从文本到图像的生成式 AI 模型现状与展望

甘晴 void（学号：202108010XXX）

摘要

近年来，生成式 AI 在文本到图像生成任务中取得了显著进展。通过结合自然语言处理与计算机视觉技术，这类模型能够生成高质量、语义准确的图像，广泛应用于艺术设计、广告制作、虚拟场景构建等领域。本文系统综述文本到图像生成模型的核心技术、代表性方法及应用现状，分析当前技术的优势与挑战，并对未来发展方向进行展望。

关键词：人工智能 深度学习 自然语言处理 计算机视觉 模型优化 无监督学习 图像合成

Abstract

In recent years, generative AI has made significant progress in the task of text-to-image generation. By combining natural language processing (NLP) and computer vision technologies, these models are capable of generating high-quality, semantically accurate images. They are widely used in fields such as art design, advertising, and virtual scene creation. This paper provides a systematic review of the core technologies, representative methods, and current applications of text-to-image generation models. It analyzes the advantages and challenges of current technologies and offers perspectives on future development directions.

Keywords: Artificial Intelligence, Deep Learning, Natural Language Processing, Computer Vision, Model Optimization, Unsupervised Learning, Image Synthesis

1. 引言

文本到图像生成是生成式人工智能的重要方向，通过结合自然语言处理与计算机视觉技术，能够根据文本描述生成高质量、语义一致的图像。这一技术在艺术创作、虚拟场景生成等领域展现出广泛的应用潜力。然而，现有方法在生成质量、语义理解与控制能力等方面仍存在不足，同时面临数据偏差与伦理风险等挑战。

本文旨在系统梳理文本到图像生成模型的技术进展与研究现状，分析现有方法的优势与局限，并探讨未来的发展方向。我将回顾模型的发展历程与技术演进，探讨模型设计的核心要素，总结研究现状与挑战，展望前沿趋势与未来方向，最后对全文进行总结。

2. 核心技术发展概述

2.1. 基础生成模型概述

自动编码器（Autoencoders）[1]是一种无监督学习模型，其核心思想是通过编码器将输入数据压缩为潜在表征（latent representation），并通过解码器从中重构原始数据，如图 1 所示。在文本到图像生成任务中，自动编码器可用于学习图像的潜在空间表示，为进一步的条件生成任务奠定基础。变种如变分自动编码器（Variational Autoencoders，VAEs）进一步增强了潜在空间的连续性，使生成的图像更加平滑、连贯。

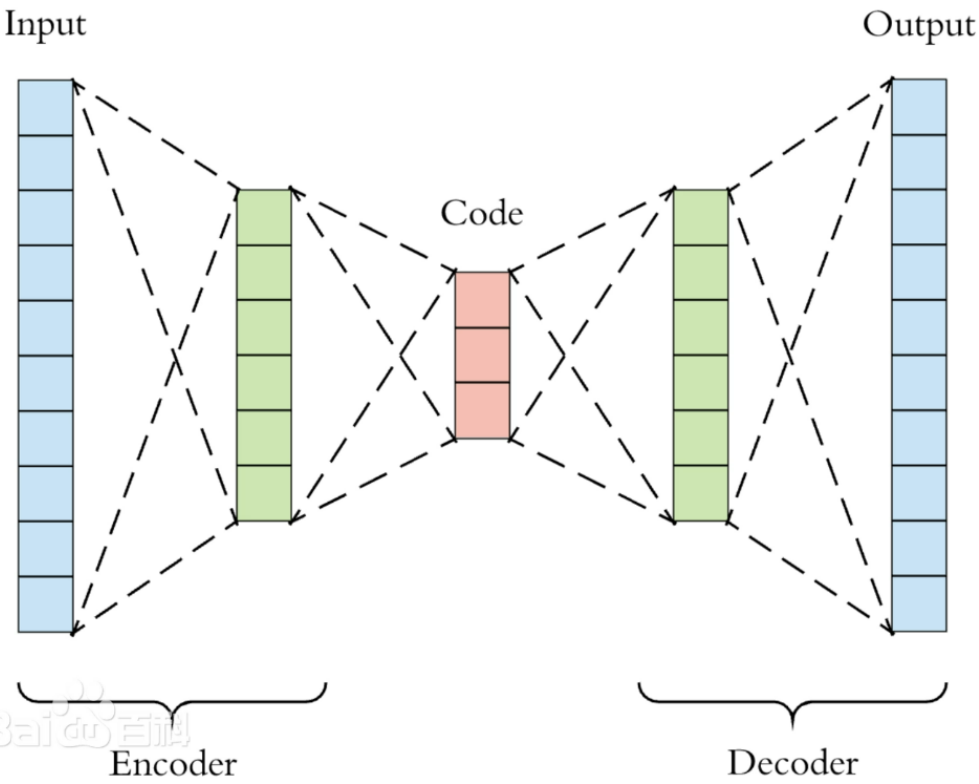


图 1

生成对抗网络（GANs）[2]由生成器（Generator）和判别器（Discriminator）组成，通过两者的博弈学习生成高质量的图像。GANs 在文本到图像生成任务中的关键作用在于能够生成视觉上逼真的图像，同时通过条件 GANs（cGANs）[3]引入文本作为条件信息，使生成的图像与输入文本语义一致。图 2 展示了 GAN 的效果。例如，AttnGAN[4]通过引入注意力机制增强文本与图像之间的语义对齐能力，是 GAN 在该领域的重要扩展。

