在文本到图像的扩散模型中添加条件控制

摘要

ControlNet 是一种神经网络架构，旨在将空间条件控制添加到大型预训练的文本到图像扩散模型中。ControlNet 将锁定预训练好的大型扩散模型，并重用由数十亿图像预先训练的强大主干中的深层和鲁棒编码层，以学习多样化的条件控制。本文基于 ControlNet 结构，以边缘控制作为条件，加入感知损失函数作为约束条件，使用一个新的小型服饰数据集训练了新的 ControlNet 模型。实验表明，新的 ControlNet 模型能够生成与原图更相似的图像，比原 ControlNet 模型生成效果要好。

关键词： ControlNet；条件控制；扩散模型

1 引言

随着文本到图像的扩散模型的出现，大型文本到图像生成模型的存在让人们意识到人工智能的巨大潜力，现在可以通过输入文本提示符来创建视觉上令人惊叹的图像。然而，文本到图像的模型在提供对图像空间构成的控制方面是有限的，仅通过文本提示来精确地表达复杂的布局、姿态、形状和形式可能是困难的。生成一个与我们的期望准确匹配的图像，通常需要多次反复的试验，比如编辑提示，检查生成的图像，然后重新编辑提示。

以端到端方式学习大型文本到图像扩散模型的条件控制是一项挑战。特定条件下的训练数据量可能明显小于一般文本-图像训练的数据。例如，针对各种特定问题（例如，物体形状，人体姿态提取等）的最大数据集，通常大小在 10 万左右，比用于训练 Stable Diffusion 的 LAION-5B [26] 数据集小 5 万倍。对数据有限的大型预训练模型进行直接微调或持续训练，可能会导致过拟合和灾难性遗忘。研究者已经表明，这种遗忘可以通过限制可训练参数的数量或排名来缓解。所以，设计更深层次或更多定制的神经结构可能是处理具有复杂形状和不同高级语义的条件生成图像的必要条件。

本文提出了一种控制网络，一种端到端神经网络架构，用于为大型预训练的文本到图像扩散模型（Stable Diffusion 模型中进行实验）学习条件控制。ControlNet 通过锁定其参数，并制作其编码层的可训练副本，从而保持了大型模型的质量和能力。该体系结构将大型预训练模型视为学习不同条件控制的强大主干。可训练的副本和原始的锁定模型与零卷积层连接，权值初始化为零，使它们在训练过程中逐渐增长。这种体系结构确保了在训练开始时不将有害的噪声添加到大扩散模型的深层特征中，并保护了可训练副本中的大规模预训练主干不被这种噪声破坏。实验表明，ControlNet 可以通过各种条件输入来控制稳定的扩散，包括边缘、霍夫线、用户涂鸦、人类关键点、分割图、形状法线、深度等。

ControlNet 以端到端的方式学习特定于任务的条件，即使训练数据集很小（<50k），学习也是稳健的，能够在个人设备上进行训练。并且，使用 ControlNet 对 Stable diffusion 进行

了增强，实现条件输入，生成高质量、详细的图像。因此，本文使用 ControlNet 结构，以边缘作为输入条件，在一个较小的服饰相关数据集上训练，通过对输入条件的控制生成更细致的服装。

2 相关工作

2.1 微调神经网络

调整神经网络的一种方法是直接用额外的训练数据继续训练它。但这种方法可能会导致过拟合、模式崩溃和灾难性的遗忘。广泛的研究集中在制定避免微调策略的问题。

超网络是一种起源于自然语言处理（NLP）领域的方法，目的是训练一个小的递归神经网络来影响一个较大的权重。它已被应用于生成式对抗网络（GANs）的图像生成。Heathen 等人 [5] 和 Kurumuz [9] 实现了 Stable Diffusion 的超网络，以改变其输出图像的艺术风格。

