第 1 章

Chapter

概&述

AI 创意绘画与视频制作
基于 Stable Diffusion
和 ControlNet

1.1 为什么使用 Stable Diffusion

ChatGPT 和 Midjourney 等 AIGC（Artificial Intelligence Generated Content，生成式人工智能）工具的兴起，给传统媒体和绘画行业造成了不小的影响。我们可以通过自然语言进行提问来获得相应的回答，其中包含了写作、编程和文艺创作。绘画艺术曾经被认为是 AI 行业最难攻克的领域，但是各种 AI 模型的迭代促进了 AI 绘画技术的升级，使它越来越接近专业画师的水平。其中以 Midjourney 为代表的 AI 程序，可以通过文字描述生成图像，它自 2022 年 7 月开放测试至今，已经迭代了 5 个版本，生成图像的质量从开始的抽象和初级的图像，到目前支持 2000+ 的不同绘画风格，包含了卡通、印象派、抽象派等风格，使得绘画的门槛降低到非专业人士也可以上手。

与商业版Midjourney对应的开源工具Stable Diffusion，提供了高度定制化和免费的方案，使得它在开源社区里得到广泛推崇。Midjourney 和 Stable Diffusion 的对比如表 1-1 所示。

表1-1 Midjourney和Stable Diffusion的对比

特　征	Midjourney	Stable Diffusion
成本	收费	免费
内容过滤器	Yes	No
图像定制	相对较低	较高
上手难度	中等	较低
生成优质图像的难度	较高	相对较低
宽高比设定	支持	支持

（续表）

特　征	Midjourney	Stable Diffusion
修复图像Inpainting	不支持	支持
模型变体	较少	很多
许可证	需要看收费账号的层级	需要看模型的许可

近年来，人工智能在艺术领域的应用已经引起越来越多的关注，其中利用人工智能进行绘画创作更是引起了设计师、艺术家和计算机科学家的极大兴趣。Stable Diffusion 和 ControlNet 是人工智能绘画创作中的两个关键技术方法，使用它们可以绘制出令人惊叹的艺术效果。如图 1-1 所示是使用 Stable Diffusion 和 ControlNet 创作的画作。

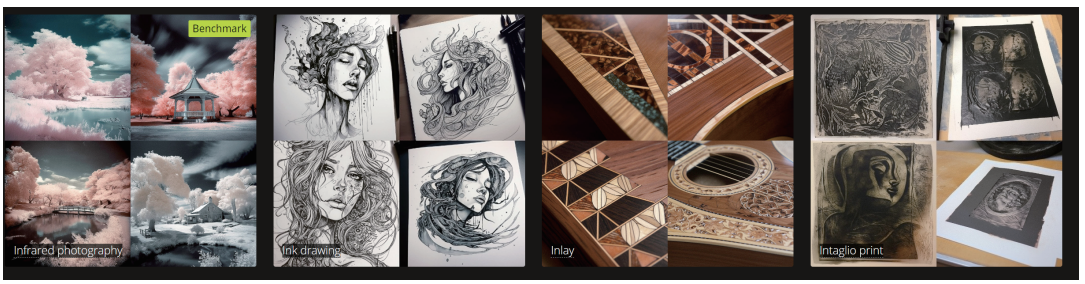

图 1-1　使用 Stable Diffusion 和 ControlNet 创作的画作

Stable Diffusion 可以在生成过程中有效地管理绘画元素的分布和形式，从而产生视觉平衡和美观的艺术作品。通过训练一个具有控制功能的神经网络，ControlNet 可以实现对生成的绘画内容的形状、颜色、纹理等方面的精细控制，从而满足艺术家或设计师对绘画创作的个性化需求。例如针对人物作图里常见的动作迁移，可以结合 ControlNet 中的 OpenPose 模型实现人体关键点识别，进行模特质态的精细调整；结合 Depth 和 Canny 模型可以对建筑物或机械类图形进行纹理调整，等等。

1.2　AI 图像生成模型介绍

随着人工智能的不断发展，图像生成技术逐渐成为热门领域。AI 图像生成模型作为一种基于深度学习的技术，已经在多个应用领域取得了显著的成果，如艺术、设计、广告和游戏等。在这一领域中，不同的 AI 图像生成模型呈现出各自独特的特点和优势，通过不同的训练方法和技术能够生成出多样化、高质量的图像内容。