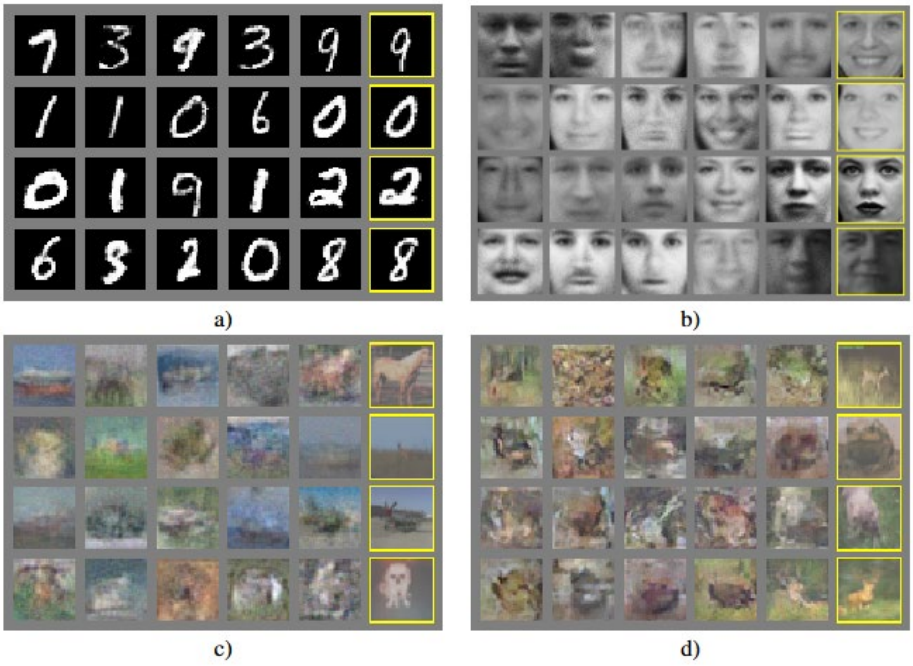
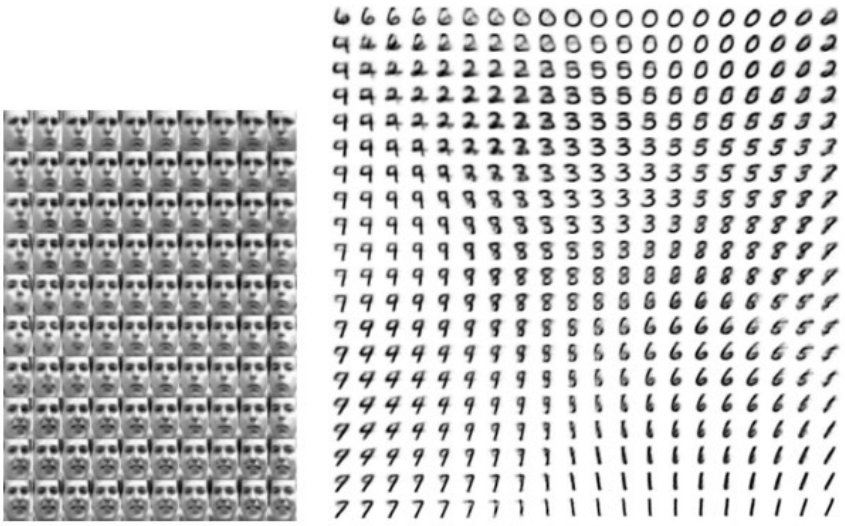


图 2

变分自动编码器（VAEs）是自动编码器的概率生成版本，通过在潜在空间上施加先验分布约束，生成的样本具有更好的分布连续性和生成多样性。其训练效果如图 3 所示。在文本到图像生成任务中，VAEs 通常作为图像生成器的一部分，与其他模型（如 GANs）结合使用，增强生成图像的多样性和语义一致性[5]。



(a) Learned Frey Face manifold　　(b) Learned MNIST manifold

图 3

自回归模型通过逐步生成序列的每一部分来建模生成过程，其典型代表包括 PixelCNN 和 PixelRNN[6]。这类模型在生成图像时能够很好地捕获局部结构特征，但生成速度较慢。尽管其在独立应用中有所局限，自回归模型的技术思想对后续的扩散模型和 Transformer 模型有重要启发作用[7]。

以上基础模型为文本到图像生成任务提供了不同的技术方向。自动编码器和 VAEs 负责潜在表征的学习，GANs 关注生成质量与语义一致性，自回归模型为生成过程建模提供了启示。这些技术的结合为更复杂的多模态生成模型奠定了基础，使得文本到图像生成任务在语义表达与视觉生成之间实现平衡[8]。

2.2 文本到图像生成模型的演进

2.2.1. 早期基于规则的方法

文本到图像生成最初基于规则和模板驱动方法，这些方法依赖对文本的显式解析，将其映射到预定义的图像元素。例如，Zhao 等（2013）提出了基于场景图的方法，将文本转化为场景描述，然后生成简单的合成图像[9]。尽管这些方法在特定领域具有一定实用性，但生成内容的多样性和复杂性较为有限，难以适应更为复杂的自然语言描述。

2.2.2. 基于统计学习的方法

随着统计学习的兴起，研究者引入特征映射技术以增强文本到图像生成的能力。Karpathy 和 Fei-Fei（2015）提出的 Deep Fragment Embeddings 方法通过文本和图像的嵌入对齐实现跨模态生成[10]。这些方法在捕获文本与图像语义相关性方面有所进步，但生成能力通常受限于训练数据集，缺乏创新性和通用性。

2.2.3. 深度学习驱动的多模态生成模型

（1） 基于生成对抗网络（GANs）的方法

生成对抗网络（GANs）的引入推动了文本到图像生成技术的快速发展。Reed 等（2016）首次使用条件 GAN（cGAN）实现文本到图像生成，将文本嵌入用作生成条件[11]。随后，StackGAN 引入了分阶段生成策略，从粗略图像逐步细化到高分辨率图像[12]。AttnGAN 则进一步引入了注意力机制，使模型能够聚焦于文本描述中的关键语义细节，从而生成更为精准的图像[13]。

（2） 基于变分自动编码器（VAEs）的方法

VAEs 在生成图像时表现出潜在空间的连续性优势。Yan 等（2016）提出的改进 VAE 方法在文本到图像生成中表现出较好的多样性[14]。

尽管生成质量不及 GANs，但结合 VAE 和 GAN 的混合模型（如 VAE-GAN）显著提升了生成效果，兼顾多样性与视觉质量。

（3） 基于扩散模型的生成方法

近年来，扩散模型（Diffusion Models）逐渐成为文本到图像生成的热门技术。Ramesh 等（2022）提出的 DALL·E 2 和 Saharia 等（2022）的 Imagen 使用扩散模型生成高质量图像[15][16]。这些模型通过逐步逆转噪声过程生成图像，解决了 GANs 训练不稳定的问题，并在生成复杂语义图像方面表现卓越。

（4） 基于 Transformer 的生成方法

Transformer 模型因其强大的多模态建模能力而在文本到图像生成中获得广泛应用。Brown 等（2021）提出的 CLIP 模型为文本和图像的跨模态对齐奠定了基础[17]。DALL·E 系列模型和 Stable Diffusion 则进一步结合扩散模型与 Transformer 的优势，在语义复杂性和生成质量上达到新高度[18]。