适配器方法在计算机视觉中，适配器用于增量学习 [21] 和领域自适应 [26]。该技术经常与 CLIP [17] 一起使用，用于将预先训练好的主干模型转移到不同的任务中。最近，适配器在视觉 transformer [11] 和视频适配器上取得了成功的结果。在与 ControlNet 的同步工作中，T2I 适配器 [14] 使 Stable Diffusion 适应外部条件。

附加学习通过冻结原始模型的权值和使用学习到的权重掩模、剪枝或硬性注意力添加少量新参数来规避遗忘。Side-Tuning [25] 使用侧分支模型来学习额外的功能，通过线性混合冻结模型和添加的网络的输出和预定义的混合权重计划。

低秩自适应 (LoRA) 基于许多过参数化模型位于低内在维子空间中的观察，通过学习参数的偏移量来防止灾难性遗忘 [6]。

零初始化层被控制网用于连接网络块。对神经网络的研究已经广泛地讨论了网络权值的初始化和操作。例如，权值的高斯初始化可能比用零初始化的风险更小。最近，Nichol 等人 [10] 讨论了如何在扩散模型中缩放卷积层的初始权值来改进训练，他们的“零模块”的实现是将权值缩放到零的一个极端情况。稳定性的模型 cards [28] 也提到了在神经层中使用零权重。在 ProGAN [7] 和 StyleGAN [8] 中也讨论了对初始卷积权值的操作。

2.2 图像扩散

图像扩散模型首先由 Sohl-Dickstein 等人 [27] 提出，最近已应用于图像生成。潜扩散模型（LDM） [19] 在潜图像空间中执行扩散步骤，降低了计算成本。文本-图像扩散模型通过 CLIP 等预先训练的语言模型将文本输入编码到潜在向量中，从而实现了最先进的图像生成结果。Glide [15] 是一个支持图像生成和编辑的文本引导扩散模型。Stable Diffusion 是潜在扩散 [19] 的一个大规模实现。Imagen [24] 直接扩散像素使用金字塔结构，而不使用潜在的图像。商业产品包括 DALL-E2 [16] 和 Midjourney [13]。

可控图像扩散模型有助于个性化、定制或特定于任务的图像生成。图像扩散过程直接提供了对颜色变化和插入绘制的一些控制。文本引导的控制方法侧重于调整提示、操作 CLIP 特性和修改交叉注意。MakeAScene [4] 将分割掩码编码为标记来控制图像的生成。SpaText [1] 将分割掩码映射到本地化的标记嵌入中。GLIGEN [12] 在扩散模型的注意层中学习新的参数，用于更真实的生成。Textual Inver-sion [3] 和 DreamBooth [22] 可以通过使用一小组用户提供

的示例图像来微调图像扩散模型来个性化生成的图像中的内容。基于提示的图像编辑提供了使用提示来操作图像的实用工具。Voynov 等人 [30] 提出了一种优化方法，适合扩散过程的草图。

2.3 图像到图像的翻译

条件 GANs [32] 和 transformers 可以学习不同图像域之间的映射，例如，Taming Transformer [2] 是一种视觉 transformer 方法；Palette [23] 是一个从头开始训练的条件扩散模型；PITI [31] 是一个基于预训练的图像到图像转换的条件扩散模型。操作预先训练好的 GANs 可以处理特定的图像到图像的任务，例如，StyleGANs 可以由额外的编码器 [18] 控制。

3 本文方法

3.1 本文方法概述

ControlNet 是一种神经网络体系结构，它可以增强具有空间定位、特定任务的图像条件的大型预训练的文本到图像扩散模型。我们首先在第 3.1 节中简单描述了本文的方法，然后在第 3.2 节中介绍了 ControlNet 的基本结构，在第 3.3 节中描述了如何将 ControlNet 应用于图像扩散模型的 Stable Diffusion [19]。在第 3.4 节中阐述模型的损失函数。

