Control Net

代码地址：github.com/lllyasviel/...

论文地址：arxiv.org/abs/2302.05...

分析：

https://juejin.cn/post/7210369671656505399

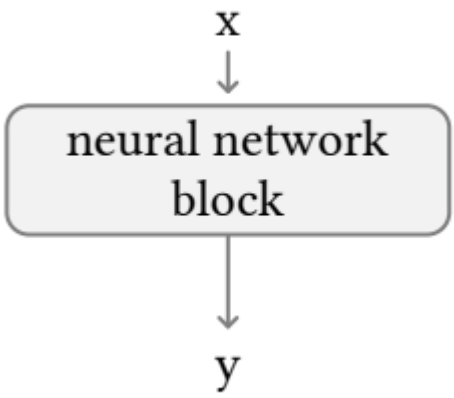
https://blog.csdn.net/qq_45752541/article/details/132619474

原理

方法

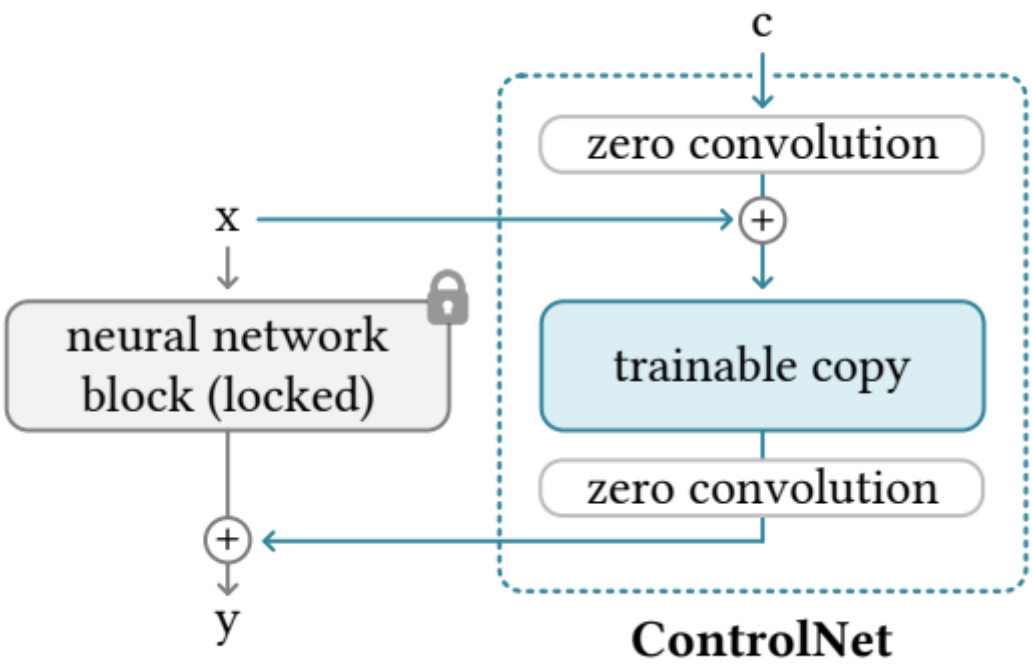
- ControlNet通过对神经网络模块的**输入条件**进行操作，从而进一步控制整个神经网络的整体行为。其中，“神经网络模块”是指将一组神经层作为一个常用单元组合在一起构建神经网络，如“resnet”块、“convn-bn-relu”块、多头注意力块、transformer模块等。
- 以二维特征为例，给定特征x∈h×w×c，{h, w, c}为高度、宽度和通道，神经网络模块F(·;Θ)和一组参数Θ将x转换为另一个特征y：

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \Theta)$$



- 如果将所有参数锁定在Θ中，然后将其**克隆**为**可训练的副本Θc**。复制的Θc使用外部条件向量c进行训练。在本文中，称原始参数和新参数为“锁定副本”和“可训练副本”。制作这样的副本而不是直接训练原始权重的动机是：避免数据集较小时的过拟合，并**保持**从数十亿张图像中学习到的**大型模型的能力**。
- 神经网络模块由一种称为“**零卷积**”的独特类型的卷积层连接，即1×1卷积层，权重和偏差都用零初始化。将零卷积运算表示为**Z(·;·)**，使用参数{Θz1，Θz2}的两个实例组成ControlNet结构：

$$\boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x}; \Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c); \Theta_{z2})$$