2.3 主要框架与技术

（1）条件生成对抗网络（cGANs）

条件生成对抗网络（Conditional GANs, cGANs）是文本到图像生成任务的重要基石，其技术核心在于通过条件信息（如文本嵌入）指导生成过程。其结构如图 4 所示。

cGANs 由生成器（Generator）和判别器（Discriminator）组成。生成器接收文本嵌入和随机噪声向量，生成与输入文本语义一致的图

像；判别器则判断生成图像的真实性和与条件的一致性。训练过程中，通过交替优化生成器和判别器的目标函数，cGANs 能够逐步生成更高质量的图像[19]。

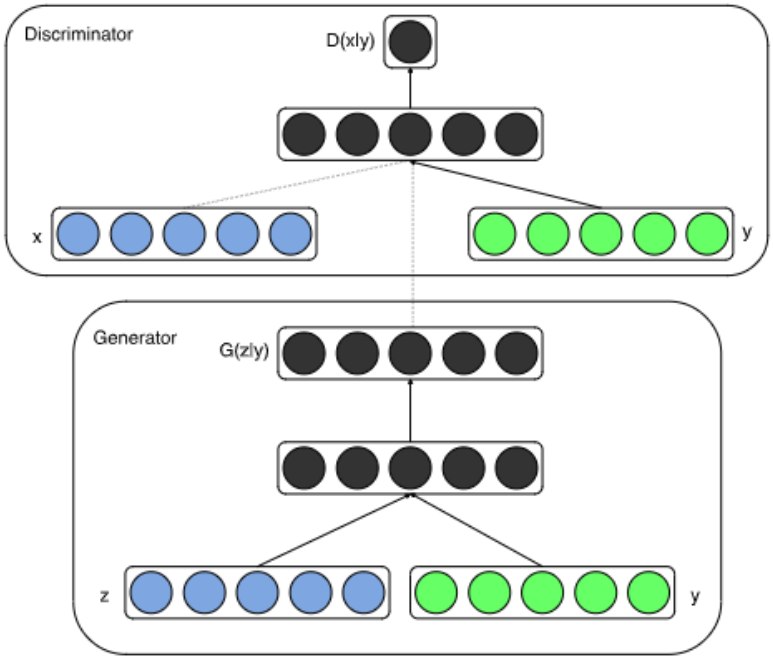


图 4

Reed 等（2016）首次将文本嵌入引入到 GAN 的生成过程中，通过训练 cGANs 实现从文本到图像的直接生成[20]。随后，StackGAN 提出分阶段生成策略，先生成粗略图像，再逐步优化细节，从而大幅提升了图像分辨率和质量[21]。AttnGAN 引入注意力机制，解决了文本与图像特征对齐不足的问题，使生成图像更精准地反映输入描述[22]。

（2）扩散模型（Diffusion Models）

扩散模型以其生成质量和训练稳定性在文本到图像任务中占据重要地位，其关键技术在于逐步逆转随机噪声过程生成图像。扩散模型的训练分为两个阶段：正向扩散过程：通过逐步向图像添加噪声，

将其转换为标准正态分布；反向生成过程：学习逆向去噪过程，通过条件信息逐步生成高质量图像。

Ramesh 等（2022）提出的 DALL·E 2 在扩散模型的基础上，通过引入 CLIP 图像-文本嵌入对，显著提升了图像生成的语义一致性和质量[23]。Saharia 等（2022）的 Imagen 进一步优化扩散模型，将语言理解和图像生成能力结合，生成的图像在细节和语义一致性上达到新的高度[24]。Stable Diffusion 则以高效的计算资源需求和开源特性获得广泛关注，通过潜在空间扩散优化生成效率[25]。部分生成的结果如图 5 所示。

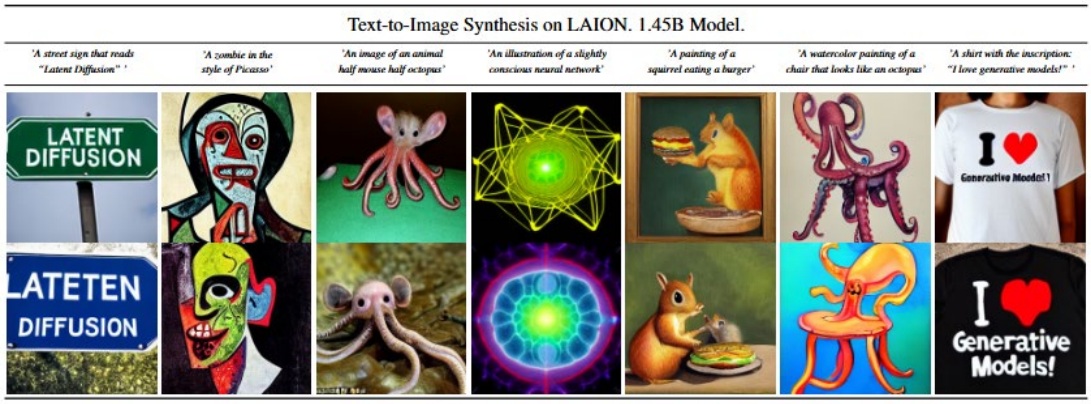


图 5

（3）多模态预训练模型（CLIP 等）

多模态预训练模型通过对文本和图像特征的联合学习，在文本到图像生成中起到了关键的桥梁作用。

CLIP（Contrastive Language - Image Pre-training）由 OpenAI 提出，通过对海量的图文对进行对比学习，训练出共享的多模态嵌入空间。模型采用两个分支架构：一个文本编码器和一个图像编码器，分别提取文本和图像特征。通过对齐不同模态特征，CLIP 实现了文本

和图像之间的语义匹配，为生成模型提供了强大的条件表示[26]。CLIP 的流程总结如图 6 所示。

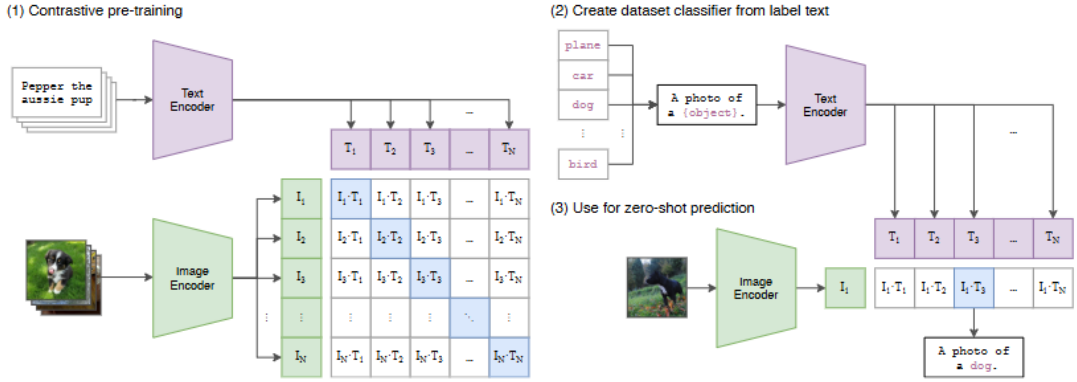


图 6

许多文本到图像生成模型结合了 CLIP 的特性。例如，DALL·E 2 利用 CLIP 嵌入作为生成条件，提升生成图像的语义对齐能力[23]。Stable Diffusion 通过潜在空间中的 CLIP 嵌入条件指导扩散过程，在生成效率和效果上均表现出色[25]。

3. 模型设计的核心要素

3.1 文本表征

在文本到图像生成模型中，文本表征是模型设计的核心要素之一。优质的文本表征能够准确捕捉输入文本的语义信息，并与图像表征进行高效对齐，从而指导生成器生成语义一致且视觉细腻的图像。本节讨论文本嵌入技术和多模态对齐的关键技术与进展。

