基于扩散模型的姿态引导人物图像合成

——论文复现：Person Image Synthesis via Denoising Diffusion Model

摘要

本文研究了基于去噪扩散模型（DDPM）的人物图像合成问题。具体来说，给定一张输入图像，以及目标姿态图（pose map），希望能在 pose map 的引导下，输出一张既符合 pose map，又和输入图像有着相同风格的新视图。由于姿态引导的人物图像生成任务需要合成任意姿态的逼真图像，而所产生的输出通常会经历变形的纹理和不现实的身体形状，尤其是在合成封闭的身体部位时，难以处理复杂的姿态和严重的遮挡。因此在新的姿态中保持连贯的结构、外观和整体的身体组成是一项具有挑战性的任务。现有方法往往使用生成对抗网络（GAN）来实现，而本文复现的论文 [1] 中，首次将扩散模型应用到人物图像生成任务中，将复杂的转移问题分解为一系列更简单的去噪扩散步骤，逐步将源图像中的一个人转移到目标姿态。该方法可以模拟人的姿势和外观之间的复杂相互作用，提供更高的多样性。然而，本文在复现过程中，发现 [1] 方法模型也存在一些问题，并对此提出了一些优化方法与思想。

关键词：扩散模型；人物图像合成；姿态引导

1 引言

随着计算机视觉和图像处理领域取得的显著进展，人物姿态转移问题也得到了广泛的研究。姿势引导的人物图像合成任务 [20, 25] 旨在根据所需的姿势和外观来渲染一个人的图像。具体而言，给定一张输入图像，以及目标姿态图（pose map），希望能在 pose map 的引导下，输出一张既符合 pose map，又和输入图像有着相同风格的新视图。如图 1所示，外观是由给定的源图像定义的，而姿态则是由一组关键点定义的。

姿态引导的人物图像合成任务在多个领域中具有广泛的应用和重要的意义。首先，在虚拟现实和增强现实应用中，合成逼真的人物图像是实现更真实、更沉浸式用户体验的关键因素。通过根据姿态引导生成的图像，用户可以在虚拟环境中看到与自己的真实姿态和动作相匹配的虚拟人物，增强了用户的身临其境感。其次，在电子商务和时尚行业中，通过该技术，用户可以根据自己的需求和喜好，预览不同姿态下穿着不同服装的效果。这对于在线购物体验的改进和个性化定制具有重要意义，能够帮助用户更好地选择适合自己的服装款式和尺寸。此外，在安全监控和人员重新识别等领域，通过合成具有不同姿态和外观的人物图像，可以增加训练数据的多样性，从而提高人员重新识别算法的准确性和鲁棒性。

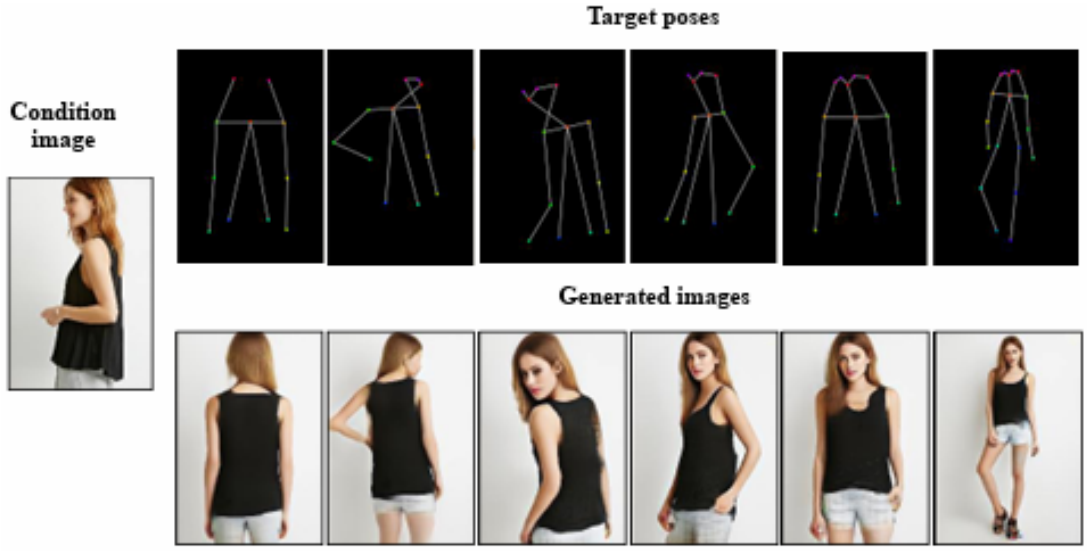


图 1. 姿态引导的人物图像合成任务，图引自 [25]

