A Comprehensive Survey on Diffusion Models: Structure, Variants, Applications, and Optimizations

李宇哲 PB21111653

摘要—扩散模型近年来在生成模型领域引起了广泛关注，其独特的生成机制和优秀的性能使其成为图像、文本、音频等多种应用的热门选择。于此，我选择对 Diffusion 模型展开调研。

本文对扩散模型进行了全面的调研，首先介绍了其基本结构和数学原理，阐明了与传统生成模型的比较。其次，我们探讨了扩散模型的各种变种，分析了每种变体的优势和适用场景。然后，重点回顾了扩散模型在实际应用中的最新进展，包括图像生成、声音合成和文本生成等领域的案例研究。最后，本文总结了当前优化策略的研究进展，讨论了如何提升扩散模型的训练效率和生成质量。

通过系统化的梳理和分析，我希望我能更理解 Diffusion 模型，并在这个方向上做一些有价值的工作。

***Index Terms*—Diffusion, GAN, Generative Model, VAE**

I. Introduction

随着深度学习的快速发展，生成模型在多个领域展现出了强大的潜力，尤其是在图像、文本和音频生成等任务中。近年来，扩散模型作为一种新兴的生成模型，凭借其独特的生成机制和优秀的性能，逐渐成为研究的热点。与传统的生成对抗网络（GAN）[4] 和变分自编码器（VAE）[7] 相比，扩散模型在生成质量和稳定性上展现出了明显的优势，吸引了众多研究者的关注。

扩散模型的核心思想是通过逐步引入噪声来训练模型，使其能够学习到数据的潜在分布。在训练阶段，模型通过反向扩散过程去除噪声，从而生成高质量的数据。这一过程不仅具有强大的理论基础，还能够有效地处理复杂的生成任务。

本文旨在对扩散模型进行全面的调研，首先介绍其基本结构和数学原理，帮助读者理解其与传统生成模型的区别。接着，本文将探讨扩散模型的各种变种，分析其各自的特点与应用场景。同时，我们将回顾扩散模型在不同领域的实际应用，包括图像生成、声音合成和文本生成等，展示其广泛的适用性和潜在的影响力。最后，本文还将总结当前针对扩散模型的优化策略，探讨如何进一步提升其性能和效率。

通过对扩散模型的系统性梳理和深入分析，我希望能让我更深入的理解 Diffusion 模型，以便更好地对 Sora 等 text-to-video 等模型进行研究。

II. Diffusion Structure

生成模型在图像合成、数据增强等任务中扮演着重要角色，其中 GAN（生成对抗网络）以其高质量的生成效果在过去几年中引领了发展。然而，GAN 存在模式崩溃（mode collapse）等问题，导致生成样本的多样性不足。为了解决这些挑战，Berkeley 的 Jonathan Ho 等人提出了 **Denoising Diffusion Probabilistic Models (DDPM)**[6]，这种模型通过一个逐步的马尔可夫过程实现数据生成。

与传统的生成模型相比，扩散模型在理论上具有更好的稳定性和生成多样性。实验结果表明，在图像生成任务中，DDPM 的效果能够超越 GAN 等传统方法。DDPM 的引入为生成模型开辟了新的思路，并为未来的研究提供了重要的基础。

A. Forward and Reverse Processes

扩散模型的核心结构包括两个主要过程：前向扩散过程和反向去噪过程。扩散模型通过将真实数据逐步添加噪声，形成一个马尔科夫链。反向过程从噪声分布逐步恢复原始数据。而模型训练的目标是最小化均方误差损失，通过优化这个损失，模型学习在不同时间步去除噪声。

a) Forward Process: 前向扩散过程是一个马尔可夫链，它逐步将噪声添加到数据样本 x_0 中，生成一系列中间状态 $x_1, x_2, \ldots, x_T$。该过程可以描述为：