第 1 章 概述

GAN（Generative Adversarial Network，生成对抗网络）是一种常用的 AI 图像生成模型，由生成器（Generator）和判别器（Discriminator）组成。生成器负责生成虚假的图像，而判别器则负责判断图像的真实性。生成器和判别器通过对抗训练的方式相互竞争，不断优化，从而使生成的图像更加真实和逼真。GAN 模型在图像生成领域取得了重要的突破，如生成高分辨率图像、风格迁移和图像编辑等。

另一种常见的 AI 图像生成模型是变分自编码器（Variational Autoencoder，VAE），它是一种生成模型，结合了自编码器（Autoencoder，AE）和概率模型的思想。VAE 的生成器将输入图像映射到潜在空间（Latent Space），再从潜在空间中采样，最终通过解码器生成图像。VAE 模型在图像生成领域具有较强的潜在表示学习能力，能够生成具有多样性和连续性的图像。

此外，还有许多其他类型的 AI 图像生成模型，如自注意力模型（Self-Attention Model）、光流模型（Flow Model）、PixelRNN 和 PixelCNN 等。这些模型都具有不同的特点和优势，并在不同的应用场景中得到了广泛应用。例如，自注意力模型在生成长文本和高分辨率图像时表现出色，流模型在处理连续生成任务时具有优势，PixelCNN 在生成像素级图像时能够保持细节和清晰度。

这些 AI 图像生成模型的训练方法也各有不同。一般来说，训练这些模型需要大量的数据和计算资源。GAN 模型通常通过交替训练生成器和判别器来进行优化，使用梯度下降等优化算法进行参数更新。VAE 模型则通过最大化对数似然函数进行训练，同时引入潜在空间的正则化项以控制生成图像的多样性。其他类型的模型也有各自的训练方法，如自注意力模型通过自注意力机制对输入序列进行编码和解码，流模型通过对概率密度函数进行建模来生成图像，PixelRNN 和 PixelCNN 则通过对图像像素的条件生成来进行训练。

除了训练方法和技术的不同，这些 AI 图像生成模型在生成图像的质量、多样性、速度和稳定性等方面也存在差异。例如，GAN 模型通常能够生成质量较高的图像，但在生成过程中可能会出现不稳定和模式崩溃的问题。VAE 模型一般能够生成具有较好多样性和连续性的图像，但在生成质量上可能稍显逊色。流模型在生成连续数据时较为出色，但在生成高分辨率图像时可能速度较慢。

总的来说，AI 图像生成模型作为一种先进的技术，已经在图像生成领域取得了重要的突破，并在多个应用场景中起到关键性的作用。

1.3 从小白到 AI 艺术家

本书将深入探讨 Stable Diffusion 和 ControlNet 在 AI 绘画创作中的应用，详细阐述这两种技术在生成绘画作品时的原理、方法和应用实践。同时，还将探讨这两种技术的优缺点、挑战和未来发展方向。

如今我们正处在整个时代技术的前沿，新的 AI 技术使得 AI 绘画的受众更广。不需要写代码，也不需要特别高深的知识，只要敢于尝试，通过学习本书，任何人都可以从一个小白成长为 AI 艺术家。

本书所有图像均使用 Stable Diffusion 现有模型通过文生图或者图生图模块生成，旨在通过图文介绍的方式让读者了解 Stable Diffusion 的功能和原理，掌握其使用方法。

1.4 总结

近年来，基于人工智能生成图像的技术取得了飞速的发展，并得到了广泛的应用。与以往我们认知中质量欠佳、搭建复杂、使用烦琐的 AI 图像生成工具相比，现今的情况已经发生了巨大变化。随着技术的不断创新和进步，AI 图像生成工具和技术有了质的飞跃，从以前的模糊、不自然，到如今的逼真、细致，生成图像的质量显著提升。与此同时，过去需要大量技术背景和复杂设置的问题，如今得到了极大简化，让更多的人能够轻松上手。AI 绘画领域的进步也得益于深度学习技术的推动，它使得生成的图像更加逼真，同时大量的数据作为训练基础也使得生成的图像更加符合人们的审美和预期。