然而，姿态引导的人物图像生成任务需要合成任意姿态的逼真图像，关键是通过对姿势和风格的控制来合成人物图像，这是一个具有挑战性的任务。首先，需要解决姿态和外观之间的复杂关系，确保合成图像与目标姿态的一致性。其次，合成图像的质量和真实感对于用户体验至关重要，需要通过先进的图像合成算法和技术来实现高度逼真的输出。

现有方法往往使用生成对抗网络（GAN） [5] 来实现，它试图使用一次正向传递来生成一个处于期望姿势的人。然而，在新的姿态中保持连贯的结构、外观和整体的身体组成是一项具有挑战性的任务。所产生的输出通常会经历变形的纹理和不现实的身体形状，尤其是在合成封闭的身体部位时，难以处理复杂的姿态和严重的遮挡。此外，在生成任务方面，主要的生成模型还有 VAE [4] 和 Flow（NICE） [9]。它们都试图通过一次向前传递将源图像的风格直接转移到给定的目标姿态中，直接捕捉空间转换的复杂结构，这样往往难以转移布料纹理图案的复杂细节。

而在本文复现的论文 [1] 中，作者基于这些观察结果，提出使用连续的中间转移步骤获得最终的图像，可以使学习任务更简单。因此首次将扩散模型 [6] 应用到人物图像生成任务中，将复杂的转移问题分解为一系列更简单的去噪扩散步骤，逐步将源图像中的一个人转移到目标姿态，而不是一次性建模复杂的转移轨迹。该方法可以模拟人的姿势和外观之间的复杂相互作用，提供更高的多样性，并在没有纹理变形的情况下产生逼真的结果，如图 2所示，它将姿态转换过程分解为几个条件去噪扩散步骤，其中每个步骤都相对容易建模。

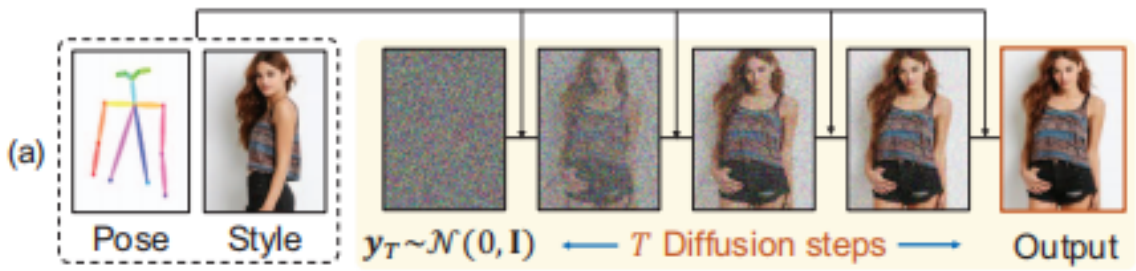


图 2. 基于扩散模型的人物图像合成任务，图引自 [1]

2 相关工作

2.1 姿态引导的人物图像合成

近年来，姿态引导的人物图像合成问题已经成为广受研究关注的领域，尤其是基于生成对抗网络（GAN）的模型 [5] 在条件图像合成方面取得了巨大的突破。早期的研究尝试 [13] 提出了一种由粗到细的方法，首先生成一个具有目标姿态的初步图像，然后通过反向细化的方式逐步改进结果。然而，这种方法简单地将源图像、源姿态和目标姿态连接在一起作为输入，导致了特征错位的问题。为了解决这个问题，Essner 等研究者 [4] 尝试采用基于变分自编码器（VAE）的设计和基于 U-Net 的跳跃连接架构来处理人物图像的外观和姿态。另外，Siarohin 等研究者 [18] 引入了可变形的跳跃连接，以对纹理进行空间变换，并通过一系列局部仿射变换来分解整体变形。