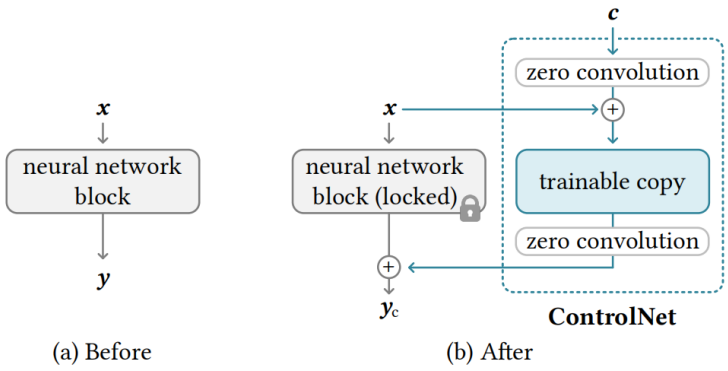

图 1. 方法示意图

3.2 ControlNet

ContolNet 向神经网络块注入额外的条件（图1），一个神经块以一个特征映射 x 作为输入，并输出另一个特征映射 y，如 (a) 所示为了向这样的块添加一个控制网，锁定原始块并创建一个可训练的副本，并使用零卷积层将它们连接在一起，即 1×1 卷积，权重和偏差都初始化为零。这里 c 是希望添加到网络中的条件反射向量，如 (b) 所示。在此，使用术语网络块来指一组神经层，这些神经层通常组合在一起形成一个神经网络单元，如 resnet 块、conv-bn-relu 块、多头注意块、transformer 块等。假设 $\mathcal{F}(\cdot;\Theta)$ 是这样一个训练过的神经块，参数为 Θ，它将输入特征映射 x 转换为另一个特征映射 y 为:

$$y = \mathcal{F}(x;\Theta). \tag{1}$$

在本文设置中，x 和 y 通常是二维特征图，即 $\boldsymbol{x} \in \mathbb{R}^{h\times w\times c}$ 分别为图中通道的高度、宽度和数量（图1a）。

为了向这样一个预先训练过的神经块添加一个控制网的神经块，本文锁定（冻结）原始块的参数 Θ，并同时将该块克隆到一个具有参数 Θ_c 的可训练副本中（图1b）。可训练的副本以一个外部条件反射向量 c 作为输入。当这种结构应用于像 Stable Diffusion 这样的大模型时，锁定参数保留了用数十亿张图像训练好的模型，而可训练副本重用这种大规模的预训练模型来建立一个深度、健壮和强大的主干，以处理不同的输入条件。

可训练的副本连接到具有零卷积层的锁定模型，表示为 $\mathcal{Z}(\cdot;\cdot)$。具体来说，$\mathcal{Z}(\cdot;\cdot)$ 是一个 1×1 的卷积层，其权值和偏差初始化为零。为了建立一个 ControlNet，本文分别使用了两个具有参数 Θ_{z1} 和 Θ_{z2} 的零卷积。然后用完整的 ControlNet 进行计算：

$$\boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x};\Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c};\Theta_{z1});\Theta_c);\Theta_{z2}) \tag{2}$$

其中，$\boldsymbol{y}_c$ 是 ControlNet 块的输出。在第一步训练中，由于零卷积层的权值和偏差参数都初始化都为零，因此式 (2) 中的 $\mathcal{Z}(\cdot;\cdot)$ 项都计算为零，并且

$$\boldsymbol{y}_c = \boldsymbol{y}. \tag{3}$$

这样，当训练开始时，有害噪声就不会影响可训练副本中神经网络层的隐藏状态。此外，由于 $\mathcal{Z}(\boldsymbol{c};\Theta_{z1}) = 0$ 和可训练副本也接收输入的图像 x，可训练副本有全部功能的，并保留了大型的预训练模型的能力，允许其作为进一步学习的强大骨干。零卷积通过消除在初始训练步骤中作为梯度的随机噪声来保护这个主干。

