3DMM 驱动的人脸旋转可控生成

冼晓乐

20240103

摘要

尽管近年来人脸旋转方面的工作已经取得不小的进展，但是高质量的人脸配对数据仍然是钳制该课题进一步发展，并且落地部署的一大原因。在现有条件下需要一种不依赖于配对人脸数据集的方法，基于无监督的生成模型生成对应正脸，为此我们提出了 3DMM 驱动下的人脸扩散生成正脸，基于该框架我们能够利用大量现有的人脸数据集作为训练数据，解决数据集瓶颈的问题。具体而言，在训练阶段我们首先使用了现有基于卷积网络 3DMM 方法拟合任意一张正脸图片的 3DMM 系数，通过修改其中的欧拉角系数在 3D 空间修改人脸的角度，得到正侧脸数据对。侧脸作为可控扩散模型的控制条件，通过学习对应的人脸先验信息，我们的模型能够生成其对应的正脸图片。即通过 3DMM 方法利用单张人脸图片生成可供自监督训练的数据对。对比现有的捕获特定概念的生成模型，在野外图像的人脸旋转上，我们的模型能生成更为真实，身份保真度更高的人脸图片。

关键词： 3DMM；可控扩散模型；自监督

1 引言

人脸转正一直是一个备受关注的研究课题，它在现实场景中有广泛的应用场景，不仅能够用于公安系统的刑事侦查，还是“美图秀秀”，“妙鸭相机”等新兴多媒体 APP 的特色功能之一。

传统的方法可以通过使用 3D 数据，对人脸进行 3D 建模来解决，进而旋转该 3D 人脸得到结果，但由于此类方法严重依赖于 3DMM 使用的数据库，难以对各个环境下的人脸进行拟合，并且需要依赖额外的方法来恢复人脸的高频细节。基于 3DMM 的方法拟合的人脸尽管是粗糙的，但是其强大的人脸拟合能力仍然有强大的潜在利用价值，例如对于一张极端姿态下的人脸....。所以我们的另一个目的为 1) 正确地利用 3DMM。# 待补充

而在深度学习领域，不少作者都针对该问题提出自己的解决方法，有的把 3DMM 与深度学习结合，使之能够产生多视角的人脸。基于 GAN 的生成技术再图像生成领域蓬勃发展，与此同时也催生了使用 GAN 相关技术来实现该目的的工作。但是基于 GAN 的方法需要正侧对的人脸数据集提供有监督训练的数据对，然而现实中极其缺少这样的数据集，或是该数据集具有单一域信息，采集条件固定，风格单一，像来自于卡耐基梅隆大学的 MultiPie 正侧对数据集是在单一环境下采集的，即使神经网络能够在该数据集的测试集上表现良好，但对于野外采集的数据集，其效果令人堪忧，这样的效果显然是难以做到落地应用，失去了该任务其本身的意义。

尽管像上述所说的基于 GAN 的方法存在其无法解决的固有问题，但其通过使用基于深度学习的图像生成来实现人脸转正的思想是可取的，这是因为深度学习的模型能够在训练阶段从大量的人脸中学习到丰富的先验知识，再基于这样强大的先验知识去生成人脸能够利用上模型的“经验之谈”。这个过程就像是人类在见识了大量人脸之后根据一张非正常姿态下的人脸图片想象该人脸的正面肖像。那么问题就转化为 2) 如何合理地利用大量存在的人脸作为模型的训练集，而不仅限于使用有限的正侧人脸数据对。

根据侧脸得到对应的正脸，在生成领域我们很容易想到这是个图像作为条件的有条件生成任务。最近，在图像生成领域，diffusion 扩散模型异军突起，已经成为生成领域的热门技术，而且在数亿计的图像上预训练 Stable Diffusion 模型的生成能力已经能够落地应用，具备商用价值。像 Textual Inversion，DiffusinRig 等工作思考了如何让 Diffusion 捕获特定的概念的物体或是特定的人脸，实现定制化内容的产生。众多的工作都表明其强大的生成能力仍然可以不断挖掘。那么 3) 如何挖掘强大的 diffusion 预训练模型为我们的任务服务是一个值得思考的角度。

结合以上问题，我们提出了 FrontDiff 模型，这个模型继承了 diffusion 模型自监督训练的优良训练方法，并且利用 3DMM 拟合的人脸指导 diffusion 的生成，达到我们人脸转正的目的。我们的模型是首先通过 Deep3Dface 来建模人脸，通过编辑 3DMM 的系数编辑人脸的角度从而得到正侧脸数据对。具体来说，一张极端姿态下的人脸通过 3DMM 方法的拟合，再通过 3DMM 系数的修改可以转为正脸，那么存在一部分在 2D 空间是被遮挡的。类似地，我们能够使用 3DMM 拟合的正脸进行旋转得到侧脸，再转回正脸，从 RGB 正脸——3DMM 侧脸——3DMM 正脸的过程中，我们成功得到了基于 3DMM 方法拟合的正侧脸数据对。接下来的任务就是如何使用这些构造的数据对训练我们的生成模型能够根据 3DMM 侧脸生成真实的正脸。我的方法是锁定 diffusion 原有的去噪网络 U-Net 来保留其作为大模型的生成能力，并且训练一个模块能够感知侧脸，控制去噪网络的前向过程来生成对应的正脸。在训练充分的条件下，我们的模型能够根据 3DMM 建模并渲染出的任意一张侧脸作为控制条件，生成对应的正脸。