（1）文本嵌入技术

文本嵌入技术通过将自然语言转化为连续的高维向量空间表示，为生成模型提供了语义信息的数值化表达。

早期的文本嵌入方法，如 Word2Vec 和 GloVe，通过统计词频和上下文信息学习词语的向量表示[27][28]。尽管这些方法在语义表示上有所突破，但其固定的嵌入无法适应文本的多义性和上下文依赖特性。

随着深度学习的发展，基于 Transformer 的模型（如 BERT 和 GPT）在文本表征上表现出显著优势。BERT（Devlin 等，2018）通过双向编码器捕捉上下文依赖特性，生成更为精准的文本表征[29]，如图 7 所示。GPT 系列模型则采用自回归方式，擅长生成任务的语言建模[30]。这些方法为文本到图像生成模型提供了丰富的语义信息，有助于捕捉复杂的文本描述。

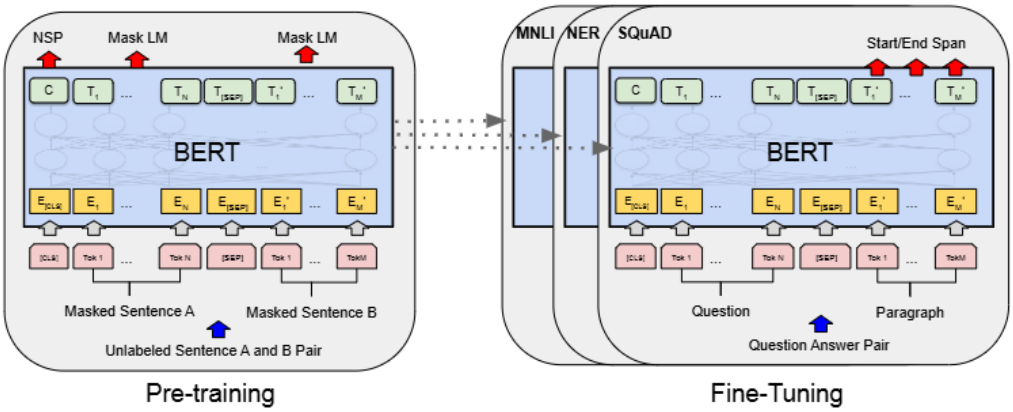


图 7

（2）多模态对齐技术

多模态对齐技术通过对文本和图像的语义嵌入进行对齐，使生成模型能够跨模态理解和处理信息。

CLIP 模型（Radford 等，2021）是多模态对齐的代表性成果。CLIP 使用对比学习策略，将文本和图像映射到共享嵌入空间中，通过最大化正样本的相似度和最小化负样本的相似度，实现高效的跨模态对齐[31]。CLIP 嵌入广泛应用于生成模型（如 DALL·E 2 和 Stable

Diffusion）中，用于引导生成过程并提升语义一致性[32][33]。

注意力机制在多模态对齐中扮演了重要角色。例如，Transformer 模型通过自注意力机制捕捉文本和图像特征之间的关联，使模型能够聚焦于与生成任务相关的语义信息[34]。AttnGAN 则通过在生成过程中动态分配注意力权重，使生成图像更符合文本描述[35]。

3.2 图像生成机制

（1）图像解码器的结构设计

传统图像解码器多采用基于卷积神经网络（CNN）的结构，利用上采样层（如反卷积或插值）逐步生成高分辨率图像。例如，GANs 家族中的生成器通常通过多层卷积模块和归一化技术生成清晰的图像[36]。此外，StackGAN 引入两阶段生成架构：第一阶段生成粗略图像，第二阶段通过细化模块增强细节[37]。

近年来，Transformer 在图像生成任务中的应用逐渐增多，其结构如图 8 所示。Imagen 等模型结合文本编码器和基于 Transformer 的图像解码器，通过多头注意力机制处理图像特征，使生成过程能够灵活适应复杂的输入描述[38]。Transformer 解码器的优势在于其全局建模能力，能够有效提升图像的语义一致性和细节表现力。

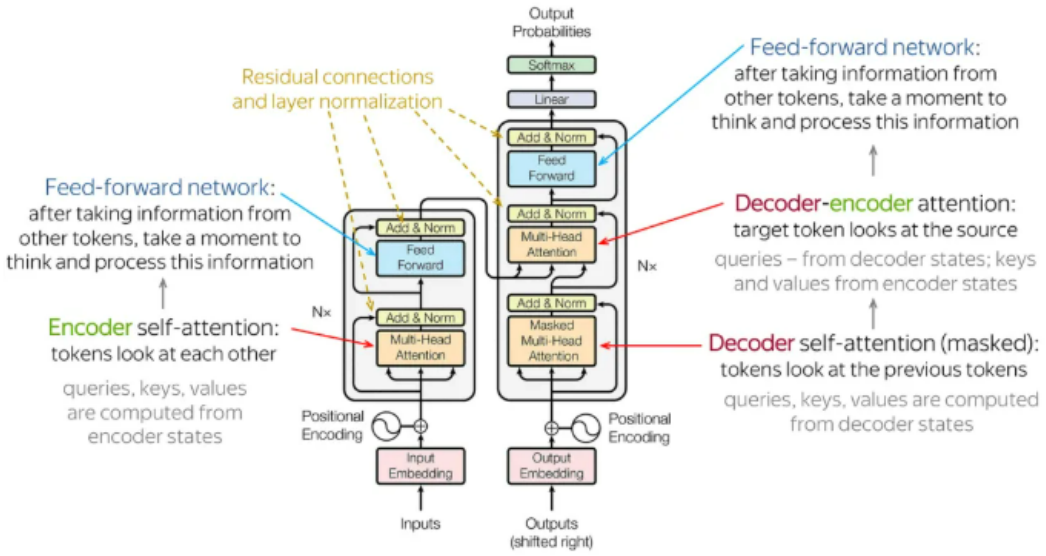


图 8

扩散模型通过反向去噪过程生成图像，显著提升了生成质量和多样性。Stable Diffusion 等模型将生成过程限制在潜在空间中，通过潜在变量的逐步解码实现高效生成，同时降低计算资源的需求[39]。这种方法能够保留图像全局语义，同时细化局部细节。

（2）生成质量评价

主观评价依赖于人类评审对生成图像的直观判断，包括图像的真实感、细节表现力和与输入文本的语义一致性。然而，主观评价由于耗时且依赖于评审者的个人经验，难以规模化应用[40]。

客观评价通过量化指标评估生成图像的质量，常用方法包括：

- 结构相似性指数（SSIM）：衡量生成图像与真实图像的结构相似性[41]。
- Frechet Inception Distance（FID）：评估生成图像和真实图像在潜在空间分布的差异，数值越小，生成质量越高[42]。