3.3 ControlNet 用于文本到图像扩散模型

本文使用 Stable Diffusion [19] 为例来展示控制网如何将条件控制添加到一个大型的预训练扩散模型中，如图2所示，图中 Stable Diffusion 的 U-net 架构与编码器块和中间块上的控制网连接。锁定的灰色块显示了 Stable Diffusion V1.5（或 V2.1 的结构，因为它们使用相同的 U-net 体系结构）。添加可训练的蓝色块和白色的零卷积层来构建一个控制网。Stable Diffusion 本质上是一个带有一个编码器、一个中间块和一个跳跃连接解码器的 U-Net [20]。编码器和解码器都包含 12 个块，完整的模型包含 25 个块。在 25 个块中，8 个块是下采样或上采样卷积层，而其他 17 个块是主块，每个块包含 4 个 resnet 网层和 2 个视觉 Transformers（ViTs）。每个 ViT 都包含了几种交叉注意力和自我注意力机制。例如，在图2a 中，“SD 编码器块 A”包含 4 个 resnet 层和 2 个 ViTs，而“×3”表示该块重复了三次。文本提示使用 CLIP 文本编码器进行编码，而扩散时间步长使用位置编码的时间编码器进行编码。

ControlNet 结构应用于 U-net 的每个编码器级别（图2b）。特别是，使用 ControlNet 创建了 Stable Diffusion 的 12 个编码块和 1 个 Stable Diffusion 中间块的可训练副本。12 个编码块有 4 个分辨率（64×64,32×32,16×16,8×8），每个分辨率复制 3 次。输出被添加到 U-net 的 12 个跳过连接和 1 个中间块中。由于 Stable Diffusion 是一个典型的 U-net 结构，这种 ControlNet 结构可能适用于其他模型。

连接 ControlNet 的方式在计算上是高效的——锁定副本的参数被冻结，在微调过程中，原始锁定编码器不需要进行梯度计算。这种方法加快了训练速度并节省 GPU 内存。

图像扩散模型学习逐步去噪图像并从训练域生成样本。去噪过程可以在像素空间或从训练数据编码的潜在空间中进行。Stable Diffusion 使用潜在图像作为训练域。Stable Diffusion 将 512 × 512 像素空间的图像转换为较小的 64 × 64 潜在图像。要将 ControlNet 添加到 Stable

Diffusion 中，首先将每个输入条件图像（例如边缘、姿势、深度等）从 512 × 512 的输入大小转换为与 Stable Diffusion 大小相匹配的 64 × 64 特征空间向量。使用一个具有四个卷积层的小网络 E(·)，这些卷积层具有 4 × 4 内核和 2 × 2 步长（通过 ReLU 激活，使用 16、32、64、128 个通道，分别初始化高斯权重并与其他完整模型联合训练），将图像空间条件 c_i 编码为特征空间条件向量 c_f。c_f 条件向量被传递到 ControlNet 中。