随后，一些研究工作 [8, 10, 17] 采用基于流的变形方法来转换源信息，以改善姿态对齐。其中 [8, 10] 使用几何模型将三维网格人体模型拟合到二维图像上，并预测三维流场以扭曲源图像的外观。而 [17] 提出了 GFLA 方法，用于生成全局流场和遮挡掩模，以扭曲源图像的局部补丁以匹配所需的姿态。另一方面，Zhen Zhu 等人 [25] 提出了一种无需任何变形操作的方法，通过一系列传输块逐步转换源图像。然而，在多次传输过程中，有用的信息可能会丢失，导致细节模糊化。ADGAN [14] 利用纹理编码器提取人体部位的样式向量，并通过多个 AdaIN 残差块合成最终图像。类似的方法，如 PISE [20]、SPGnet [12] 和 CASD [24]，利用解析映射来生成最终图像。CoCosNet [21, 23] 则通过基于注意力的操作提取跨域图像之间的密集对应关系。最近，Ren 等人 [16] 提出了一种基于神经纹理提取和分布操作的框架 NTED，取得了卓越的效果。

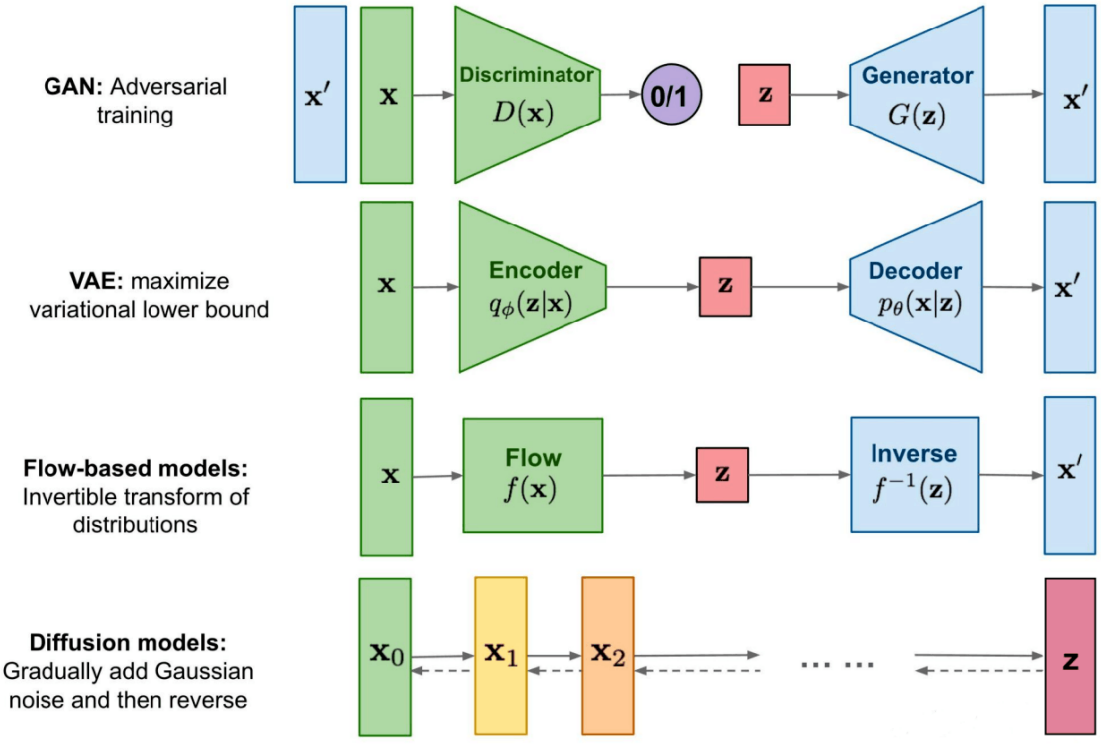


图 3. 生成模型主要框架对比

2.2 扩散模型

扩散模型最初在 2015 年的 [19] 中提出。其基本思想是，通过迭代的正向扩散过程破坏数据分布中的结构，然后学习一个反向扩散过程以恢复数据中的结构，即算法核心为估计一个马尔可夫扩散链的反转。但在当时，这个扩散模型并没有立刻得到广泛的关注。直至 2020 年的 [6] 提出去噪扩散概率模型（DDPM）在更加庞大的数据集上展现出与当时最优秀的生成对抗网络 GAN 模型相媲美的性能，于是引起了扩散模型的浪潮。图 3简要展示了其他生成模型与扩散模型的区别。可以看到，前三种生成模型均从隐变量 Z 生成目标数据 X。它们假设隐变量服从某种常见的概率分布（比如正态分布），然后希望训练一个模型 $X = g(Z)$ 将原来的概率分布映射到训练集的概率分布，也就是分布的变换，其本质都是概率分布的映射。而 Diffusion model 将 Z 作为类似于变分后验的马尔可夫链的平稳分布。