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

其中 α_t 控制每一步的噪声添加程度，使得在较长的步骤后，数据接近标准正态分布。

b) Reverse Process: 反向去噪过程则是通过学习参数化的神经网络模型 p_θ，逐步从噪声中恢复数据样本 x_0。该过程由以下条件分布定义：

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

反向过程通过逐步减少噪声来接近真实数据分布。

B. Training Objective

训练扩散模型的目标是通过最小化均方误差（MSE）损失来学习去噪过程。该损失函数表示为：

$$\mathcal{L} = E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

其中 ϵ 是从标准正态分布中采样的噪声，ϵ_θ 是神经网络预测的噪声。通过优化此损失，模型学习在各个时间步如何有效去除噪声。

C. Model Properties

扩散模型具备以下特性：

- **稳定性**：与 GAN 不同，扩散模型在训练过程中不容易出现模式崩溃（mode collapse）的问题。
- **高质量生成**：实验表明，扩散模型能够在多个基准数据集上生成质量优于传统生成模型的图像。
- **逐步生成**：尽管生成过程可能较慢，但每一步的去噪过程确保了高精度和生成样本的多样性。

通过理解扩散模型的前向和反向过程以及其训练目标，我们可以更好地掌握该模型在生成任务中的独特优势。

III. Variants and Applications of Diffusion Model

随着扩散模型的发展，研究人员针对其生成效率、质量和适应性进行了广泛探索，提出了多种变体。这些变体在基本结构上引入了不同的创新，以解决传统扩散模型中的计算瓶颈、训练稳定性和生成速度等问题。以下是几种关键的扩散模型变体及其详细描述。

A. Denoising Diffusion Probabilistic Models (DDPM)

DDPM[6] 是扩散模型的基础变体，首次由 Ho et al. 在 2020 年提出。DDPM 的主要贡献在于通过逐步添加高斯噪声将数据点转化为标准正态分布，然后通过学习反向过程逐步去噪，恢复出高质量的样本。这种逐步去噪的生成方式使得模型在理论上更为稳定，避免了像 GAN 那样的模式崩溃问题。DDPM 在图像生成任务中展示了强大的表现，其生成质量在多个基准数据集（如 CIFAR-10、CelebA）上与 GAN 相当甚至超越 GAN。

B. Score-Based Generative Models

Score-Based Generative Models[15] 由 Song. 提出，是扩散模型的变体，旨在通过学习数据分布的梯度信息来进行高效采样。这类模型的核心思想是利用分数匹配（score matching）技术，估计数据分布的分数函数（即对数概率的梯度）。分数函数提供了样本如何在数据空间中移动的信息，使模型能够通过高效的路径生成数据。

a) 工作原理: Score-based generative models 通过两个关键步骤来实现生成任务：

1) **分数估计**：模型首先通过神经网络来学习分数函数 $\nabla_x \log p(x)$，这意味着它在给定的输入数据上预测如何移动才能增加样本的可能性。训练时，模型使用分数匹配来最小化分数估计与真实数据分布之间的误差。
2) **随机微分方程（SDE）采样**：在生成过程中，这些模型采用随机微分方程（SDE）来从噪声分布逐渐转移到数据分布。SDE 是一种连续时间的采样方法，允许模型通过引入随机扰动来实现逐步去噪并生成样本。此过程可以看作是扩散过程的推广，包含了更灵活的采样路径。

b) 采样过程: Score-based models 在采样时通过解决反向 SDE，从标准噪声分布 $p(x_T)$ 开始，逐渐演化到目标数据分布 $p(x_0)$。这种方法能够减少采样步骤的数量，从而提升生成速度。与传统扩散模型不同，score-based 模型的采样过程可以在不同分布间灵活切换，这为高分辨率和复杂结构数据的生成提供了更高的适应性。

c) 优势:

- **灵活性和高效性**：通过 SDE 方法，score-based models 能够在不同尺度上调整生成路径，使生成过程更为高效。
- **高质量生成**：由于模型在采样过程中可以动态调整生成路径，生成的样本在视觉质量和细节上表现出色。实验表明，score-based models 在高分辨率图像合成上具备显著优势。
- **理论支持**：这种方法具有坚实的理论基础，基于分数匹配和概率论，确保了训练和采样的稳定性。

d) 应用: Score-based generative models 已被成功应用于多种任务，如高分辨率图像生成、图像复原和音频生成。其灵活的 SDE 采样方法使其在处理复杂数据结构和需要高细节表现的任务中尤为有效。