1.5 练习

（1）目前市面上使用 AIGC 生成图片的主流工具有哪几种？

（2）通过 AI 生成图片的技术框架有哪几种？

第 2 章 Chapter 人工智能与图像生成技术

随着机器学习和深度学习技术的发展，传统的强调知识积累和创造性的行业也被 AI 技术突破。本章将介绍人工智能的发展及其在图像生成技术方向上用到的技术，包括人工智能的发展历程、图像生成技术的基本原理、PyTorch 框架和 AIGC 技术框架。这些技术之间的交叉和融合也为机器学习和计算机视觉领域的进一步发展提供了新的机会和挑战，同时也奠定了 AIGC 技术大量应用的基础。

2.1 人工智能的发展历程

人工智能是一种通过计算机模拟人类智能，使机器能够像人类一样思考、学习、推理和决策的技术。人工智能起源于 20 世纪 50 年代，从那时起，它一直是计算机科学领域的重要研究领域。本节，我们将介绍人工智能的基本概念和发展历程，并重点关注图像生成技术的历史和发展。

人工智能的基本概念涵盖了一系列模拟人类智能的技术，其目的是使计算机系统能够处理各种智能任务。这些任务包括语音识别、自然语言处理、图像识别、机器学习和决策制定等。人工智能技术可以分为强人工智能和弱人工智能。强人工智能是一种完全智能的系统，能够像人类一样思考和解决问题。弱人工智能是一种专门针对特定任务进行优化的系统，例如语音识别系统或机器人控制系统。

人工智能的发展历程可以追溯到 20 世纪 50 年代，当时计算机科学家开始尝试模拟人类智能。早期的人工智能系统主要依赖于规则，这些规则指导计算机系统进行逻辑推理和问题解决。然而，这种方法很快显示出了局限性，因为它无法解决复杂的现实世界的问题。

随着时间的推移，人工智能研究逐渐转向了基于知识的系统，这些系统使用预先编写的知识库来解决问题。这种方法提高了人工智能系统的表现，但仍然受到知识表示和知识获取的限制。

到了 20 世纪 80 年代，机器学习成为人工智能领域的重要研究方向。机器学习是一种自我学习的方法，它使用大量数据来训练计算机系统，并通过经验不断改进性能。机器学习技术是现代人工智能的核心，是许多 AI 应用的基础。

近年来，生成式人工智能（Generative AI）已经成为人工智能领域的一项重大突破。它能够创造出各种类型的数据，包括图像、视频、音频、文本和 3D 模型。生成式模型的基本原理是通过大量的输入数据进行学习，并生成与原始数据相似度极高的内容。生成式人工智能能够创造出高度逼真和复杂的内容，能够模仿人类的创造力，因此，它在游戏、娱乐和产品设计等许多行业中被广泛应用，并成为宝贵的工具。

在生成式人工智能中，不得不提的是 Transformer，它是一种非常强大的神经网络模型，在自然语言处理、图像音频处理等各种生成式任务中具有广泛的应用。该模型最主要的特点是采用了一种被称为注意力机制的技术，这种技术能够帮助模型更好地理解输入数据的上下文关系和重要性。ChatGPT 就是通过 Transformer 模型进行训练的，ChatGPT 能够有效地生成具有语义连贯性和上下文一致性的文本响应。Transformer 模型的出现为自然语言生成任务带来了重大的突破和进步。

2.2 图像生成技术的基本原理

图像生成技术是一种人工智能技术，用于生成逼真的图像和视频。该技术通常使用深度神经网络来学习图像特征，并生成外观逼真的新图像。随着人工智能的发展，图像生成技术也取得了显著的进展，并在多个领域得到了广泛应用。

2.2.1 深度神经网络图像生成技术

在过去的几年中，人工智能技术在图像生成方面取得了令人瞩目的进展。其中，DALL-E 2、Midjourney 和 Stable Diffusion 都是最新的图像生成技术，均采用深度神经网络模型来生成逼真的图像。然而，它们在设计和实现上存在一些显著的区别。