$$c_{\mathrm{f}} = \varepsilon(c_i). \tag{4}$$

条件向量 c_f 被传递到 ControlNet 中。

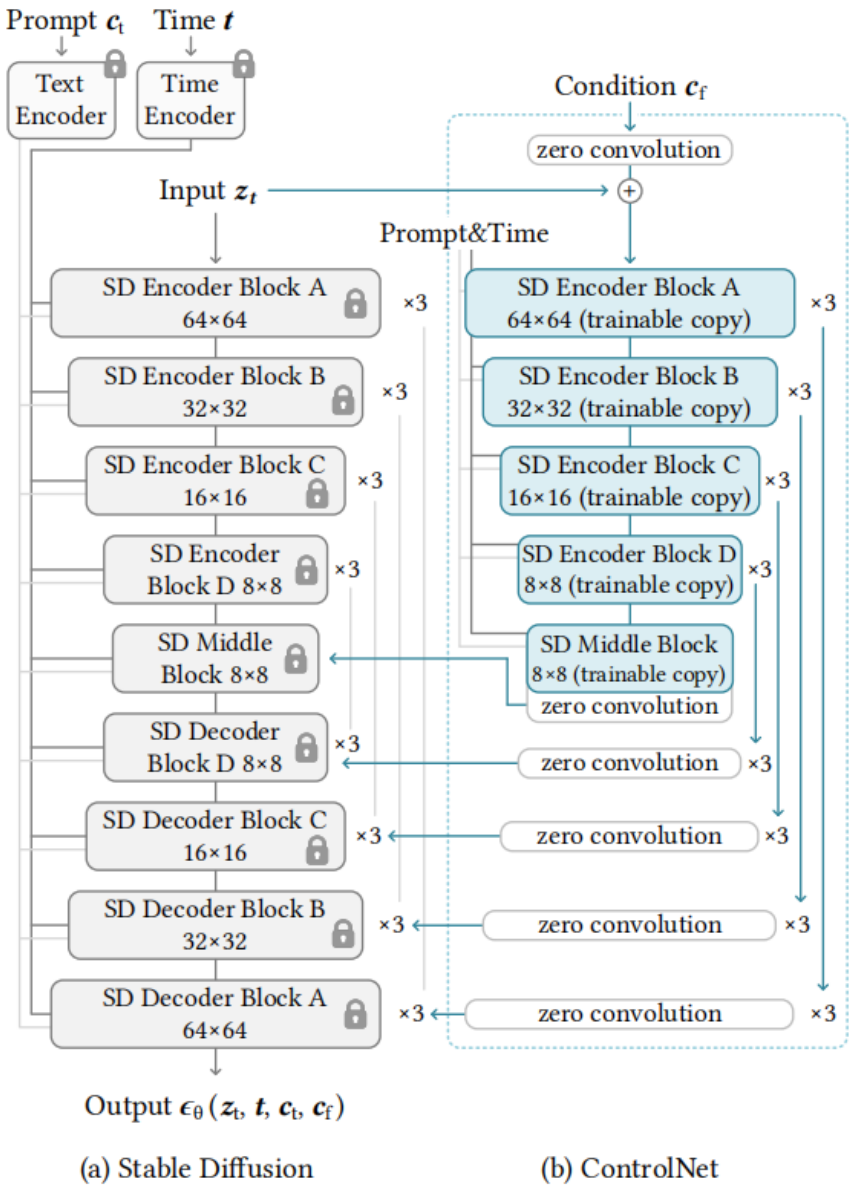

图 2. ControlNet 用于扩散模型方法示意图

3.4 损失函数定义

给定一个输入图像 z_0，图像扩散算法会逐步向图像添加噪声，并生成一个噪声图像 z_t，其中 t 表示添加噪声的次数。给定一组条件，包括时间步长 $\boldsymbol{t}$、文本提示 c_t 以及特定于任务的条件 c_f，图像扩散算法会学习一个网络 ϵ_θ 来预测添加到噪声图像 z_t 上的噪声。

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)} \Big[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f})) \|_2^2 \Big]. \tag{5}$$

其中，$\mathcal{L}$ 为整个扩散模型的整体学习目标。该学习目标直接用于用 ControlNet 微调扩散模型。

在训练过程中，本文随机用空字符串替换 50% 的文本提示 c_t。这种方法增加了控制网直接识别输入条件反射图像中的语义的能力（例如，边缘、姿态、深度等），作为提示符的替代品。

4 复现细节

4.1 与已有开源代码对比

本文的开源代码地址：https://github.com/lllyasviel/ControlNet。
使用开源代码中 ControlNet 模型的可训练副本框架代码，自己编写其他部分内容的代码：