- 其中yc成为该神经网络模块的输出。因为零卷积层的权值和偏差都初始化为零，所以在第一个训练步骤中，有：

$$\begin{cases} \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}) = \boldsymbol{0} \\ \mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c) = \mathcal{F}(\boldsymbol{x}; \Theta_c) = \mathcal{F}(\boldsymbol{x}; \Theta) \\ \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c); \Theta_{z2}) = \mathcal{Z}(\mathcal{F}(\boldsymbol{x}; \Theta_c); \Theta_{z2}) = \boldsymbol{0} \end{cases}$$

- 这可以转换为yc = y
- 表明，在第一个训练步骤中，神经网络块的可训练副本和锁定副本的所有输入和输出都与它们的状态一致，就像ControlNet不存在一样。换句话说，当一个ControlNet应用于一些神经网络块时，在进行任何优化之前，它不会对深层神经特征造成任何影响。任何神经网络块的能力、功能和结果质量都得到了完美的保留，任何进一步的优化都将变得像微调一样快(与从零开始训练这些层相比)。
- 下面简单地推导零卷积层的梯度计算。考虑权值W和偏差B的1 × 1卷积层，在任意空间位置p和通道索引i处，给定输入特征i∈h×w×c，正向通过可写成

$$\mathcal{Z}(\boldsymbol{I}; \{\boldsymbol{W}, \boldsymbol{B}\})_{p,i} = \boldsymbol{B}_i + \sum_j^c \boldsymbol{I}_{p,i} \boldsymbol{W}_{i,j}$$

- 零卷积有W = 0和B = 0(在优化之前)，对于Ip,i非零的任何地方，梯度变为：

$$\begin{cases} \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{B}_i} = 1 \\ \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{I}_{p,i}} = \sum_j^c \boldsymbol{W}_{i,j} = 0 \\ \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{W}_{i,j}} = \boldsymbol{I}_{p,i} \neq \boldsymbol{0} \end{cases}$$

- 可以看到，尽管零卷积可以导致特征项i的梯度变为零，但权值和偏差的梯度不受影响。在第一次梯度下降迭代中，只要特征I是非零，权重W就会被优化为非零矩阵。值得注意的是，在例子中，特征项是从数据集中采样的输入数据或条件向量，这自然地确保了I不为零。例如，考虑一个具有总体损失函数L和学习率βlr的经典梯度下降，如果"外部"梯度∂L/∂Z(I;{W, B})不为零，有：

$$\boldsymbol{W}^* = \boldsymbol{W} - \beta_{\text{lr}} \cdot \frac{\partial \mathcal{L}}{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})} \odot \frac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})}{\partial \boldsymbol{W}} \neq \boldsymbol{0}$$

- 其中W∗是一阶梯度下降后的权值;是Hadamard乘积。在这一步之后，可得到：

$$\frac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W}^*,\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{I}_{p,i}} = \sum_j^c \boldsymbol{W}^*_{i,j} \neq \boldsymbol{0}$$

- 获得非零梯度，神经网络开始学习。通过这种方式，零卷积成为一种独特的连接层类型，以学习的方式逐步从零增长到优化参数。

Related Work

链接：https://www.zhihu.com/question/614056414/answer/3273259612

2.2. 图像扩散Image Diffusion

图像扩散模型最早由Sohl- Dickstein等人[80]引入，最近被应用于图像生成[17, 42]。潜在扩散模型（LDM）[71]在潜在图像空间[19]中执行扩散步骤，这降低了计算成本。文本到图像扩散模型通过预先训练的语言模型（如CLIP [65]）将文本输入编码为潜在向量，从而实现了最先进的图像生成结果。Glide [57]是一个支持图像生成和编辑的文本引导扩散模型。Disco Diffusion [5]在clip引导下处理文本提示。Stable Diffusion [81]是潜在扩散[71]的大规模实现。Imagen [77]使用金字塔结构直接扩散像素，而不使用潜在图像。商业产品包括DALL-E2[61]和Midjourney[54]。

Image Diffusion Models were first introduced by Sohl-Dickstein et al. [80] and have been recently applied to image generation [17, 42]. The **Latent Diffusion Models (LDM)** [71] perform the diffusion steps in the latent image space [19], which reduces the computation cost. **Text-to-image diffusion models** achieve state-of-the-art image generation results by encoding text inputs into latent vectors via pretrained language models like CLIP [65]. **Glide** [57] is a text-guided diffusion model supporting image generation and editing. **Disco Diffusion** [5] processes text prompts with CLIP guidance. **Stable Diffusion** [81] is a large-scale implementation of latent diffusion [71]. **Imagen** [77] directly diffuses pixels using a pyramid structure without using latent images. Commercial products include **DALL-E2** [61] and **Midjourney** [54].

