolei Huang 和 Dimitris N Metaxas。2018。StackGAN++：利用堆叠生成对抗网络实现逼真图像合成。《IEEE模式分析与机器智能汇刊》41，8（2018），1947–1962。[524] Kunai Zhang、Da Huang 和 David Zhang。2017。一种基于图像清晰度的优化掌纹识别方法。《模式识别快报》85（2017），65–71。[525] Mingju Zhang、Lei Zhang、Yanfeng Sun、Lin Feng 和 Weiying Ma。2005。数码照片的自动裁剪。载于《2005 IEEE多媒体与博览会国际会议》。电气电子工程师学会，4–pp。[526] Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、Lei Zhang、Lijuan Wang、Yejin Choi 和 Jianfeng Gao。2021。VinVL：重新审视视觉语言模型中的视觉表征。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。5579–5588。[527] Weixing Zhang。2022。人工智能在机器人体育新闻写作中的应用与发展。载于《2022 IEEE第二届数据科学与计算机应用国际会议（ICDSCA）》。电气电子工程师学会，869–872。

[528] Wenlong Zhang、Yihao Liu、Chao Dong 和 Yu Qiao。2019年。RankSRGAN：结合排序器进行图像超分辨率的生成对抗网络。载于 IEEE/CVF国际计算机视觉会议论文集。3096–3105。[529] Xiaobo Zhang、Xiangchu Feng、Weiwei Wang、Shunli Zhang 和 Qunfeng Dong。2013年。用于图像去噪的基于梯度的维纳滤波器。《计算机与电气工程》39卷第3期（2013年），934–944。[530] Xuying Zhang、Xiaoshuai Sun、Yunpeng Luo、Jiayi Ji、Yiyi Zhou、Yongjian Wu、Feiyue Huang 和 Rongrong Ji。2021年。RSTNet：利用自适应注意力处理视觉与非视觉词语的图像描述。载于 IEEE/CVF计算机视觉与模式识别会议论文集。15465–15474。[531] Yichi Zhang、Zhijian Ou、Huixin Wang 和 Junlan Feng。2020年。一种面向半监督学习、具有潜在信念状态的概率式端到端面向任务的对话模型。arXiv 预印本 arXiv:2009.08115（2020年）。[532] Yizhe Zhang、Siqi Sun、Michel Galley、Yen-Chun Chen、Chris Brockett、Xiang Gao、Jianfeng Gao、Jingjing Liu 和 Bill Dolan。2019年。DialoGPT：用于对话响应生成的大规模生成式预训练。arXiv 预印本 arXiv:1911.00536（2019年）。[533] Zheng Zhang、Ryuichi Takanobu、Qi Zhu、MinLie Huang 和 XiaoYan Zhu。2020年。面向任务的对话系统的最新进展与挑战。《中国科学：技术科学》63卷第10期（2020年），2011–2027。[534] Mo Zhao、Gang Cao、Xianglin Huang 和 Lifang Yang。2022年。用于真实图像去噪的混合 Transformer-CNN。《IEEE 信号处理快报》（2022年）。[535] Qian Zhao、Hao Yang、Dongming Zhou 和 Jinde Cao。2022年。通过 CNN-Transformer 多尺度混合架构重新思考图像去模糊。《IEEE 仪器与测量汇刊》（2022年）。[536] Wentian Zhao、Yao Hu、Heda Wang、Xinxiao Wu 和 Jiebo Luo。2021年。利用多模态知识图谱增强实体感知图像描述。arXiv 预印本 arXiv:2107.11970（2021年）。[537] Rui Zhen、Wenchao Song、Qiang He、Juan Cao、Lei Shi 和 Jia Luo。2023年。人机交互系统：说话人头部生成综述。《电子学》12卷第1期（2023年），218。[538] Ce Zhou、Qian Li、Chen Li、Jun Yu、Yixin Liu、Guangjing Wang、Kai Zhang、Cheng Ji、Qiben Yan、Lifang He 等。2023年。预训练基础模型综合综述：从 BERT 到 ChatGPT 的发展历程。arXiv 预印本 arXiv:2302.09419（2023年）。[539] Hang Zhou、Yu Liu、Ziwei Liu、Ping Luo 和 Xiaogang Wang。2019年。通过对抗解耦的视听表征生成说话人脸。载于 AAAI人工智能会议论文集，第33卷。9299–9306。[540] Jingyuan Zhou、Chaktou Leong、Minyi Lin、Wantong Liao 和 Congduan Li。2022年。结合退化因素的图像复原任务自适应网络。载于 IEEE/CVF 冬季计算机视觉应用会议论文集。1–8。[541] Li Zhou、Jianfeng Gao、Di Li 和 Heung-Yeung Shum。2020年。共情型社交聊天机器人小冰的设计与实现。《计算语言学》46卷第1期（2020年），53–93。[542] Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason Corso 和 Jianfeng Gao。2020年。用于图像描述和视觉问答的统一视觉语言预训练。载于 AAAI人工智能会议论文集，第34卷。13041–13049。[543] Le Zhou、Qiu-Feng Wang、Kaizhu Huang 和 Cheng-Hung Lo。2019年。一种用于中国山水画文档的交互式生成方法。载于2019年国际文档分析与识别会议（ICDAR）。电气电子工程师学会，819–824。[544] Jiapeng Zhu、Yujun Shen、Deli Zhao 和 Bolei Zhou。2020年。用于真实图像编辑的域内 GAN 反演。载于欧洲计算机视觉会议。施普林格，592–608。[545] Jun-Yan Zhu、Philipp Krähenbühl、Eli Shechtman 和 Alexei A Efros。2016年。在自然图像流形上进行生成式视觉操控。载于欧洲计算机视觉会议。施普林格，597–613。[546] Qingfu Zhu、Lei Cui、Weinan Zhang、Furu Wei 和 Ting Liu。2018年。用于神经响应生成的检索增强对抗训练。arXiv 预印本 arXiv:1809.04276（2018年）。[547] Barret Zoph 和 Kevin Knight。2016年。多源神经机器翻译。arXiv 预印本 arXiv:1601.00710（2016年）。[548] Zhengxia Zou、Tianyang Shi、Shuang Qiu、Yi Yuan 和 Zhenwei Shi。2021年。风格化神经绘画。载于 IEEE/CVF计算机视觉与模式识别会议论文集。15689–15698。