- Timestep Embedding 模块将时间步长编码成长为 1280 的嵌入。
- HintBlock 模块在输入的图像与其他特征融合前先提取特征。
- ResBlock 模块融合时间步的嵌入和上一层的输出。
- SpatialTransformer 模块融合提示嵌入和上一层的输出。
- SD Encoder 是 Stable Diffusion 编码阶段的模块，实现 ResBlock 和 SpatialTransformer 的堆叠实现了时间步长、提示图像、和提示嵌入的特征融合，并进行下采样增加特征图的通道数。
- SD Decoder 是 Stable Diffusion 解码阶段的模块，实现了时间步长、提示图像、和提示嵌入的特征融合，进行上采样减少特征图的通道数，冻结 SD Encoder 和 Decoder 部分的参数。
- 最后将感知损失函数加入到源代码的损失函数中。

4.2 实验环境搭建

- 服务器版本：Ubuntu 16.04.1 LTS
- GPU：Tesla P100
- python：3.8.5
- PyTorch: 1.13.1
- torchvision：0.13.1
- opencv-python：4.8.1.78

- transformers：4.35.2
- diffusers：0.14.0

4.3 ControlNet Demo 界面分析与使用说明

图3是 ControlNet 模型的测试界面展示图，在图中，1 处输入边缘提示图像，2 处输入描述文本，然后点击 Run，就会在 3 处显示生成的图像。

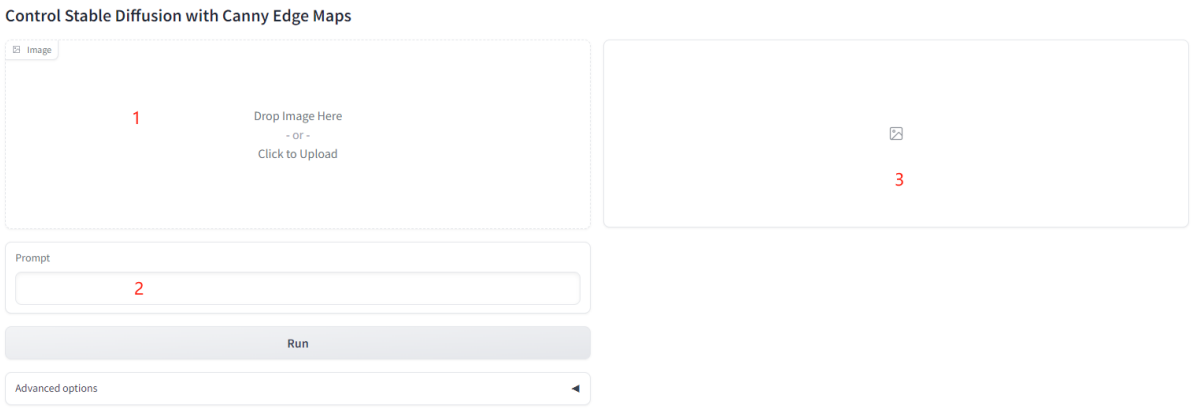

图 3. 操作界面示意

4.4 创新点

数据预处理由于本文训练 ControlNet 模型使用的数据集没有公开而且数据量太大，在现有的实验室机器上无法实现所用数据集的重新训练，考虑到实验室服务器的限制，仅使用了服装数据集 Re-PolyVore 中的上衣和裙子数据，共 11307 张图片。为了使用 ControlNet 模型生成更细粒度的图片，对每一张图片都加以文字描述作为输入的提示（原数据集不带文字说明），使用 BLIP-2 方法获取图片的文字描述，BLIP-2 是一种通用且计算效率高的视觉语言预训练方法，它利用了冻结的预训练图像编码器和大语言模型，可以很好的理解图像包含的内容。ControlNet 模型可以测试包括 Canny Edge、Depth Map、Normal Map、M-LSD lines、HED soft edge、ADE20K segmentation、Openpose 和用户草图作为条件输入，功能强大。鉴于是在服装数据集上训练该模型，则只选择了 Canny Edge 作为输入的控制条件进行训练。训练数据包含 3 种文件，原图、边缘图和对应的文本描述，以下是实验对数据进行文本描述提取和 Canny Edge 处理的效果示例图，如图4所示：