my:

Image Diffusion Models

Diffusion Models

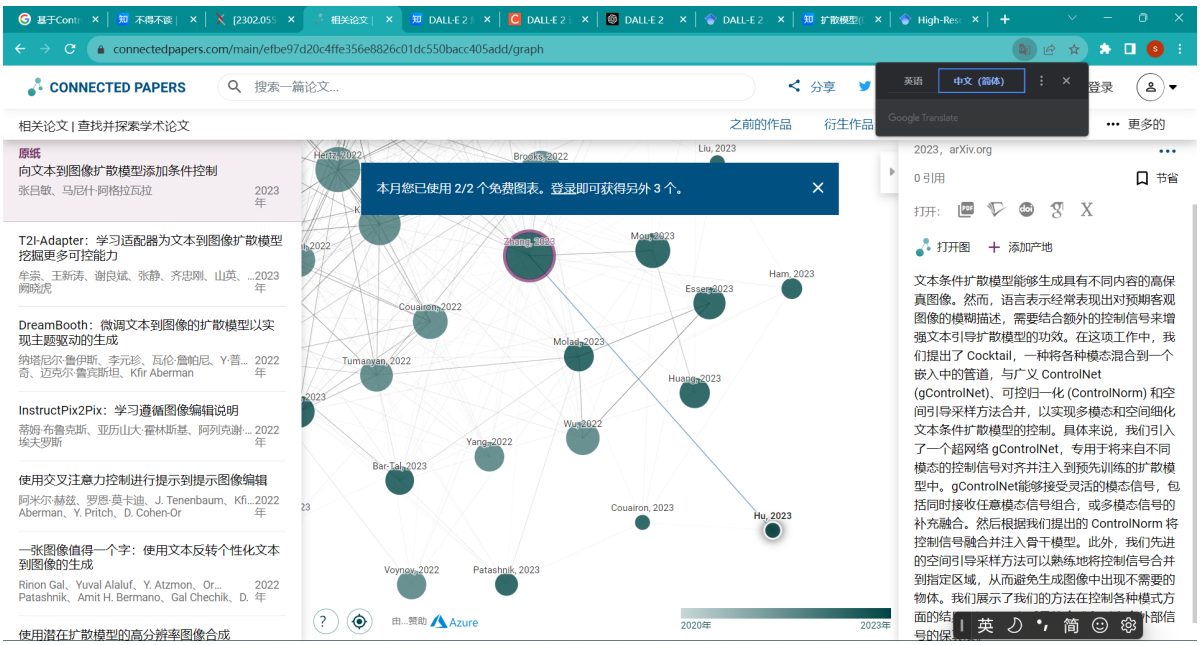
优化 或者基于此的工作

https://cloud.tencent.com/developer/article/2328560

gControlNet

SD controlNet work?

Extending Text2Video-Zero for Multi-ControlNet



其他

t2i

近年来，扩散模型[12]在图像合成领域取得了巨大成功。它旨在通过迭代去噪过程从高斯噪声中生成图像。其实现建立在严格的物理原理基础上[38, 39]，包括扩散过程和逆过程。在扩散过程中，图像X0通过T次迭代添加随机高斯噪声转换为高斯分布XT。逆过程则是通过多个去噪步骤从XT中恢复X0。

近年来，许多扩散方法集中在文本到图像（T2I）生成任务上。例如，Glide [23]提出将文本特征融入去噪过程中的变换器块。随后，DALL-E [30]、Cogview [6]、Make-a-scene [10]、Stable Diffusion [32]和Imagen [34]显著提高了T2I生成性能。广泛采用的策略是在特征空间进行去噪，并通过交叉关注模型将文本条件引入去噪过程。尽管它们取得了令人满意的合成质量，但仅靠文本提示无法提供可靠的结构指导。