C. Latent Diffusion Models (LDM)

Latent Diffusion Models (LDM)[12] 是一种计算优化的扩散模型变体，由 Rombach et al. 提出。LDM 的主要创新在于将扩散过程从数据空间移入潜在空间中进行。这一设计思想来源于高维数据（如图像）在扩散过程中计算复杂度较高、资源消耗大的问题。通过在较低维度的潜在空间进行操作，LDM 能够有效地减少计算负担，显著降低内存和计算成本，同时保持生成质量。

a) 基本思想: LDM 的核心思想是首先使用一个预训练的自动编码器（autoencoder），将输入数据映射到潜在空间，从而获得低维度的潜在表示。在潜在空间中，扩散模型对潜在表示执行逐步的去噪过程，生成新的潜在编码。最后，将处理后的潜在表示通过解码器映射回原始数据空间，从而获得高质量的生成样本。

b) 模型结构: LDM 的结构由三个主要部分组成：

- **编码器 (Encoder)**：将输入数据（如高分辨率图像）压缩到低维潜在空间。编码器通常采用卷积神经网络 (CNN) 架构，旨在保留数据的关键特征，减少冗余信息。
- **潜在扩散模型 (Latent Diffusion Model)**：在编码后的潜在空间中执行逐步的扩散过程。与传统扩散模型不同，LDM 在维度较低的空间中工作，因此计算效率更高。潜在扩散模型在每一步中去除潜在表示中的噪声，逐步将其还原为清晰的潜在表示。
- **解码器 (Decoder)**：将去噪后的潜在表示解码为原始数据的高分辨率版本。解码器通过反卷积或其他上采样方法将低维潜在表示还原为生成样本。

c) 计算效率: 由于扩散过程在低维潜在空间中进行，LDM 大大降低了模型的计算复杂度。例如，对于高分辨率图像生成任务，直接在数据空间进行扩散计算的成本非常高，而在潜在空间进行操作可以节省大量计算资源，使得训练和推理过程更加高效。此外，LDM 的结构也有助于减少显存占用，使得模型更易于在普通硬件上运行。

d) 应用场景: LDM 在多个生成任务中表现出色，特别适合处理高分辨率图像和复杂样本的生成任务，包括：

- **图像合成 (Image Synthesis)**：通过学习潜在空间中的数据分布，LDM 能够生成高清晰度的图像。
- **风格转换 (Style Transfer)**：LDM 能够在潜在空间中捕捉图像风格和内容特征，实现高质量的图像风格转换。
- **超分辨率 (Super-Resolution)**：通过在低维潜在空间中逐步去噪，LDM 可以生成比传统方法更细腻的高分辨率图像。

e) 实验结果: Rombach 等人对 LDM 进行了实验，表明在图像生成任务上，LDM 在生成质量上接近甚至超过传统的扩散模型，并在计算效率上有显著优势。实验结果表明，LDM 能够在较低的计算成本下生成高分辨率图像，并且其在不同数据集上的生成质量具有很好的稳健性。

f) 总结: LDM 的提出有效解决了高维数据扩散模型在计算成本和存储需求上的限制。通过在潜在空间中进行扩散，LDM 提供了一种高效的生成建模方法，适用于各种需要高质量输出的生成任务。它展示了潜在空间扩散模型在计算资源受限的情况下的巨大潜力，是扩散模型领域的一项重要进展。

D. Conditional Diffusion Models

Conditional Diffusion Models[8] [10] 是扩散模型的一种变体，通过引入条件输入（如文本描述、类别标签或其他模态数据），使得生成过程可以受到特定信息的指导。这种模型在生成过程中使用条件信息，使得模型不仅能够生成符合数据分布的样本，还能够生成符合特定要求或语义的样本，从而拓展了扩散模型的应用范围。

a) 基本工作原理: 在条件扩散模型中，噪声去除过程被调整为与条件信息共同进行。在每个去噪步骤中，模型接受条件输入 y 和当前状态 x_t，利用条件输入对去噪过程施加约束。假设条件输入为文本描述 y，那么模型会在去噪过程中加入文本信息，使生成样本更符合文本内容。具体来说，条件信息可以通过以下方式融入扩散模型：