1 DALL-E 2

DALL-E 2 是 OpenAI 最新的图像生成模型，它可以通过自然语言描述生成逼真的图像。作为 DALL-E 的升级版本，DALL-E 2 具有更高的保真度，能够在图像中保留语义相关性和高保真度的细节。因此，DALL-E 2 是目前最先进的图像生成技术之一，具有极高的创造力和准确性。

2 Midjourney

Midjourney 是一种基于变分自编码器的生成模型，它可以从图像中学习潜在的变量并生成新的图像。与其他生成模型不同，Midjourney 采用一种新颖的训练方法，即从训练数据中随机选择两幅图像并将它们组合成一幅新图像，然后训练模型生成这幅新图像。这种方法使得 Midjourney 在生成新图像时具有更高的创造力和多样性。

3 Stable Diffusion

Stable Diffusion 是一种最近推出的图像生成技术，它通过控制生成过程的稳定性来生成逼真的图像。与其他生成模型不同，Stable Diffusion 将生成过程分成多个步骤，每个步骤都是一种稳定的演化，使得生成过程更加可控和稳定。这种方法使得 Stable Diffusion 生成的图像具有更高的质量和稳定性。

DALL-E 2、Midjourney 和 Stable Diffusion 虽然在训练方法、生成过程的稳定性和创造力等方面存在区别，但它们都代表着人工智能技术的最新发展趋势。

2.2.2 Stable Diffusion 的关键组件

Stable Diffusion 是一种从噪声或不完整输入中生成高质量图像样本的方法，它依赖于 3 个关键组件：ClipText、UNet + 调度器和自编码器解码器。

1 ClipText

ClipText 是一种文本编码器，它将输入文本映射到一个固定长度的特征向量。ClipText 模型是在一幅大型的图像和它们关联字幕的数据集上进行训练的，学习将每幅图像与相应的文本描述关联起来。

ClipText 绘画生成技术如图 2-1 所示。

图 2-1 ClipText 绘画生成技术

CLIP（Contrastive Language-Image Pretraining）是一种通过将图像和文本描述相互连接并使用评分方法来衡量它们之间的相似性的方法。CLIP 利用了互联网上亿级的图像和对应的描述数据，并使用 CLIP 对生成的图像（例如使用 GAN 方式）进行评分，以提高 CLIP 的准确性。目前使用 CLIP 技术的有 DepDaze、ApephImage、Disco Diffusion（diffusion model）等。

2 UNet + 调度器

UNet 是一种常用于图像分割任务的神经网络架构。在稳定扩散的背景下，UNet 用于填补输入图像中的缺失或损坏部分。该方法的调度器组件会在训练过程中调整扩散程度。

UNet 也被称为 U-Net 架构，是一个卷积神经网络（CNN），由 Olaf Ronneberger、Philipp Fischer 和 Thomas Brox 在 2015 年发表的《U-Net：应用于生物医学图像分割的卷积网络》论文中阐述。该网络已被证明对医学图像分割任务极为有效，特别是在生物医学图像分析领域。

U-Net 由 3 个关键部分组成：收缩层、瓶颈层和扩展层。

- 收缩层（也称为编码器）的作用是逐渐减小输入图像的大小，并增加通道的数量。通过一系列的卷积层和下采样操作，它能够提取出图像的局部特征，并转化为更高级别的抽象特征。
- 瓶颈层是 U-Net 结构的核心部分，它的目标是捕捉输入图像的高级特征。由多个卷积层组成，这一层有助于减少特征图的维度，并保留重要的空间信息。通过瓶颈层，U-Net 能够整合全局和局部信息，以获取图像中的细节和上下文关系。
- 扩展层（也称为解码器）的任务是将特征图进行上采样，并恢复到原始图像的尺寸。通过一系列的上采样操作和卷积层，U-Net 能够逐步重建图像的空间分辨率。U-Net 中的跳跃连接是一项关键技术，它允许将来自收缩层和扩展层的特征图进行连接。这种连接方式有助于保留细粒度的空间信息，提高分割结果的准确性和稳定性。

U-Net 架构中的跳过连接提供了一种将编码器路径的特征图与解码器路径的特征图相结合的方法。这使得网络能够学习并纳入来自多个尺度的信息，从而提高了准确性和得到了更好的分割结果。