PITI [43]提出通过缩小其他类型条件的特征与文本条件之间的距离来提供结构指导。[42]提出利用目标草图和中间结果之间的相似度梯度来约束最终结果的结构。还有一些方法[11, 9, 1]旨在调制预训练的T2I模型中的交叉关注图，以指导生成过程。这种方法的一个优点是它们无需单独训练，但在复杂场景中仍然不太实用。作为并发工作，[45]学习了专门的控制网络以实现预训练T2I模型的条件生成。[14]提出基于一组控制因素重新训练扩散模型。

zhihu

最近，扩散模型[12]在图像生成领域取得了巨大成功。图像生成领域最常见生成模型有GAN和VAE，2020年，DDPM（Denoising Diffusion Probabilistic Model）被提出，被称为扩散模型（Diffusion Model），同样可用于图像生成。和其他生成模型一样，实现从噪声（采样自简单的分布）生成目标数据样本。扩散模型包括两个过程：前向过程（forward process）和反向过程（reverse process），其中前向过程又称为扩散过程（diffusion process），图像X0通过T次迭代添加随机高斯噪声转换为高斯分布XT。其中反向过程可用于生成数据样本，是通过多个去噪步骤从XT中恢复X0。

OK1:

Recently, diffusion models [1] have achieved tremendous success in the field of image generation. The most common models for image generation are GAN and VAE. In 2020, DDPM (Denoising Diffusion Probabilistic Model) was introduced, referred to as the diffusion model, and it can also be used for image generation. Like other generative models, it aims to generate target data samples from noise (sampled from a simple distribution).

The diffusion model comprises two processes: the forward process and the reverse process. The forward process, also known as the diffusion process, transforms the image X0 into a Gaussian distribution XT through T iterations by adding random Gaussian noise. The reverse process can be used to generate data samples by recovering X0 from XT through multiple denoising steps.

优化：

Recently, diffusion models [1] have achieved significant success in the field of image generation. The most common models for image generation are GAN and VAE. In 2020, DDPM (Denoising Diffusion Probabilistic Model) was introduced, known as the diffusion model, and it can be used for image generation as well. Like other generative models, its goal is to generate target data

samples from noise, which is sampled from a simple distribution.

The diffusion model consists of two processes: the forward process and the reverse process. The forward process, also referred to as the diffusion process, transforms the initial image X0 into a Gaussian distribution XT through T iterations by adding random Gaussian noise. The reverse process can be employed to generate data samples by recovering X0 from XT through multiple denoising steps

[1]Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851, 2020. 3

ddim:

回到2020年的十月，斯坦福大学的研究人员Jiaming Song提出了**DDIM**（Diffusion Denoising Implicit Model），在提升了DDPM采样效率的基础上，仅用50步就能达到1000步采样的效果。DDIM不仅实现了高效率的采样方法，其作为确定性的采样方法还为后续的研究开创了一种类似于**GAN Invesion**的方法，用于实现各种真实图像的编辑与生成。

继之而来的，是2021年五月OpenAI所发布的**“Classifier Guidance”**（亦被称为Guided Diffusion）。这篇论文提出了一项重要的策略，即通过基于分类器的引导来指导扩散模型生成图像。借助其他多项改进，扩散模型首次成功击败了生成领域的巨头“GAN”，同时也为OpenAI的**DALLE-2**（一个图像和文本生成模型）的发布奠定了基础。

2022年的四月，来自OpenAI的**DALLE-2**横空出世，通过利用扩散模型以及海量数据，DALLE-2呈现出了前所未有的理解和创造能力。

Stable Diffusion release,其工作更为强大的3D生成等领域，将图像生成再度推进，使其更加贴近人类需求。

在2020年10月，斯坦福大学的研究员宋佳明介绍了DDIM（Diffusion Denoising Implicit Model）。在DDPM的效率改进基础上，DDIM仅需50步就实现了1000步采样的效果。DDIM不仅实现了高效采样，还开创了一种类似GAN逆向的确定性采样方法，用于图像编辑和生成。

随后，在2021年5月，OpenAI发布了“分类器引导”（也称为Guided Diffusion）。这篇论文介绍了一种关键策略，即通过基于分类器的引导来指导扩散模型生成图像。结合各种其他增强措施，扩散模型成功超越了生成领域的巨头，特别是“GAN”。此外，这项工作为OpenAI随后发布的DALL-E 2奠定了基础，这是一种强大的图像和文本生成模型。

在2022年4月，OpenAI推出了DALL-E 2，利用扩散模型和大规模数据展示了前所未有的理解和创造能力。

2022年发布的“Stable Diffusion”进一步推动了图像生成，尤其是在3D生成等领域，使其更符合人类需求。

In October 2020, Jiaming Song, a researcher at Stanford University, introduced the **DDIM** (Diffusion Denoising Implicit Model). Building on the efficiency improvements of DDPM, DDIM achieved the effect of 1000-step sampling in just 50 steps. DDIM not only enabled efficient sampling but also pioneered a deterministic sampling method, reminiscent of **GAN Inversion**, for image editing and generation.