DDPM [6] 基于非平衡热力学的扩散原理，缓慢地添加噪声到输入样本（正向传递），然后从噪声（反向传递）重建所需的样本。如图 4，在前向扩散过程 q 中（虚线部分），在原始图像 X_0 上逐步增加噪声，每一步得到的图像 X_i 只和上一步的结果相关，即 $q(X_t|X_{t-1})$，直至第 T 步的图像 X_T 变为纯高斯噪声图。这个过程主要是训练过程，训练 U-Net 网络预测噪点的能力。在反向去噪扩散过程 p 中（实线部分），是生成图像的过程，图像生成是靠不断地去除噪点。首先给定一个全高斯噪点图，通过训练好的 U-Net 网络预测的噪声逐步去噪，直至最终复现出图像 X_0。

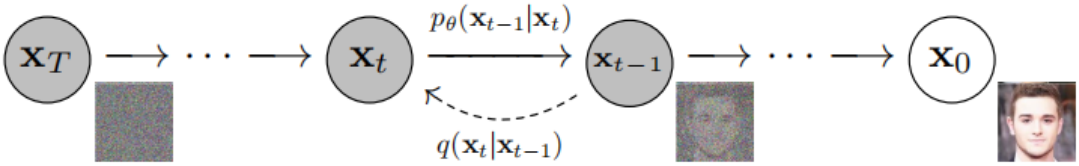


图 4. DDPM 的扩散过程，图引自 [6]

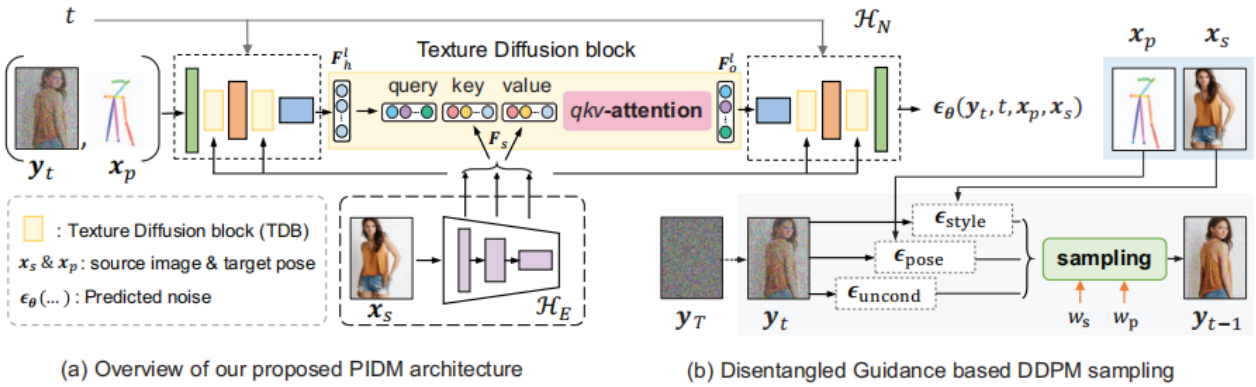
扩散模型可以合成高质量的图像。在无条件生成的成功之后，这些模型被扩展到在条件生成设置中工作，显示出比 GANs 具有竞争力甚至更好的性能。对于类条件生成，Dhariwal 等人 [3] 引入了分类器引导扩散，以使条件反射对 CLIP 文本表示。最近， [7] 提出了一种无分类器指导方法，可以在不需要分类器预训练的情况下进行调节。

3 本文方法

3.1 本文方法概述

基于现有相关方法在姿态引导的人物图像合成任务中的优缺点，以及新起的扩散模型，在本文复现的论文 [1] 中，作者提出了一个基于扩散模型的人物图像合成框架（PIDM），如图 5所示。总的来看，PIDM 由两大核心模块组成，即用于执行去噪扩散过程的 $\mathcal{H}_N$ 和提取纹理信息的 $\mathcal{H}_E$。x_s 为输入图像，y_t 表示第 t 步对应的噪声图像，x_p 表示用于引导视角生成的姿态图 pose map。为了在去噪过程中穿插输入图像 x_s 的风格，使网络能够更好地利用输入图像和输出图像之间纹理的对应关系，作者设计了 TDB 模块，即交叉注意力层。其中总共采用了五个 TDB 块，并用在 U-Net 三个特征分辨率的位置。对于第 t 步去噪过程，令 $\mathcal{H}_N$ 中第

l 层输出的噪声特征为 F_h^l，并从中得到 Attention 模块的 Q(query)。而 K(keys) 和 V(value) 则来自提取到的多尺度特征 F_s。