- **直接拼接**：将条件信息直接拼接到扩散过程的输入中，使其在每个时间步均包含条件约束。
- **条件嵌入**：将条件信息编码为嵌入向量，然后在每个去噪步骤中结合嵌入信息，指导生成过程。
- **跨模态自注意力机制**：在多模态任务中（如文本到图像生成），条件扩散模型可以通过跨模态自注意力机制，使不同模态的数据互相关联，从而生成更加符合条件的样本。

b) 应用场景: 条件扩散模型在多模态生成任务中表现出色，能够在以下任务中实现条件生成：

- **文本到图像生成 (Text-to-Image Generation)**：模型根据输入的自然语言描述生成相应的图像。典型应用包括 OpenAI 的 DALL · E[11] 和 Google 的 Imagen[13] 等。这些模型能够将文本信息编码到潜在空间中，并通过逐步的去噪生成与文本描述一致的图像。
- **类别条件图像生成 (Class-Conditional Image Generation)**：根据类别标签生成特定类别的图像。例如，在 CIFAR-10 数据集上，模型可以生成特定类别（如 “猫” 或 “汽车”）的图像。通过类别标签的约束，模型能够更有效地生成特定类别的样本。
- **视频生成和动画生成**：条件扩散模型还可用于根据给定的场景描述生成视频或动画序列。例如，通过提供视频片段的前后帧信息，模型可以生成连续且符合条件的视频序列。

c) 典型模型:

- **DALL · E 2**[11]：由 OpenAI 提出的文本到图像生成模型，使用条件扩散模型在潜在空间中生成高分辨率图像。DALL · E 2 能够基于文本描述生成复杂且语义丰富的图像，为多模态生成任务提供了全新解决方案。
- **Imagen**[13]：Google 提出的文本到图像生成模型，结合了条件扩散和高分辨率生成技术。在多数据集上，Imagen 展现了出色的生成质量，生成的图像细节和逼真度优于传统生成模型。

d) 优势: 条件扩散模型在条件生成任务中具备多项优势：

- **生成质量**：通过逐步去噪过程，条件扩散模型能够生成高质量且符合条件的样本，避免了模式崩溃的问题。
- **多模态适用性**：条件扩散模型适用于多模态生成任务，能够处理文本、图像、视频等多种数据模态。
- **灵活性**：条件信息可随任务需求灵活调整，使得模型可以应用于从文本到图像、视频预测等多种生成任务。

e) 研究挑战: 尽管条件扩散模型在多模态生成任务中表现出色，但其在训练过程中仍面临一些挑战：

- **计算复杂度**：条件扩散模型在每一步中都需要处理条件输入，计算开销较大，尤其是在处理高分辨率图像和视频任务时。
- **条件一致性**：确保生成结果与条件信息的一致性较为困难，尤其在多模态任务中，条件和生成的跨模态一致性需要模型设计和训练策略的优化。

f) 总结: 条件扩散模型通过引入条件信息，显著扩展了扩散模型的应用场景，为多模态生成任务提供了新方法。这些模型在图像合成、视频生成和多模态生成中展现出极大潜力，推动了生成建模的前沿发展。

E. Fast Sampling Techniques

扩散模型的一大挑战在于生成过程的采样步骤通常非常多，通常需要数百到上千个时间步来逐步去噪以达到高质量的生成结果。如此多的采样步骤会导致生成过程较慢，限制了扩散模型在实际应用中的效率。为了解决这一问题，研究人员开发了多种 **Fast Sampling Techniques**[9]，以减少采样步骤数，同时保持生成质量。

a) DDIM (Denoising Diffusion Implicit Models): DDIM[14] 是一种著名的快速采样方法，首次由 Song. 提出。与标准扩散模型不同，DDIM 在反向扩散过程中使用了一种确定性采样方法，可以跳过一部分时间步，从而减少采样次数。DDIM 的核心思想是在反向去噪过程里引入隐式建模，使得生成路径可以以更少的步骤完成。具体而言，DDIM 使用一个参数化公式将反向过程简化为确定性映射，避免了对每个时间步进行随