Following this, in May 2021, OpenAI released **"Classifier Guidance"** (also known as Guided Diffusion). This paper introduced a crucial strategy of guiding diffusion models to generate images using classifier-based guidance. Coupled with various other enhancements, diffusion models successfully surpassed the giants in the generative field, notably "GAN."

In April 2022, OpenAI unveiled **DALL-E 2**, which leveraged diffusion models and massive data to exhibit unprecedented levels of understanding and creative capabilities.

The release of Stable Diffusion in 2022 further propelled image generation, , making it more aligned with human needs.

优化去掉日期：

Jiaming Song, a researcher at Stanford University, introduced the **DDIM** (Diffusion Denoising Implicit Model). Building on the efficiency improvements of DDPM, DDIM achieved the effect of 1000-step sampling in just 50 steps. DDIM not only enabled efficient sampling but also pioneered a deterministic sampling method, reminiscent of **GAN Inversion**, for image editing and generation.

Following this, OpenAI released **"Classifier Guidance"** (also known as Guided Diffusion). This paper introduced a crucial strategy of guiding diffusion models to generate images using classifier-based guidance. Coupled with various other enhancements, diffusion models successfully surpassed the giants in the generative field, notably "GAN."

In April 2022, OpenAI unveiled **DALL-E 2**, which leveraged diffusion models and massive data to exhibit unprecedented levels of understanding and creative capabilities.

The release of Stable Diffusion in 2022 further propelled image generation, making it more aligned with human needs.

优化去掉加粗：

Jiaming Song, a researcher at Stanford University, introduced the DDIM (Diffusion Denoising Implicit Model). Building on the efficiency improvements of DDPM, DDIM achieved the effect of 1000-step sampling in just 50 steps. DDIM not only enabled efficient sampling but also pioneered a deterministic sampling method, reminiscent of GAN Inversion, for image editing and generation.

Following this, OpenAI released "Classifier Guidance" (also known as Guided Diffusion). This paper introduced a crucial strategy of guiding diffusion models to generate images using classifier-based guidance. Coupled with various other enhancements, diffusion models successfully surpassed the giants in the generative field, notably "GAN."

In April 2022, OpenAI unveiled DALL-E 2, which leveraged diffusion models and massive data to exhibit unprecedented levels of understanding and creative capabilities.

The release of Stable Diffusion in 2022 further propelled image generation, making it more aligned with human needs.

总的

Diffusion Models

Recently, diffusion models [1] have achieved tremendous success in the field of image generation. The most common models for image generation are GAN and VAE. In 2020, DDPM (Denoising Diffusion Probabilistic Model) was introduced, referred to as the diffusion model, and it can also be used for image generation. Like other generative models, it aims to generate target data samples from noise (sampled from a simple distribution).

The diffusion model comprises two processes: the forward process and the reverse process. The forward process, also known as the diffusion process, transforms the image X0 into a Gaussian distribution XT through T iterations by adding random Gaussian noise. The reverse process can be used to generate data samples by recovering X0 from XT through multiple denoising steps.

Jiaming Song, a researcher at Stanford University, introduced the DDIM (Diffusion Denoising Implicit Model)[2]. Building on the efficiency improvements of DDPM, DDIM achieved the effect of 1000-step sampling in just 50 steps. DDIM not only enabled efficient sampling but also pioneered a deterministic sampling method, reminiscent of GAN Inversion, for image editing and generation.

Following this, OpenAI released "Classifier Guidance" (also known as Guided Diffusion)[3]. This paper introduced a crucial strategy of guiding diffusion models to generate images using classifier-based guidance. Coupled with various other enhancements, diffusion models successfully surpassed the giants in the generative field, notably "GAN."

In April 2022, OpenAI unveiled DALL-E 2, which leveraged diffusion models and massive data to exhibit unprecedented levels of understanding and creative capabilities.

The release of Stable Diffusion in 2022 further propelled image generation, making it more aligned with human needs[4].

[1]Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851, 2020. 3

[2]Song J, Meng C, Ermon S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.

[3]Ho J, Salimans T. Classifier-free diffusion guidance[J]. arXiv preprint arXiv:2207.12598, 2022.

[4]Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

Plan

https://zhuanlan.zhihu.com/p/617017935

受ControlNet启发，我们提出CC_ControlNet.