(a) Overview of our proposed PIDM architecture (b) Disentangled Guidance based DDPM sampling

图 5. 基于扩散模型的人物图像合成框架（PIDM），图引自 [1]

3.2 纹理条件下的扩散模型

PIDM 的生成建模方法是基于去噪扩散概率模型（DDPM） [6]。其思想是设计一个扩散过程，逐渐向从目标分布 $y_0 \sim q(y_0)$ 采样的数据中添加噪声，而后向去噪过程则试图学习反向映射。去噪扩散过程最终将一个各向同性高斯噪声 $y_T \sim \mathcal{N}(0, I)$ 转换为第 T 步中的目标数据分布。DDPM 的正向扩散过程是一个具有以下条件分布的马尔可夫链:

$$q(y_t|y_{t-1}) = \mathcal{N}(y_t; \sqrt{1-\beta_t}y_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

其中，$t \sim [1, T]$ 和 β_1，β_2，$\cdots\cdots$，β_T 是一个使用 $\beta_t \in (0, 1)$ 的固定方差设置。使用符号 $\alpha_t = 1 - \beta_t$，可以得到任意时间步长 t 的对应的噪声图像 y_t。真正的后验条件 $q(y_{t-1}|y_t)$ 可以通过一个深度神经网络来近似，以预测 y_{t-1} 的平均值和方差，通过以下参数化:

$$p_\theta(y_{t-1}|y_t, x_p, x_s) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, t, x_p, x_s), \Sigma_\theta(y_t, t, x_p, x_s)) \tag{2}$$

在去噪扩散模块 $\mathcal{H}_N$ 中，将第 t 步的噪声 y_t 与目标姿态 x_p 拼接起来，并送到 $\mathcal{H}_N$ 中，x_p 将指导去噪过程，并确保中间噪声表示和最终图像遵循给定的骨架结构。这样便可以预测该步对应的噪声。

3.3 损失函数定义

为了训练去噪过程，作者首先通过将高斯噪声 ϵ 添加到 y_0 中生成有噪声样本，然后训练条件去噪模型使用标准 MSE 损失来预测添加的噪声:

$$\mathcal{L}_{\text{mse}} = E_{t\sim[1,T], y_0\sim q(y_0), \epsilon} \|\epsilon - \epsilon_\theta(y_t, t, x_p, x_s)\|^2. \tag{3}$$

Nichol 等人 [15] 提出了一种有效的学习策略，作为 DDPM 的改进版本，所需的步骤更少，并应用了一个额外的损失项 $\mathcal{L}_{\text{vib}}$ 来学习方差。作者将两者进行融合作为总体损失函数:

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{vib}}. \tag{4}$$

4 复现细节

4.1 实验环境搭建

本文的实验在服务器 RTX-3090-1 卡条件下进行，python 环境如下：

- Python == 3.7, pytorch-cuda=11.7

同时环境中安装以下 python 库：
imageio==2.10.3
lmdb==1.2.1
opencv-python==4.5.4.58
Pillow==8.3.2
PyYAML==5.4.1
scikit-image==0.17.2
scipy==1.5.4
tensorboard==2.6.0
tqdm==4.62.3
wandb
tensorfn

备注：原文使用 python3.6，但在该实验中发现其中用到的库 wandb 只支持 python3.7 以上，故本实验将其环境改为 python3.7 。

4.2 数据集与评价指标

本文使用服装检索基准数据集 DeepFashion [11] 和 Market-1501 [22] 数据集。DeepFashion 数据集包含 52,712 张时装模特的高分辨率图片，将该数据集分成训练和测试子集，分别为 101,966 对和 8,570 对。Market-1501 包含 32,668 张低分辨率图像，这些图像在不同的视点、背景、照明等方面都有所不同。对于这两个数据集，训练集和测试集的个人身份不重叠。姿态图（pose map）由 OpenPose [2] 提取。

在评价指标方面，使用三个不同的度量标准来评估该模型，分别为 FID（用于测量生成图像的真实性）、SSIM（计算像素级的图像的相似度）和 LPIPS（计算生成的图像与参考图像在感知域之间的距离）。

4.3 模型参数设置

在本实验中，对于 DeepFashion 数据集，使用 256×176 张和 512×352 张图像来训练模型。对于 Market-1501 数据集，使用 128×64 张图像。此外，batch-size 大小设置为 8，学习速率设置为 2e−5，噪声步长 T 设置为 1000。优化器选择 Adam 优化器。对于采样，w_p 和 w_s 设置为 2.0。