总而言之，U-Net 是一种强大而高效的卷积神经网络结构，被广泛应用于图像分割任务等。

3 变分自编码器

在机器学习中，变分自编码器是由 Diederik P. Kingma 和 Max Welling 提出的一种人工

神经网络结构，属于概率图模式和变分贝叶斯方法。变分自编码器用于从 ClipText 和 UNet 产生的编码特征向量中生成高质量图像样本，它将这些特征向量作为输入，并产生相应的图像作为输出。

2.3 深度学习框架 PyTorch 基础

PyTorch 是一种用于构建深度神经网络模型的开源机器学习框架。它由 Facebook 的人工智能研究团队开发，并于 2017 年首次发布。PyTorch 提供了一组灵活且高效的工具，可以让开发者轻松创建、训练和部署深度学习模型。

在 PyTorch 中，核心数据结构是张量（Tensors），它类似于多维数组。张量可以在 CPU 或 GPU 上运行，并且支持各种数学操作。与 NumPy（一种基于 PyTorch 语言的科学计算工具）数组操作类似，PyTorch 中的张量操作也非常便捷，而且还可以利用 GPU 加速计算。

PyTorch 使用动态计算图来跟踪计算过程，这是框架的一大特点。相比于静态计算图，动态计算图允许开发者使用常规的 Python 控制流程语句（如循环和条件语句），而无须预先定义静态计算图。这种设计使得模型的定义和调试更加灵活和直观。

PyTorch 的另一个重要功能是自动求导（Automatic Differentiation），它能够自动计算张量操作的梯度。通过调用 .backward() 方法，可以方便地计算相对于模型参数的梯度，这对于训练神经网络模型非常有用。自动求导还支持高阶导数和向量化操作。

PyTorch 提供了丰富的工具和模块来构建深度神经网络模型。可以通过继承 torch.nn.Module 类来定义自己的模型，并且可以使用各种预定义的层（如全连接层、卷积层、循环神经网络等）来组成模型。此外，PyTorch 还提供了方便的初始化方法、损失函数和优化器等。

在数据加载和处理方面，PyTorch 提供了 torch.utils.data 模块，用于加载和处理训练和测试数据。PyTorch 还可以自定义数据集类，并使用数据加载器进行批量数据加载和随机化。此外，PyTorch 还提供了各种数据变换和增强的功能，如随机裁剪、翻转和归一化等。

PyTorch 还具有 GPU 加速的能力，可以利用 GPU 来加速深度学习模型的训练和推断。通过使用 .to(device) 方法，可以将模型和数据移动到 GPU 上，并利用 GPU 进行并行计算。这种 GPU 加速对于处理大规模数据和复杂模型非常重要。

总之，PyTorch 是一种功能强大且灵活的深度学习框架，它提供的丰富的工具为 Stable Diffusion 的实现展现了可能和便利。

2.4 AIGC 技术框架介绍

生成式人工智能 AIGC（Artificial Inteligence Generated Content）是人工智能发展到今天的重要成果。

常见的 AIGC 技术框架有：

- GAN，由生成器和判别器组成的图像生成模型，已在当前图像生成等领域获得长足进展。
- VAE，可变分自编码器，使得数据可以从图像空间转换到潜在空间中，使得扩散模型性能得到提升。
- NeRF，一种用来渲染三维场景的技术，通过神经网络对场景深度和颜色进行建模，生成高质量的三维模型。
- CLIP，用于建立图像和文本之间的关联，也是奠定文本生成图像方案的基础。它的出现极大推动了文本和图像之间的跨模态交互和应用。
- CodeFormer，这是集成在 Stable Diffusion 中的一种常见的人脸清晰化模型，通过一键勾选的方式在生成人物图像时为需要进行面部重建的部分提供极大的便利。

下面对上述几个技术框架进行详细介绍。

2.4.1 GAN 对抗网络

GAN 是一种深度学习模型，由生成器和判别器两个相互对抗的网络组成。GAN 的初衷是让生成器能够产生与真实数据相似的新样本。

生成器的任务是接收一个随机噪声向量作为输入，并生成看起来像真实数据的样本。而判别器则被训练来区分生成器生成的样本和真实数据样本。通过生成器和判别器的对抗学习，二者不断优化自身，使得生成器生成的样本更加逼真，而判别器能更加准确地区分真实样本和虚假样本。