我们计划实现CC_ControlNet，在文生图的基础加入额外的编辑方案，控制扩散模型，使其生成更接近用户需求。我们的主要目的是实现动漫角色或卡通角色等风格化角色的自定义换装。我们将分为几个阶段来实现，第一阶段，输入动漫角色图片、及角色图片及对需要替换的衣服进行手动涂抹做mask操作后的图片，以及一段对新衣装的描述，实现涂抹部分根据文本信息进行生成。第二阶段，可以给定特定服装，对特定角色进行换装，并自动检测服装位置，无需手动提供mask后的图片。、

Based on ControlNet inspiration, we propose CC_ControlNet.We plan to implement CC_ControlNet, incorporating additional editing schemes on the basis of document-based graph, controlling the diffusion model to generate outputs more closely aligned with user requirements. Our primary objective is to achieve custom outfitting for stylized characters such as anime or cartoon characters. We'll execute this in several stages. In the first phase, we'll input anime character images, images of the characters with manually edited mask operations for the clothes to be replaced, and a textual description of the new attire. This stage aims to generate the edited portions based on the provided textual information. In the second phase, specific attire will be applied to designated characters, automatically detecting clothing positions without the need for manually provided masked images.

我们计划实现CC_ControlNet，在文生图的基础加入额外的编辑方案，控制扩散模型，使其生成更接近用户需求。我们的主要目的是实现动漫角色或卡通角色等风格化角色的自定义换装。我们将分为几个阶段来实现，第一阶段，输入动漫角色图片、及角色图片及对需要替换的衣服进行手动涂抹做mask操作后的图片，以及一段对新衣装的描述，实现涂抹部分根据文本信息进行生成。第二阶段，可以给定特定服装，对特定角色进行换装，并自动检测服装位置，无需手动提供mask后的图片。、

OK1:

Drawing inspiration from ControlNet, we introduce CC_ControlNet.Our plan involves implementing CC_ControlNet, which incorporates additional editing schemes based on text-to-image generation. This will allow us to control the diffusion model, generating outputs that closely align with user requirements. Our primary goal is to achieve customized outfitting for stylized characters, such as anime or cartoon characters. We will carry out this project in multiple stages. In the initial phase, we will input anime character images, along with images of characters that have undergone manual mask operations for the clothes to be replaced. Additionally, we will provide a textual description of the new attire. The objective of this stage is to generate the edited portions based on the provided textual information. In the second phase, specific attire will be automatically applied to the characters, with clothing positions detected automatically, eliminating the need for manually provided masked images。

以Stable Diffusion为例，我们介绍在角色换装这个特定任务条件下，利用CC_ControlNet对大型扩散模型进行控制的方法。我们将使用 CC_ControlNet来控制Stable Diffusion的U-net的各个层。CC_ControlNet将通过操作神经网络的输入条件来控制神经网络的行为。我们将各个encoder层进行拷贝，而decoder部分进行skip connection。通过迭代的过程，**重复应用ControlNet操作来优化神经网络块。**

Using Stable Diffusion as an example, we introduce the method of controlling large-scale diffusion models with CC_ControlNet for the specific task of character outfit swapping. We will employ CC_ControlNet to regulate the various layers of the U-net in Stable Diffusion. CC_ControlNet will manipulate the behavior of the neural network by operating on its input conditions. We will duplicate each encoder layer and establish skip connections within the decoder section. Through

an iterative process, we will repeatedly apply ControlNet operations to optimize the neural network blocks.

总的：

Drawing inspiration from ControlNet, we introduce CC_ControlNet.Our plan involves implementing CC_ControlNet, which incorporates additional editing schemes based on text-to-image generation. This will allow us to control the diffusion model, generating outputs that closely align with user requirements. Our primary goal is to achieve customized outfitting for stylized characters, such as anime or cartoon characters. We will carry out this project in multiple stages. In the initial phase, we will input anime character images, along with images of characters that have undergone manual mask operations for the clothes to be replaced. Additionally, we will provide a textual description of the new attire. The objective of this stage is to generate the edited portions based on the provided textual information. In the second phase, specific attire will be automatically applied to the characters, with clothing positions detected automatically, eliminating the need for manually provided masked images。

Using Stable Diffusion as an example, we introduce the method of controlling large-scale diffusion models with CC_ControlNet for the specific task of character customization. We will employ CC_ControlNet to regulate the various layers of the U-net in Stable Diffusion. CC_ControlNet will manipulate the behavior of the neural network by operating on its input conditions. We will duplicate each encoder layer.In the decoder section, skip connections are established. Through an iterative process, we will repeatedly apply CC_ControlNet operations to optimize the neural network blocks.

Character Customization With Stable Diffusion