- CLIP Score：基于 CLIP 嵌入的语义一致性评估，衡量生成图像与输入文本的语义相关性[43]。

近年来，针对生成图像语义一致性的评价方法不断涌现。例如，AlignDraw 引入跨模态评价机制，通过比较图像与文本的语义相关性分数直接优化生成过程[44]。这种方法为进一步提升生成模型的性能提供了新的研究方向。

3.3 条件生成与控制性增强

条件生成与控制性增强是文本到图像生成模型的重要研究方向，其目标是提高生成图像的可控性，使模型能够准确响应输入条件并支持复杂语义的细粒度控制。

（1）条件生成机制

条件生成对抗网络（cGANs）通过在生成器和判别器中加入条件向量（如文本嵌入或标签）增强生成的语义一致性。例如，Mirza 和 Osindero（2014）提出的 cGAN 框架，首次在生成过程中引入条件变量，使生成器能够根据输入文本生成与之匹配的图像[45]。进一步的改进如 StackGAN 和 AttnGAN，通过分层生成或注意力机制提升生成图像的细节质量和语义一致性[46][47]。

扩散模型在条件生成中的表现尤为突出。例如，Imagen 和 Stable Diffusion 等模型将文本条件编码为潜在向量，通过在反向去噪过程中融入条件信息实现高质量的图像生成[48][49]。特别是 Stable Diffusion 利用 CLIP 文本嵌入作为条件向量，为生成过程

提供了细粒度的语义指导，并通过潜在空间的灵活性提升了生成图像的多样性和细节质量[50]。

（2）控制性增强方法

细粒度语义控制技术通过更精细的条件表示，使生成图像能够捕捉复杂的语义细节。例如，ControlNet 提出了在扩散模型框架下添加额外的条件路径（如边缘检测或姿态信息），实现对生成图像的精确控制[51]。此外，方法如 SEGA（Selective Guidance for Diffusion Models）在生成过程中动态调整条件权重，从而在特定细节上进行控制[52]。

为支持多种条件的组合控制，研究者提出了多模态条件生成方法。例如，DALL·E 2 能够结合文本和图像作为条件，为生成过程提供更全面的指导[53]。此外，Prompt-to-Prompt 方法允许用户通过修改文本条件的方式实时调整生成图像的内容，实现对生成结果的多样化控制[54]。

（3）应用场景与未来方向

条件生成与控制性增强技术广泛应用于艺术创作、产品设计和医学影像生成等领域。例如，通过控制生成结果的风格和细节，设计师可以生成特定需求的艺术作品[55]。未来研究可以在以下方面深入探索：条件控制的通用性：开发适应多种条件类型的生成框架。交互式控制增强：支持用户实时调整生成过程中的细节内容。生成安全性与公平性：在复杂条件下确保生成结果的语义一致性和公平性。

4.研究现状与挑战

近年来，文本到图像生成模型在生成质量和效率上取得了显著进展，但仍存在一些挑战和局限性。本节从生成质量的提升和效率优化两个方面探讨当前研究的进展与不足。

4.1 模型性能提升

（1）生成质量的提升

高质量图像生成依赖于强大的解码器和训练策略。例如，StyleGAN 系列通过多尺度特征提取和渐进式训练，实现了高分辨率图像的细节增强[56]。扩散模型（如 Imagen 和 Stable Diffusion）进一步优化了生成过程，结合高质量的文本嵌入（如 CLIP）实现了语义一致性和细节表现力的提升[57][58]。

然而，生成质量在某些场景中仍存在不足：复杂场景生成：当输入文本描述高度复杂的场景时，模型容易丢失细节或生成语义冲突的内容[59]。多样性和平衡性：某些生成模型在优化质量的同时可能牺牲多样性，导致生成图像的分布与真实图像分布不一致[60]。

文本到图像生成任务的核心在于确保生成图像与输入文本的语义高度一致。DALL·E 2 通过引入层次化的潜在空间和文本条件嵌入，在生成过程中实现了更高的语义一致性[61]。此外，AlignDraw 等模型利用交叉注意力机制在多模态对齐中取得了良好的效果[62]。

尽管如此，语义一致性仍面临挑战：长文本描述的处理：当文本描述较长或包含复杂语义时，生成图像可能无法准确捕捉所有关键信

息[63]。

（2）效率优化的进展

生成模型通常需要大量的计算资源和训练时间。为提升训练效率，研究者提出了一些改进：小样本学习：通过少量标注样本进行高效训练，Meta-Diffusion 提供了一种潜在空间小样本优化方法[64]。参数共享和模型压缩：研究者通过权重共享和量化技术大幅降低模型的存储和计算需求，例如 DreamBooth 中使用的高效微调方法[65]。

扩散模型的推理过程通常需要多步去噪迭代，导致生成时间较长。最近的研究通过以下方式优化推理效率：加速去噪过程：DDIM（Denoising Diffusion Implicit Models）通过去噪步骤的简化实现了快速推理[66]。高效潜在空间生成：Stable Diffusion 通过在潜在空间中进行计算，显著减少了内存和时间消耗[67]。

（3）当前不足与未来方向

尽管性能不断提升，当前方法在以下方面仍有改进空间：

- 生成效率与质量的权衡：如何在生成高质量图像的同时保持较低的计算资源需求是一个重要挑战[68]。
- 跨模态语义理解：提高文本和图像的深度语义对齐能力仍然是未来研究的重点[69]。
- 通用化与泛化能力：当前模型往往对特定场景优化，未来需要开发适应多样化场景的通用生成框架[70]。

4.2 数据与训练

文本到图像生成模型的性能很大程度上取决于所使用的数据集的规模、质量和多样性。然而，数据集中的偏差问题也可能对生成结果产生负面影响。

（1）数据集规模与质量

在文本到图像生成任务中，大规模、高质量的多模态数据集是模型训练的基础。例如，COCO、ImageNet 和 OpenAI 的 LAION 数据集，广泛用于训练和评估多模态生成模型[71]。这些数据集提供了大量标注的图像-文本对，使得模型能够学习从文本描述中生成对应的图像。高质量的数据集通常具备清晰的文本描述和多样化的图像内容，有助于提高生成图像的多样性和真实感。其中，图 9 为部分 COCO 数据集。