但是经过仔细核对 BLIP-2 提取的文字描述信息和原始的图片内容，存在少数文本描述包含了不属于图像内容的信息，例如图4中第四个示例的文本描述为：a woman wearing a red dress with bow tie，但是原图中只有一条裙子没有人的存在。总的来说，BLIP-2 是能够获取图像中服饰的关键特征的，同时需要探索另一种方案对文本描述进行筛选清洗，进一步提升文本和图像的一致性。

图 4. 实验数据示意

目标损失如第 3.4 节所述，扩散模型隐式地学习从高斯噪声中重建图像。网络 ϵ_θ 估计当前输入的噪声图像 x_t 中的噪声。DDIM 的训练目标 (等式 (5))，并没有明确地处理条件约束。因此，除了使用等式 (5) 外，还使用了感知损失（等式（6)），以控制图像的合成。为此，根据估计的噪声 $\hat{\epsilon}_t$，在每个时间步 t 得到重构图像 $\hat{x}_0$，感知损失的计算方法如下：

$$\mathcal{L}_t^{\text{perc}} = \mathbb{E}_m \left\| \psi_m \left(\hat{\mathbf{x}}_0 \right) - \psi_m \left(\mathbf{x}_0 \right) \right\|_2 . \tag{6}$$

其中，ψ_m 表示 VGG 的第 m 层。根据 [29]，在等式 (6) 中使用了 $relu1\_2$ $relu2\_2$ $relu3\_2$ $relu4\_2$ 和 $relu5\_2$ 的层。基于感知损失的总体训练目标如下：

$$\mathcal{L}_{all} = \mathcal{L} + \lambda_p \mathcal{L}_t^{\text{perc}}. \tag{7}$$

其中，λ_p 是感知损失的平衡权重。在实验中，将等式中 λ_p=0.01。

5 实验结果分析

可视化分析本文基于 11307 张上衣和裙子数据，训练加入了感知损失的 ControlNet 模型，并将其命名为 ControlNet-P 模型。由于服务器资源有限，训练的 batch size 大小只能设为 1，所以训练一个 epoch 需要大约 6 小时。当训练一个 epoch 时，生成的图片还包含很多其他多余背景，所以训练 1 万步是不够的。当训练达到 30 个 epoch 时，能够很好的学习到服装的特点和外观，但是进一步加大 epoch 训练时，模型对服装的颜色进行过度解读，导致生成的衣服颜色完全不一致。最终将训练了 30 个 epoch 的 ControlNet-P 模型保存下来，图5和6是使用训练好的 ControlNet-P 进行测试，并和原来训练好的 ControlNet 模型生成效果进行比较。

图 5. 实验结果示意

图 6. 实验结果示意

从图5和图6中可以看出 ControlNet-P 模型的生成效果要比 ControlNet 生成的与原图更相似，ControlNet-P 模型是可以学习到原始服饰的细节的，对一些比较明显突出的图案样式能够很好的学习到它们的特征，例如图5中的熊猫图案。但是对于设计更复杂的图案样式该模型的学习能力还是有所欠缺的，例如图6中第三个例子生成的菱形之间没有黑色边界，衣领上的纹路也没有学习到。所以对于服饰的 ControlNet 模型，还需要进一步思考怎样才能让其学习到更加细粒度的图案样式，并且要避免颜色的失真。

表 1. 定量分析

方法	FID↓	LPIPS↓	CLIP-score↑
ControlNet-P	**42.37**	**0.6912**	25.60
ControlNet	43.81	0.6986	**27.95**