4.4 与已有开源代码对比

本文在已有的开源代码基础上，主要进行以下几个方面的工作：

（1）在图像数据预处理方面，加入将训练数据集进行降维操作的模块。

```
class DimensionalityReduction:
    def __init__(self, n_components):
        self.n_components = n_components
        self.pca = None
        self.scaler = None

    def fit_transform(self, X_train):
        self.scaler = StandardScaler()
        X_train_scaled = self.scaler.fit_transform(X_train)
        self.pca = PCA(n_components=self.n_components)
        X_train_reduced = self.pca.fit_transform(X_train_scaled)

        return X_train_reduced

    def transform(self, X_test):
        X_test_scaled = self.scaler.transform(X_test)
        X_test_reduced = self.pca.transform(X_test_scaled)

        return X_test_reduced
```

（2）在学习率调整方面，使用了学习率衰减策略（余弦退火法）。

```
class CosineAnnealingLR:
    def __init__(self, optimizer, initial_lr, T_max, eta_min=0):
        self.optimizer = optimizer
        self.initial_lr = initial_lr
        self.T_max = T_max
        self.eta_min = eta_min
        self.last_epoch = -1
        self.lr = initial_lr

    def step(self, epoch):
        if epoch != self.last_epoch:
            self.last_epoch = epoch
            self.lr = self.eta_min + (self.initial_lr - self.eta_min)
            * \(1 + math.cos(math.pi * epoch / self.T_max)) / 2
            for param_group in self.optimizer.param_groups:
                param_group['lr'] = self.lr
```

此改进效果：原模型训练速度在 RTX-3090-1 卡条件下，训练一次迭代大约三个小时，调整后训练速度在 RTX-3090-1 卡条件下，训练一次迭代约提高 15 分钟。如图 6所示，为模型训练过程。

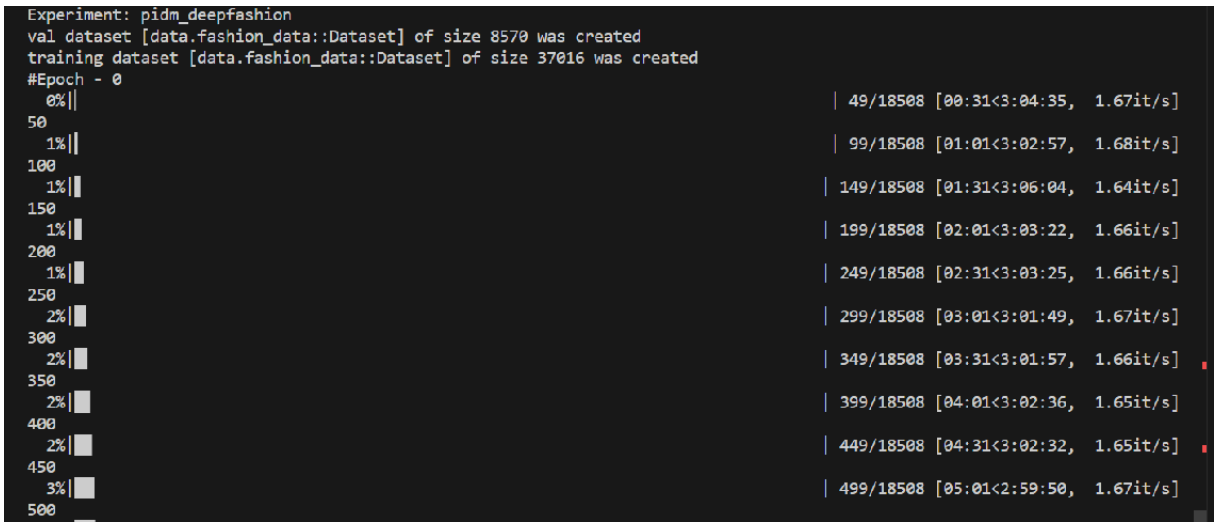


图 6. 训练过程展示

4.5 创新点

本文的核心创新思想有两个：

（1）针对扩散模型的训练速度：在数据预处理方面，加入了将训练数据集进行降维操作的模块；在学习率调整方面，使用了学习率衰减策略（余弦退火法）；在模型改造方面，将基于 Rectified Flow [9] 改造扩散模型，其思想为基于一个简单的常微分方程（ODE），通过构造一个“尽量走直线”的连续运动系统来产生想要的数据分布。