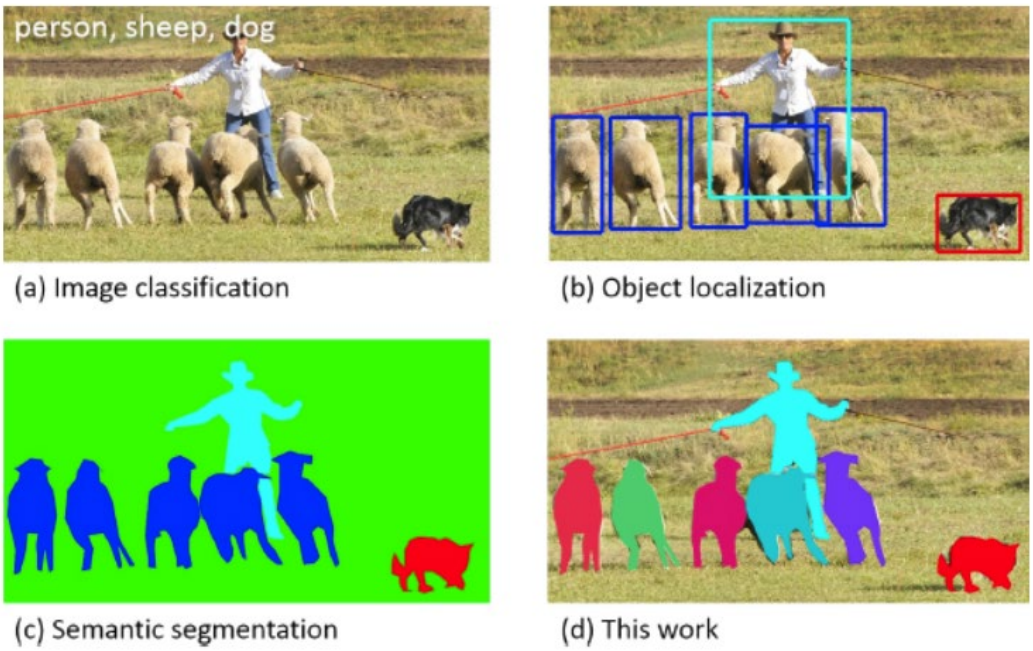


图 9

然而，数据集的质量并非总是理想。文本与图像之间的对齐问题可能导致模型无法准确捕捉语义。例如，某些图像描述可能存在歧义，

导致生成图像的多样性受限[72]。此外，训练集中的噪声和错误标注也可能影响模型的学习效果[73]。

（2）数据集偏差问题

数据集的偏差是影响生成图像质量的关键因素之一。如果训练数据中的图像和文本描述存在固有的偏差，生成模型可能会在生成过程中复制这些偏差。例如，如果训练集主要包含某些群体或文化的图像，模型可能会在生成图像时缺乏多样性和包容性[74]。为了减轻这种偏差问题，研究者提出了数据增强和多样化技术，如对抗性训练和公平性约束，以确保生成模型的公平性和普适性[75]。

4.3 计算资源与成本

随着模型规模的不断增加，训练大型文本到图像生成模型需要大量的计算资源和时间，这给研究和开发带来了巨大的挑战。如何优化计算资源的使用，降低训练和推理成本，成为了当前研究的一个重要方向。

（1）大规模训练的资源需求

训练如 DALL·E 2、Stable Diffusion 等大型模型需要庞大的计算资源。例如，DALL·E 2 的训练依赖于数百个高性能 GPU 和数 TB 的数据存储资源[76]。对于大规模模型而言，计算时间和内存消耗是不可忽视的瓶颈。每次迭代的训练和推理过程可能需要几小时到几天的计算时间，尤其是当模型需要处理大规模数据集时。

（2）计算资源优化方法

为了降低计算资源的需求，研究者提出了多个优化方法：

- 混合精度训练：通过使用低精度（如 16 位浮点数）训练，能够显著减少内存占用和计算量[77]。
- 模型压缩与知识蒸馏：通过模型压缩技术，可以将大型模型的计算量压缩到较小规模，同时保持相对较高的性能。知识蒸馏方法将大模型的知识转移到较小的学生模型中，从而降低计算成本[78]。
- 分布式训练：通过分布式训练，将训练任务分配到多个计算节点，能够大幅度缩短训练时间并提高资源的利用效率[79]。

这些方法的应用显著减少了训练时间和计算资源的消耗，推动了文本到图像生成模型的可扩展性。

4.4 应用与伦理问题

随着文本到图像生成模型在各行各业的广泛应用，伦理和法律问题也逐渐浮出水面。如何确保这些模型在实际应用中符合道德标准，避免对社会产生负面影响，是当前亟需解决的问题。

（1）应用领域

文本到图像生成技术在多个领域展现出了巨大的潜力，包括：

- 创意产业：如艺术创作、广告设计和虚拟现实内容生成[80]。
- 教育与科研：通过生成与教育内容相关的图像或示意图，增强学习和研究的互动性和可视化效果。

- 医疗影像：在医学领域，生成模型有望用于生成医学影像，辅助医生诊断和治疗[81]。
- 商业与营销：通过根据文本描述生成广告素材、产品设计图等，降低创意成本，提高营销效率[82]。

（2）生成内容的伦理风险

文本到图像生成模型在应用过程中可能带来一些伦理风险，主要表现在以下几个方面：

- 虚假信息与误导性内容：生成的图像可能被用于虚假宣传、造谣和误导公众。例如，生成的假冒产品图像可能会被用于欺诈，或者通过虚构的图像误导公众[83]。
- 偏见与刻板印象：如果训练数据存在性别、种族或文化偏见，模型可能会生成带有偏见的图像，甚至可能无意中强化这些偏见[84]。例如，某些生成模型可能会生成性别或种族刻板化的图像，影响社会公平和多样性[85]。
- 隐私问题：生成技术也可能被恶意用于制造侵犯隐私的图像，特别是在深度伪造（Deepfake）领域，生成的虚假图像可能被用于恶意操控[86]。

（3）伦理规范与解决方案

为了应对这些伦理问题，研究者和开发者提出了一些规范和解决方案：

- 公平性与多样性约束：通过对抗性训练和公平性优化，减少模型的偏见和不公平性[87]。

- 透明度与可追溯性：确保生成模型的使用透明化，允许对生成过程进行追溯，保证生成内容的合法性和道德性[88]。
- 生成内容的水印技术：为生成内容添加水印，以便于追踪和识别其来源，防止其被用于不当用途[89]。

5. 前沿研究与趋势

近年来，文本到图像生成技术在多个方面取得了显著进展。特别是无监督与半监督学习的探索、跨模态生成的扩展以及强化学习与人类反馈的结合，成为当前的前沿研究方向。这些进展不仅推动了生成模型的应用多样性，也为模型的性能提升提供了新的思路。

5.1 无监督与半监督学习在文本到图像生成中的探索

无监督学习和半监督学习为解决大规模标注数据稀缺问题提供了新的可能性，特别是在多模态生成任务中。通过利用无标注数据的潜力，模型能够在不依赖大规模人工标注数据的情况下，从数据中学习更有意义的表示。

（1）无监督学习

近年来，研究者们探索了基于无监督学习的文本到图像生成模型，这类模型不依赖于文本-图像配对的训练数据，而是通过其他方式从大量的图像数据中自动生成语义丰富的文本描述。例如，UNet 架构与生成对抗网络（GANs）结合，通过图像的自我表示学习，在无监督条件