GAN 的训练过程可以简单概括如下：

步骤 01 初始化生成器和判别器的参数。

步骤 02 从真实数据中随机选择一批样本作为判别器的真实数据输入。

步骤 03 从噪声分布中随机采样一批噪声向量作为生成器的输入，生成一批虚假样本。

步骤 04 将判别器分别对真实数据和虚假样本进行分类，并计算它们的损失（通常使用二分类的交叉熵损失）。

步骤 05 对判别器进行反向传播和参数更新，以提高对真实和虚假样本的分类准确性。

步骤 06 固定判别器的参数，更新生成器的参数，使得生成器生成的样本更容易被判别器误认为真实数据。

步骤 07 重复步骤 02 至步骤 06，进行多次迭代训练，直到生成器生成的样本质量满足预期。

通过生成器和判别器的对抗学习，GAN 能够学习数据的分布特征并生成逼真的新样本。GAN 广泛应用于图像生成、图像编辑、生成对抗攻击、数据增强等领域。其重要的分支 RealESRGAN 和 ESRGAN（Enhanced Super-Resolution Generative Adversarial Network）用来进行高分辨率处理，GFPGAN 用来进行面部修复。

1 ESRGAN 超分辨率

ESRGAN 即增强型超分辨率生成对抗网络，是一种令人惊叹的深度学习模型，专为图像超分辨率而设计。超分辨率意味着通过引入像素级的细节提升，让图像展现出更为清晰和细腻的魅力。

ESRGAN 以生成对抗网络为基础，独特而出色地生成出高质量的超分辨率图像。它的目标是令那些低分辨率的图像以一种更高的品质展现于世。这一目标通过生成器网络和判别器网络的相互协作、相互竞争来实现。

在 ESRGAN 的训练过程中，生成器和判别器相互对抗、相互学习。生成器扮演着一个巧妙的“骗子”，力图让判别器分不清生成的图像与真实高分辨率图像的差异。而判别器则是一位精明的辨别者，努力学习如何分辨真实和生成的图像，并向生成器提供改进建议。这种博弈、对抗的学习过程不断推动着生成器提升生成图像的质量。

2 GFPGAN 人脸修复算法

人脸修复是指从低分辨率的人脸图像中恢复出高清晰度的人脸图像。目前，GFPGAN 是一种开源的人脸修复算法，已经被集成到 Stable Diffusion Web UI 中，用于重新绘制面部。该算法通过在训练过程中对低质量人脸图像进行预处理来保留面部的基本信息。同时，通过引入具有辨别性的面部损失（Facial Component Loss）来判断哪些细节需要保留，之后通过保留损失（Identity Preserving Loss）来保持面部特征。

2.4.2 VAE 变分自编码器

VAE 是一种与自编码器密切相关的模型。尽管 VAE 与自编码器在结构上有一定的相似性，但在目标和数学表述上存在显著差异。VAE 属于概率生成模型（Probabilistic Generative Model），神经网络是其中的一个组件。根据其功能的不同，VAE 可分为编码器和解码器。

编码器的主要功能是将输入变量映射到潜在空间，与变分分布的参数相对应。这样做的结果是可以生成多个遵循同一分布的不同样本。相反，解码器的功能是从潜在空间映射回输入空间，以生成数据点的表示。

在 VAE 中，我们追求的目标是最大化观察数据的边际对数似然。为了达到这个目标，VAE 使用变分推断的方法来近似潜在空间的后验分布。它通过最大化似然下界（ELBO）来进行优化。这个下界是通过对潜在表示进行采样后的期望得到的，并且通常使用重参数化技巧（Reparameterization Trick）进行训练。

总之，VAE 是一种概率生成模型，与自编码器密切相关。它通过将输入数据映射到潜在空间并利用变分推断的方法，实现了对数据分布的建模。通过从学习到的分布中进行采样，VAE 能够生成新的样本。值得注意的是，重参数化技巧是训练 VAE 时常使用的技术之一，它可以有效地优化模型。