机采样。DDIM 采样过程不仅提升了生成速度，而且在视觉质量上能够保持与标准扩散模型相近的表现。

b) 局部更新策略: 一种常用的快速采样策略是采用局部更新，即在每个时间步中仅对一部分数据进行更新，而不是对整个样本进行全局去噪。这种方法利用局部信息来加速采样过程，并且减少了计算负担。例如，在高分辨率图像生成中，局部更新可以在保持全局一致性的前提下，以更少的步骤完成去噪过程。

c) 动态时间步调整: 动态时间步调整是一种通过优化时间步长来加速采样的技术。传统的扩散模型采用均匀的时间步更新，但在实际操作中，噪声逐步去除的速率可以根据生成过程的需要进行调整。动态时间步调整通过在初期使用较大的时间步，而在接近数据分布时使用较小的时间步，从而在不损失质量的情况下大幅减少生成时间。该方法在保持生成样本细节和多样性的同时，大大提高了采样效率。

d) 模型蒸馏 (Model Distillation): 模型蒸馏 [5] [10] 是一种减少扩散模型计算负担的策略。该方法通过使用大型扩散模型的输出作为目标，训练一个较小的学生模型，使其在更少的时间步内生成接近的样本。模型蒸馏能够在不显著降低生成质量的情况下，显著提高生成速度。这种方法在需要快速推理的应用中非常有价值，例如实时生成和在线图像合成。

e) 多尺度生成 (Multi-Scale Generation): 多尺度生成是一种在不同分辨率下进行逐步生成的方法。在多尺度生成中，模型首先在低分辨率下生成样本的粗略结构，然后在更高的分辨率上逐步细化。这种策略减少了采样步骤，避免了在高分辨率空间中进行密集采样。通过多尺度生成，扩散模型可以在保证生成质量的同时大幅缩短生成时间。

f) 优势和应用场景: 快速采样技术为扩散模型在实际应用中提供了巨大的优势。这些技术减少了模型生成所需的步骤，使得扩散模型能够用于实时生成、在线内容生成和交互式生成任务。例如，DDIM 等方法在图像和视频生成任务中展现出高效性，使扩散模型能够在有限的时间内生成高质量结果。

g) 研究挑战: 尽管快速采样技术能够有效提升扩散模型的效率，但仍存在一些挑战。例如，减少时间步可能导致生成质量下降或生成样本失真。因此，如何在减少采样步数的同时保持高质量生成，仍然是快速采样技术的重要研究方向。

h) 总结:

F. Conclusion

扩散模型的变体展示了其在生成任务中的多样性和灵活性。无论是通过潜在空间扩散提高计算效率，还是通过条件生成扩展应用场景，这些变体不断推动着扩散模型在不同任务中的应用和性能提升。随着未来研究的深入，扩散模型将继续在生成建模中发挥关键作用，并提供更广泛的创新机会。

IV. Optimization in Diffusion Model

扩散模型在生成质量和稳定性方面表现出色，但其生成过程的计算复杂度较高，尤其是在处理高分辨率数据或需要快速生成时。因此，研究人员提出了多种优化技术来提高扩散模型的计算效率、生成速度和样本质量。以下是一些常见的优化技术及其详细说明。

A. Fast Sampling Techniques

快速采样技术旨在减少扩散模型生成过程中所需的时间步数。传统扩散模型通常需要数百到上千个时间步来完成去噪过程，这限制了其在实时应用中的实用性。为此，诸如 **DDIM（Denoising Diffusion Implicit Models）** 等方法通过引入确定性采样过程，将反向过程简化为更少的时间步数，从而显著加快生成速度。其他技术如多尺度生成和局部更新策略也在减少采样步骤方面展示了良好的效果。

B. Efficient Noise Scheduling

高效噪声调度是一种优化扩散过程的方法，旨在提高模型的采样效率和生成质量。传统扩散模型通常采用固定的噪声调度，但研究发现，通过自适应调整噪声水平可以显著提高生成效果。动态噪声调度允许模型在生成的早期阶段使用较大的步长，而在接近最终输出时使用更小的步长，以保留样本细节并提高生成质量。