下生成更高质量的图像[90]。一些研究还使用对比学习方法（如 SimCLR 和 MoCo）在无监督环境下进行视觉特征学习[91]。这些方法不仅提高了生成图像的多样性，也加强了模型的鲁棒性。

（2）半监督学习

在半监督学习中，文本-图像对的标注数据相对较少，而未标注数据则占据了大部分。通过利用少量的标注数据和大量的未标注数据，模型能够学习到更加丰富的文本与图像之间的关系。例如，CLIP（Contrastive Language-Image Pretraining）通过对比学习对文本和图像进行联合训练，在不完全标注的情况下，增强了图像与文本的表示能力[92]。此外，基于生成模型的半监督方法，如基于 VAE（变分自编码器）和 GAN 的组合，已被广泛应用于多模态生成任务，提供了一种解决数据标注稀缺问题的有效途径[93]。

5.2 跨模态生成任务的扩展

跨模态生成是指在多个模态之间进行生成任务的扩展，包括文本、图像、音频、视频等。随着多模态模型的快速发展，跨模态生成已经成为一个重要的研究方向，尤其是在自然语言处理与计算机视觉的结合方面，许多研究探索了如何在多个模态之间进行联动生成。

（1）文本、图像、音频的联动生成

多模态生成不仅限于文本与图像之间的转换，音频和图像的联动生成也是一个前沿方向。例如，研究者提出了音频到图像的生成模型，将语音描述转换为图像，这在自动视频生成、虚拟助手等领域具有重要

应用[94]。近年来，基于 Transformer 架构的多模态学习方法（如 M4C）逐渐成为主流，通过多模态联合表示实现不同模态之间的协同生成[95]。

（2）视频生成与扩展

随着计算资源的不断增强，视频生成任务逐渐成为跨模态生成的重要拓展。生成模型不仅要处理图像的静态信息，还需要理解视频中的动态变化和时间关联。例如，基于文本描述的视频生成模型通过引入时间维度，使得生成过程能够处理时序信息，生成更符合自然规律的视频内容[96]。此外，VQ-VAE-2 等变分自编码器模型在视频生成中的应用，也为生成高质量视频提供了新的技术路径[97]。

5.3 强化学习与人类反馈的结合（如 RLHF）

强化学习与人类反馈（RLHF）结合已经成为提升生成模型质量的重要手段，特别是在解决生成内容的质量和控制性方面。RLHF 通过引入人类反馈，使得模型能够根据用户需求进行个性化优化，进一步提升生成内容的质量和多样性。RLHF 的一般流程如图 10 所示。

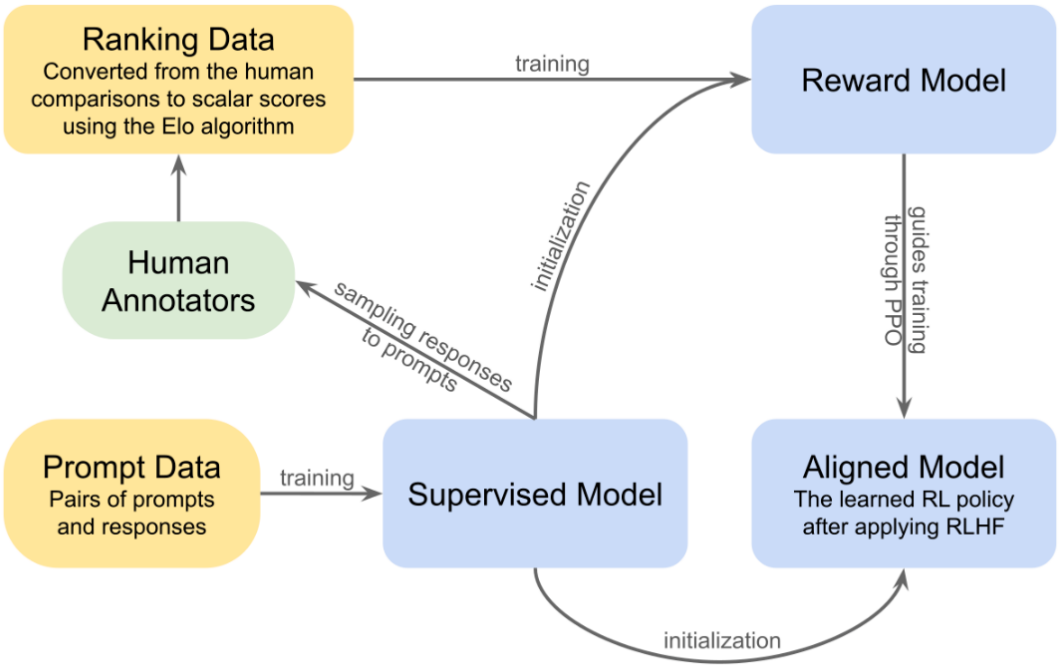


图 10

（1）RLHF 的基本概念与应用

RLHF 结合了传统的强化学习与人类的评价反馈，使得模型能够在训练过程中更加注重人类偏好。例如，OpenAI 的 ChatGPT 就使用了人类反馈来训练对话生成模型，通过奖励函数调整生成过程中的决策[98]。在文本到图像生成领域，RLHF 可以帮助模型学习如何生成更符合用户需求的图像。例如，Stable Diffusion 通过优化用户对生成图像的反馈，逐步提高了生成图像的质量和多样性[99]。

（2）RLHF 在多模态生成中的应用

RLHF 不仅在单一模态的生成任务中取得了成功，还在跨模态生成任务中发挥了重要作用。例如，模型通过人类反馈学习在生成图像时如何更好地理解和执行文本描述，或者在生成视频时如何更精确地控制时序信息[100]。结合 RLHF 的方法能够更好地处理生成图像的细粒度控

制，提高生成内容的控制性和可定制性[101]。

6.结论

本文系统梳理了文本到图像生成模型的技术发展与应用现状，分析了当前研究的技术难点与潜在风险。未来，应重点关注模型的生成质量、语义一致性与控制能力，同时解决数据偏差与伦理问题。随着技术的持续发展，文本到图像生成模型将在更多领域发挥重要作用。