（2）针对图像背景干扰问题：本文通过对 Diffusion Model 扩散过程的观察发现，它的算法实现中，都是对整张图像进行像素级扩散，进而容易将背景与目标对象融合生成。因此，为消除背景对生成任务的干扰，本文考虑先将图像进行局部分割，使得生成模型将只关注目标物体的区域，从而实现图像局部扩散。

5 实验结果分析

5.1 实验结果的定量分析

在本文的结果展示中，利用原论文提供的预训练模型，首先是将源代码的测试图像输入到模型中，从而合成了姿态引导的人物图像，如图 7和图 8所示。实验（1）的测试图像来自数据集 DeepFashion 256×176，实验（2）的测试图像来自数据集 DeepFashion 512×352，其中四个姿态图 pose map 为每次生成过程中从 pose 数据集随机抽取，因此可能会出现相同姿态引导生成的情况。通过观察实验（1）和实验（2）的结果，可以看到生成的图像质量都很好，并且其风格和姿态与源图像的外观和目标姿态紧密匹配。

然而，为验证该模型的泛化能力，本文从 DeepFashion 数据集里随机抽取了一些图像进行多次测试，发现其生成结果出现了一些问题。

图 7. 实验（1）

图 8. 实验（2）

首先，在实验（3）和实验（4）中（见图 9和图 10），对输入的测试图像出现了生成图像中男女混乱的现象，违背了“保持源图像人物身份来生成”的意愿。这说明此方法模型还不够稳定，或者训练时出现过拟合现象。针对这些问题，在训练模型时可以引入身份特征嵌入来帮助模型更好地保持源图像的人物身份。这可以通过在网络架构中添加一个身份编码器，并将其与生成器进行联合训练来实现。或者对模型进行调优和正则化。这包括对网络架构进行改进，使用合适的正则化方法（如 dropout、权重衰减等），以及调整超参数和训练策略。这样可以提高模型的稳定性，减少过拟合现象。

图 9. 实验（3）

图 10. 实验（4）

其次，在实验（4）和实验（5）中（见图 10和图 11），其生成结果的服装纹理甚至形状都发生了扭曲，但姿态的生成基本匹配。通过观察输入图像的特点，发现这两组实验的输入图像均为长裙。再观察发现，原论文给出的结果示例中，确实没有输入长裙的情况，基本都是裤子或者短裙。基于此，本文认为，在实现对齐姿势时，强度过大，导致对长裙生成造成的干扰（长裙会对腿部遮挡，影响姿态引导）。针对这个问题，在生成过程中，可以调整姿态引导的强度，使其在对齐姿势时不会过于强烈。或者在训练和生成过程中，使用服装分割掩模来指导模型更好地生成服装的纹理和形状。

图 11. 实验（5）

最后，是一个最主要的问题，在实验（5）和实验（6）中（见图 11和图 12），其生成的结果显而易见更为杂乱，这是由于输入的图像存在背景干扰。原论文给出的结果示例中，输入的图像背景均极为干净，故不会出现这样的情况。本文对 Diffusion Model 扩散过程的观察发现，它的算法实现中，都是对整张图像进行像素级扩散，进而容易将背景与目标对象融合生成。因此，为消除背景对生成任务的干扰，考虑是否可以先将图像进行局部分割，使得生成模型将只关注目标物体的区域。