定量评价表1显示了定量评价，其中使用 FID、LPIPS 和 CLIP-score 三个指标来评估 ControlNet-P 的性能，并与 ControlNet 作比较。FID 和 LPIPS 测量特征空间的距离，与 FID 关注的总体分布统计生成/合成图像和地面真理，而 LPIPS 计算每对合成图像之间的距离和相应的真实图像，更低的 FID 和 LPIPS 值代表更高的图像质量。相比之下，CLIP-score 衡量语义对应，即合成图像与其对应文本描述之间的余弦相似度，得分越高表示对齐效果越好。与原 ControlNet 相比，ControlNet-P 在数据集上表现要好，FID 和 LPIPS 均比前者低，但是 CLIP-score 分数前者较高，可见使用 BLIP-2 来提取图片信息往往是不充分的。

6 总结与展望

总的来说，ControlNet 是一种新颖的神经网络架构，可以通过微调预训练模型来适应特定任务。该方法使用条件控制技术将输入条件与预训练模型进行连接，并将其作为额外的输入信息传递给神经网络，从而帮助神经网络更好地理解输入条件，并取得更好的效果。它重用源模型的的大规模预训练层来构建一个深度且强大的编码器，以学习特定条件。原始模型和可训练的副本通过“零卷积”层连接，这些“零卷积”层可以消除训练期间的噪音。同时，ControlNet 还探讨了不同因素对其性能的影响，例如不同输入条件、不同预训练模型和不同微调策略等。此外，使用了一个新的服饰数据集并构建其边缘图像和文本提示，给 ControlNet 模型加上感知损失函数，来训练新的 ControlNet 模型。实验结果表明，新的 ControlNet 模型能够生成与原图更相似的图像，比原 ControlNet 模型生成效果要好。

不足之处在于，使用 BLIP-2 方法对服饰进行描述还是不能够获得非常精准的，且能充分描述细节的文本信息，由于资源的限制模型训练的数据集数量偏少，可能导致模型的泛化性能较低。在将感知损失函数加入到总损失函数后，新的 ControlNet 模型虽然能够生成与原图更相似的服饰，然而还是没有很好的解决细粒度生成的问题，文本描述和原图的一致性也有待提高。

ControlNet 提供了多种可控的生成方式，使得用户可以更好的根据自己的需求来生成图像，如何充分利用多种可控条件和模型中的注意力机制来实现更细粒度的服饰生成，并结合使用 BLIP-2 之类的图生文方法提取文本提示信息，是未来需要解决的关键问题，这有利于服装设计师们仅用文字描述和简单的条件控制（如边缘）就能够很快实现与期望相符的服饰设计。

参考文献

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for

controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023.

[2] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[4] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[5] Heathen. Hypernetwork style training, a tiny guide,stable-diffusion-webui, 2022. https://github.com/automatic1111/stable-diffusion-webui/discussions/2670.

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[9] Kurumuz. Novelai improvements on stable diffusion, 2022. https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac.

[10] Wenjie Li, Chi-hua Wang, Guang Cheng, and Qifan Song. International conference on machine learning. *Transactions on machine learning research*, pages 8162–8171, 2023.

[11] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[12] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[13] Midjourney, 2023. https://www.midjourney.com/.

[14] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[15] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[16] OpenAI. Dall-e-2, 2023. https://openai.com/product/dall-e-2.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[18] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[21] Amir Rosenfeld and John K Tsotsos. Incremental learning through deep adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):651–663, 2018.

[22] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[23] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[25] Alexander Sax, Jeffrey Zhang, Amir Zamir, Silvio Savarese, and Jitendra Malik. Side-tuning: Network adaptation via additive side networks. 2019.

[26] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[28] Stability. Stable diffusion v2 model card, stable-diffusion-2-depth, 2022. https://huggingface.co/stabilityai/stable-diffusion-2-depth.

[29] Zhengwentai Sun, Yanghong Zhou, Honghong He, and PY Mok. Sgdiff: A style guided diffusion model for fashion synthesis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8433–8442, 2023.

[30] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[31] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

[32] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.