7.参考文献

[1] Kingma, D.P., & Welling, M. Auto-Encoding Variational Bayes. ICLR, 2014.
[2] Goodfellow, I., et al. Generative Adversarial Nets. NeurIPS, 2014.
[3] Mirza, M., & Osindero, S. Conditional Generative Adversarial Nets. NeurIPS, 2014.
[4] Xu, T., et al. AttnGAN: Fine-Grained Text to Image Generation with Attention Mechanism. CVPR, 2018.
[5] Kingma, D.P., & Welling, M. Auto-Encoding Variational Bayes. ICLR, 2014.
[6] Salimans, T., et al. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. NeurIPS, 2017.
[7] van den Oord, A., et al. Pixel Recurrent Neural Networks. ICML, 2016.
[8] Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision. NeurIPS, 2021.
[9] Zhao, J., et al. From text descriptions to visual denotations: Generating visual scenarios from natural language (2013).
[10]Karpathy, A., & Fei-Fei, L. Deep Fragment Embeddings for Text-to-Image Alignment (2015).
[11]Reed, S., et al. Generative adversarial text to image synthesis (2016).
[12]Zhang, H., et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks (2017).
[13]Xu, T., et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks (2018).
[14]Yan, X., et al. Attribute2Image: Conditional image generation from visual attributes (2016).
[15]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2) (2022).
[16]Saharia, C., et al. Imagen: Photorealistic text-to-image diffusion models

with deep language understanding (2022).
[17]Brown, T., et al. CLIP: Connecting text and images through natural language understanding (2021).
[18]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion) (2022).
[19]Goodfellow, I., et al. Generative Adversarial Nets (2014).
[20]Reed, S., et al. Generative adversarial text to image synthesis (2016).
[21]Zhang, H., et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks (2017).
[22]Xu, T., et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks (2018).
[23]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2) (2022).
[24]Saharia, C., et al. Imagen: Photorealistic text-to-image diffusion models with deep language understanding (2022).
[25]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion) (2022).
[26]Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) (2021).
[27]Mikolov, T., et al. Efficient Estimation of Word Representations in Vector Space (2013).
[28]Pennington, J., et al. GloVe: Global Vectors for Word Representation (2014).
[29]Devlin, J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018).
[30]Radford, A., et al. Language Models are Few-Shot Learners (GPT-3) (2020).
[31]Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) (2021).
[32]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2) (2022).
[33]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion) (2022).
[34]Vaswani, A., et al. Attention Is All You Need (2017).
[35]Xu, T., et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks (2018).
[36]Goodfellow, I., et al. Generative Adversarial Nets (2014).
[37]Zhang, H., et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks (2017).
[38]Saharia, C., et al. Imagen: Photorealistic text-to-image diffusion models with deep language understanding (2022).
[39]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion) (2022).
[40]Wang, Z., et al. Image Quality Assessment: From Error Visibility to Structural Similarity (2004).

[41]Salimans, T., et al. Improved Techniques for Training GANs (2016).
[42]Heusel, M., et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (2017).
[43]Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) (2021).
[44]Mansimov, E., et al. Generating Images from Captions with Attention (2015).
[45]Mirza, M., & Osindero, S. Conditional Generative Adversarial Nets (2014).
[46]Zhang, H., et al. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks (2017).
[47]Xu, T., et al. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks (2018).
[48]Saharia, C., et al. Imagen: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (2022).
[49]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion) (2022).
[50]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2). NeurIPS, 2022.
[51]Zhang, X., et al. ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models (2023).
[52]Avrahami, O., et al. SEGA: Selective Guidance for Diffusion Models (2023).
[53]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2) (2022).
[54]Hertz, A., et al. Prompt-to-Prompt Image Editing with Cross Attention Control (2022).
[55]Gal, R., et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion (2023).
[56]Karras, T., et al. A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN) (2019).
[57]Saharia, C., et al. Imagen: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (2022).
[58]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion) (2022).
[59]Xu, T., et al. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks (2018).
[60]Brock, A., et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis (2018).
[61]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2) (2022).
[62]Mansimov, E., et al. Generating Images from Captions with Attention (2015).
[63]Nichol, A., et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models (2021).
[64]Xiao, S., et al. Meta-Diffusion: Few-Shot Image Synthesis Using Diffusion Models (2022).
[65]Ruiz, N., et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models

for Subject-Driven Generation (2023).
[66]Song, J., et al. Denoising Diffusion Implicit Models (DDIM) (2021).
[67]Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models (2022).
[68]Ho, J., et al. Denoising Diffusion Probabilistic Models (2020).
[69]Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) (2021).
[70]Zhang, H., et al. Cross-Modal Pre-training with Contrastive and Masked Autoencoders (CoMAE) (2023).
[71]Lin, T., et al. Microsoft COCO: Common Objects in Context (2014).
[72]Wu, L., et al. Image-Text Pretraining with Multi-Granular Contrastive Learning (2021).
[73]Johnson, J., et al. Image Generation from Captions with Attention (2015).
[74]Buolamwini, J., & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (2018).
[75]Wang, X., et al. Fairness in Machine Learning: A Survey (2020).
[76]Ramesh, A., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2) (2022).
[77]Micikevicius, P., et al. Mixed Precision Training (2018).
[78]Hinton, G., et al. Distilling the Knowledge in a Neural Network (2015).
[79]Li, Z., et al. Large-Scale Distributed Deep Learning: A Survey (2020).
[80]Elgammal, A., et al. CAN: Creative Adversarial Networks for Artistic Image Generation (2017).
[81]McKinney, S., et al. International Evaluation of Breast Cancer Screening with Deep Learning (2020).
[82]Chen, M., et al. Generating Advertising Content with Neural Networks (2019).
[83]Cummings, M., et al. The Ethics of Artificial Intelligence: A Survey (2020).
[84]Caliskan, A., et al. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases (2017).
[85]Binns, R. Fairness in Machine Learning: A Survey (2020).
[86]Westerlund, M. The Ethics of Deepfake Technology: A Review (2021).
[87]Zhang, L., et al. Fairness Constraints in Generative Models (2020).
[88]Sandvig, C., et al. Ethical Guidelines for AI in Journalism (2019).
[89]Dolgov, I., et al. Watermarking Deep Neural Networks (2019).
[90]Liu, X., et al. Self-supervised Learning for Text-to-Image Generation: A Unified Framework (2021).
[91]Chen, X., et al. SimCLR: A Simple Framework for Contrastive Learning of Visual Representations (2020).
[92]Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) (2021).
[93]He, Y., et al. Semi-supervised Text-to-Image Synthesis with GANs (2021).
[94]Chen, Y., et al. Audio-Visual Scene Analysis with Self-Supervised

Learning (2022).
[95]Su, P., et al. M4C: Multimodal Modeling with Contrastive Learning for Visual Question Answering (2021).
[96]Xu, T., et al. Video-to-Text: Generating Video Descriptions from Textual Inputs (2021).
[97]Razavi, A., et al. Generating Diverse High-Fidelity Images with VQ-VAE-2 (2021).
[98]Christiano, P., et al. Deep Reinforcement Learning from Human Preferences (2021).
[99]Rombach, R., et al. Stable Diffusion: A Latent Text-to-Image Diffusion Model (2022).
[100] Li, X., et al. Text-to-Image Generation with Reinforcement Learning and Human Feedback (2022).
[101] Stiennon, N., et al. Learning to Summarize with Human Feedback (2022).