2.4.3 NeRF 辐射神经网络

NeRF（Neural Radiance Fields）是一种基于神经网络的方法，用于对三维场景进行建模和渲染。它使用神经网络来表示场景中每个点的辐射强度和体素颜色。换句话说，NeRF 试图从输入图像中学习场景的几何结构和光照属性。通过训练网络，NeRF 能够估计场景中任意点的辐射强度和颜色，从而实现高质量的渲染。为了训练 NeRF，需要将输入图像与场景中的真实数据进行匹配，以优化网络参数。通常使用光线追踪等技术来生成训练数据。

NeRF 的优势在于能够生成高度逼真的三维渲染结果，包括光照、阴影和反射等效果。它已经被广泛应用于计算机图形学、虚拟现实和增强现实等领域。目前可以通过 Stable Diffusion 来生成同一物体的不同角度的 2D 照片，并通过 NeRF 进行 3D 建模渲染。

2.4.4 CLIP 对比性语言－图像预训练模型

CLIP 的全称为 Contrastive Language-Image Pre-training，是一种文字－图像对的预训

练方法。作为一种对比学习的多模态模型，它的训练目标是根据图像和对应的文字描述，通过大量的训练以及提取的文字和图像特征找到文字－图像对的关联关系。CLIP 能够将图像和文本映射到共享的潜在空间，并具备理解和推理图像与文本之间联系的能力。这使得 CLIP 成为处理图像与文本语义关系的重要工具，并在计算机视觉和自然语言处理等领域取得了重要的进展。CLIP 的重要应用有图像分类、通过图像生成对应的描述语、通过文字描述生成图像（Stable Diffusion 使用 CLIP 模型从文本中生成对应的高保真图像）。

2.4.5 CodeFormer 人脸清晰化模型

CodeFormer 是一种强大的面部恢复算法，旨在处理旧照片和 AI 生成的图像面部。

CodeFormer 人脸清晰化模型的过程如下：

步骤 01 会通过一系列学习过程来训练一个离散的 codebook 和一个解码器。这样做的目的是通过自重构学习，将面部图像的高质量视觉部分存储起来。通过这个过程，我们能够掌握如何有效地表示和保存面部图像中的关键视觉特征。

步骤 02 将使用事先确定好的 codebook 和解码器。我们引入一个称为 Transformer 模块的组件用来对低质量输入的全局人脸组成进行建模。这个模块的作用是通过编码序列预测来处理输入数据。通过这种方式，我们可以更好地理解和建模低质量人脸图像的整体结构和组成部分。

步骤 03 引入可控特征转换模块。这个模块的作用是控制从低质量编码器（LQ Encoder）到解码器的信息流。通过调整这个信息流，我们可以控制图像重建和转换过程中的特征变化。这样的设计使得我们能够根据需要调整图像的某些特征，例如亮度、对比度和姿态等，以获得更加满意的结果。

综上所述，CodeFormer 人脸清晰化模型的过程包括学习 codebook 和解码器以存储高质量视觉部分，使用 Transformer 模块对低质量输入的人脸组成进行建模，以及应用可控特征转换模块来控制信息流动。这一模型的设计旨在改善低质量人脸图像，并提供一种灵活的方式来控制图像特征。

CodeFormer 用来人脸清晰化的效果如图 2-2 所示。从图中可以看出，低质量的人脸图像得到重建和改进，细节和纹理变得更加清晰可见，面部轮廓和特征也更加清晰和鲜明。

此外，CodeFormer 模型还可以帮助纠正模糊或失真的图像部分，使人脸图像整体上更加自然和真实。

图 2-2 CodeFormer 人脸清晰化的效果

2.5 总结

人工智能从发展初期到如今，已经在各个领域取得一定的突破。本章通过介绍常见的AIGC 技术框架，探讨其在图像生成方向上的应用和前景。正是前沿的技术之间的相互作用，为 AIGC 技术的广泛应用奠定了坚实的基础。此外，本章通过聚焦于图像生成技术的基本原理，还探讨了其背后的工作机制和实现原理。

2.6 练习

（1）列举 Stable Diffusion 框架的主体结构以及相关功能。

（2）描述 PyTorch 框架的特点和优势。