C. Memory and Computation Optimization

内存和计算优化是另一个关键的研究领域，特别是在高分辨率生成任务中。诸如 **FlashAttention**[3] 等技术通过优化自注意力机制，利用 GPU 共享内存和分块计算来提高效率，减少内存占用。此类技术通过减少冗余计算和高效数据流策略，使得扩散模型在处理大型图像和复杂数据时表现更优。

D. Latent Space Diffusion

潜在空间扩散（如 LDM）通过在潜在空间而非数据空间中进行扩散过程，显著降低了计算复杂度。该方法首先将输入数据映射到低维潜在空间，在该空间中执行扩散和去噪操作，最后将结果解码回数据空间。这种策略不仅减少了内存使用，还加速了生成过程，使高分辨率图像生成和其他复杂任务变得更加高效。

E. Model Distillation

模型蒸馏技术通过训练一个简化的“学生模型”来模仿复杂的“教师模型”，从而减少生成步骤和计算负担。扩散模型的模型蒸馏允许在保持较高生成质量的同时显著减少时间步数，使其更适合实时应用和资源受限的环境。

F. Hybrid Models

混合模型将扩散模型与其他生成模型（如 GAN）结合，结合两者的优点来提高生成性能和效率。GAN 的快速采样能力与扩散模型的稳定性互补，混合模型可以在减少采样时间的同时生成高质量的样本。此类模型展示了在生成多样性和采样效率之间取得平衡的潜力。

G. Sparse Attention Mechanisms

稀疏注意力机制 [2] [1] [16] 通过减少全局自注意力计算的范围，降低了计算复杂度。扩散模型中的稀疏注意力技术能够智能地选择输入序列中最相关的部分进行处理，从而减少内存和时间消耗。这在高分辨率图像生成和长序列处理任务中尤为有效。

H. Adaptive Step Size

自适应步长优化了扩散模型在采样过程中的步长调整。通过在生成的初期采用较大的步长来加快进度，而在接近最终输出时缩小步长以提高细节质量，自适应步长优化提高了模型的灵活性和生成速度。

V. Conclusion

扩散模型作为生成模型领域的前沿研究，凭借其稳定的生成过程和强大的性能，逐渐成为图像、文本和音频生成任务中的重要工具。本文对扩散模型的基本结构、变体、实际应用以及优化技术进行了全面的调研和分析。

在探讨扩散模型的基本结构时，我们详细分析了其前向和反向过程，阐述了如何通过逐步添加和去除噪声，使模型能够学习复杂的数据分布并生成高质量样本。与传统生成模型相比，扩散模型在训练稳定性和生成样本多样性方面展现出明显优势，为处理复杂生成任务提供了可靠的解决方案。

随着研究的不断深入，各种变体如 DDPM、Score-Based Generative Models、Latent Diffusion Models 和 Conditional Diffusion Models 相继提出，这些变体通过创新设计提高了计算效率、生成速度和应用适应性。我们还探讨了扩散模型在图像生成、文本到图像生成以及视频生成等实际场景中的成功应用，显示了其在多模态生成任务中的广泛适用性。

在优化技术方面，我调研了多种提升扩散模型效率和性能的方法，如快速采样技术、模型蒸馏、稀疏注意力机制和高效的噪声调度。这些技术显著减少了计算资源的使用和生成时间，使扩散模型在处理高分辨率图像和实时应用中变得更为实用。

尽管扩散模型已经取得了诸多进展，但仍存在一些挑战，如生成速度的进一步提升、内存优化和多模态生成中的一致性问题。未来的研究方向可能包括更高效的采样方法、结合其他生成模型（如 GAN）的混合方法，以及在潜在空间中更复杂的扩散建模。

总体而言，扩散模型作为生成模型领域的一项重要突破，展示了其在生成任务中的巨大潜力和创新机会。通过不断探索新的优化技术和扩展应用场景，扩散模型有望在未来的生成建模和多模态应用中发挥更加关键的作用。

参考文献

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. In *arXiv preprint arXiv:2004.05150*, 2020.

[2] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[3] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[8] Alex Nichol and Prafulla Dhariwal. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[9] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Raphael Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[15] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[16] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.