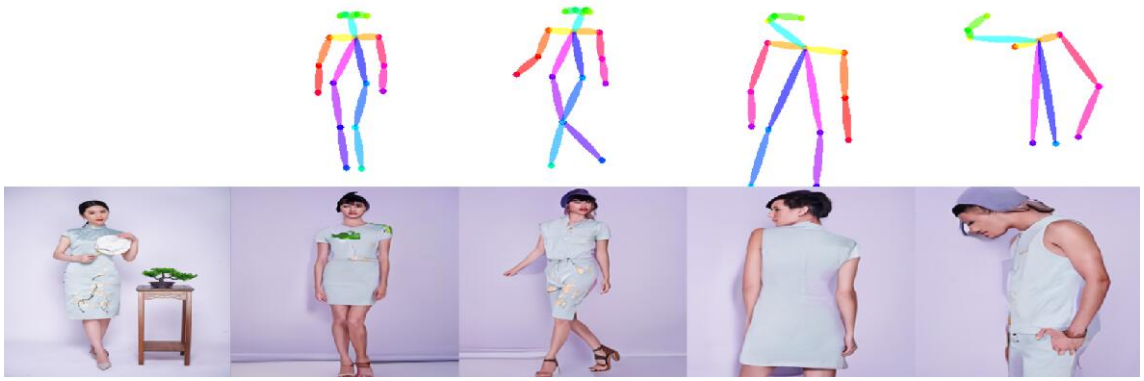


图 12. 实验（6）

5.2 实验结果的定性分析

表 1展示了以上在对应数据集（测试图像）实验的三个评价指标上的具体表现。在表中，黑色加粗字体表示原论文给出的实验效果指标。绿色加粗字体表示实验（1）和实验（2）表现效果良好的情况。红色加粗字体表示以上提到实验效果出现偏差的情况。从实验的评价指标结果可见，其评价结果与 5.1 节中可视化效果的趋势接近，当生成效果出现较大偏差时，其对应的评价指标也会有较大差异。

总的来说，该方法模型的生成效果在很大程度上依赖于输入图像。其中，在姿态引导的生成效果方面，基本不受输入图像的影响。然而，姿态的引导生成过程会影响图像服装质量的生成。而在图像人物以及服装的生成方面，受输入图像的影响较大，不同的输入，其生成的效果也可能会差异较大。这表明了输入图像的选择对于生成结果的质量和一致性非常重要。进一步的研究可以探索如何提高模型对于不同输入图像的鲁棒性和生成效果的一致性。

表 1. 实验的评价指标对比

Dataset	实验（x）	FID(↓)	SSIM(↑)	LPIPS(↓)
DeepFashion (256 x 176)	**PIDM**	**6.3671**	**0.7312**	**0.1678**
	(1)	**6.3671**	**0.7312**	**0.1678**
	(3)	**21.367**	**0.5638**	**0.2585**
	(4)	**23.484**	**0.5039**	**0.2617**
DeepFashion (512 x 352)	**PIDM**	**5.8365**	**0.7419**	**0.1768**
	(2)	**5.8412**	**0.7401**	**0.1781**
	(5)	**22.340**	**0.4988**	**0.2623**
其他	(6)	**21.831**	**0.5571**	**0.2573**

6 总结与展望

在生成任务领域，[1] 为我们开启了新的视野。在近年来的计算机视觉研究中，对于新视角图像合成的探索一直在持续进行。同时，许多相关工作利用姿势图作为引导来生成新的视角图像，其中大部分方法使用了生成对抗网络（GAN）。然而，这些方法在处理复杂姿势和严重遮挡时存在困难。值得注意的是，[1] 首次将扩散模型应用于人物图像生成任务，这种方法能够有效地学习数据分布，并显著提高生成样本的多样性，这是扩散模型的固有特性。

而本文在复现 [1] 的研究成果时发现，扩散模型直接应用到生成任务中，仍存在许多问题。具体来说，以上实验结果揭示了其不稳定性（生成男女混乱现象）和非自适应性（姿态对长裙生成的干扰）等方面存在的挑战。此外，还有两个重要的问题需要解决：模型训练速度（模型复杂度）和背景干扰。针对扩散模型的训练速度问题，需要对模型进行轻量化改造。一种研究方向是基于 Rectified Flow [9] 对扩散模型进行改进。该方法基于简单的常微分方程（ODE），通过构建一个“尽量走直线”的连续运动系统，以实现所需的数据分布。通过这种改进，可以加快模型的训练速度。另外，针对图像背景干扰问题，考虑先对图像进行局部分割，使生成模型只关注目标物体的区域，从而实现图像局部扩散。这样的处理方式可以减少背景干扰对生成结果的影响，提高生成图像的质量和真实性。

综合而言，本文的复现工作展示了对 [1] 的研究成果，并在此基础上提出一些优化方法。同时，发现了扩散模型在生成任务中的问题。为了解决这些问题，提出了改进扩散模型的训练速度（模型复杂度）和处理图像背景干扰的方法思想。这些改进方法思想有望推动扩散模型在生成任务领域的研究进展，并提高生成模型的性能和适用性。

参考文献

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[4] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8857–8866, 2018.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[8] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019.

[9] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[10] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[11] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[12] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10806–10815, 2021.

[13] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

[14] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020.

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

[16] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022.

[17] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020.

[18] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018.

[19] 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.

[20] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021.

[21] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[22] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[23] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021.

[24] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision*, pages 161–178. Springer, 2022.

[25] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.