2 相关工作

三维人脸重建一直是计算机视觉和计算机图形学领域的重要任务之一。三维可变形人脸（3DMM）在三维人脸建模中发挥了至关重要的作用。Basel Face Model [12] 建立了一个统计模型，能够表示人脸的形状和纹理变化，被广泛应用于人脸识别、人脸对齐、表情合成等领域，不少三维人脸重建的工作都是基于该统计模型展开 [2], [11], [26]。随着深度学习技术的发展，神经网络的技术被用来作为估计 3DMM 系数的回归器。[24] 提出了“渐进形状回归”的方法，该方法逐步地从粗糙到细致地预测 3D 形状参数，通过这种渐进方式，网络可以逐步提取更加准确的 3D 形状信息，[6] 则是通过弱监督的方式尝试性地解决数据集较少的问题。

3DMM 估计人脸的技术也逐渐成熟，这类方法已经逐渐出现在其他人脸的视觉任务作为先验信息辅助完成其他任务了。[30] 的工作中基于 3DMM 系数的旋转估计和渲染重建，单视角图像实现逼真的人脸旋转渲染。[7] 使用了 DECA [8] 估计的人脸进一步生成物理渲染作为人脸生成的先验知识。[20] 通过与 3DMM 对齐的表示实现头部的全局和局部编辑。越来越多的工作都在基于 3DMM 作为先验信息展开，也在各自的人脸任务上取得了不错的效果。

2.1 图像可控生成扩散模型

文本引导的控制方法侧重于调整提示、操作 CLIP 特征和修改交叉注意力 [1] [4] [9] [21] [15]，难以得到准确的定制化效果，所以出于方便个性化、定制化或任务特定的图像生成，不少工作在 LDM [25] 的基础上加入不同种类的控制条件。例如，图像扩散过程直接对颜色变化 [22]，Textual Inversion [10] 通过找到用户给定实例图像对应的词向量来捕获特定概念的生成，IDiff [3] 使用人脸识别网络捕获人脸特征作为嵌入向量指导人脸用于生成具有真实身份变化的合成人脸，Paint-by-example [27] 提出了无监督框架提取物体的隐空间特征作为交叉注意力的输入实现局部重绘。这些工作的内核都是利用交叉注意力的机制来控制图像生成，用图像特征替代了替代了文本的输入，依然是基于调整扩散提示来实现可控生成，难以做到精细化的要求。Composer [16] 解耦了控制图像的各个条件的概念，并且使用各个条件的组合实现可控生成。T2I [23] 微调了感知控制图像的适配器，并注入到扩散模型上采样的过程控制扩散过程。[29] 通过微调一个原有网络的副本分支来感知额外的图像条件，实现在分割图，深度图，线稿图等多种条件的可控生成，与此同时还设计了零卷积来防止模型受到有害噪声的影响。像这样直接把图像提取的特征注入到扩散模型的网络（U-Net）虽然抛弃了 Attention 语义信息的处理机制，但却能够在精细化定制生成内容得到更好的效果。

3 本文方法

接下来会从多个方面讲述该方法的主体框架，包括如何使用 3DMM 的方法来构造正侧脸数据对实现自监督训练，怎么在锁定原有 diffusion 参数的情况下训练一个模块能够合理地感知 3DMM 建模并渲染得到的侧脸，并且根据感知到的信息参与到原有 diffusion 去噪的前向过程。

3.1 基于 3DMM 的自监督数据对构造

通过使用 3DMM，任意一张人脸均可表示为：

$$\begin{aligned} \mathbf{S} &= \mathbf{S}(\alpha, \beta) = \bar{\mathbf{S}} + \mathbf{B}_{id}\alpha + \mathbf{B}_{exp}\beta \\ \mathbf{T} &= \mathbf{T}(\delta) = \bar{\mathbf{T}} + \mathbf{B}_{t}\delta \end{aligned} \tag{1}$$

其中 $\bar{\mathbf{S}}$ 为 2009 BFM [12] 数据库中的均值人脸形状，$\bar{\mathbf{T}}$ 为 FaceWarehouse 数据库中的均值纹理。$\mathbf{B}_{id}$，$\mathbf{B}_{exp}$，$\mathbf{B}_{t}$ 分别代表该 PCA 后的 80 个表情基，64 个表情基，80 个纹理基。α, β, δ 则分别为这些基的系数向量。

对于一张给定的 RGB 人脸图像，我们通过 CNN 网络来回归得到这些人脸的系数向量。值得注意的是，这些系数中还会包含人脸在 3D 空间中的欧拉角系数 $\mathbf{R}$。我们可以根据 α, β, δ,$\mathbf{R}$ 进一步渲染得到脸部 $\mathbf{F}$。

$$\mathbf{F} = Render(S, T, \mathbf{R}) \tag{2}$$

所以任意侧脸的 3DMM 系数，我们能够根据 3DMM 系数渲染得到侧面的 2D 脸部 $\mathbf{F}$，但其欧拉角系数 $\mathbf{R}$ 可以人为地修改，修改过后的系数再渲染就可以得到我们对应角度对应的 2D 脸部。

现在我们基于现有的 3DMM 拟合提出假设，对于估计人脸 3DMM 系数的 CNN 网络，如果是一张侧脸的真实场景下的 RGB 人脸图像作为输入，缺失被遮挡的部分脸部信息，那么 3DMM 在这一部分的拟合是有偏差的。基于这样的假设，通过上述所说的通过修改其欧拉角并渲染这样的方式得到的其他角度的 2D 脸部，存在拟合有所偏差的部分，那么我们认为在 3D 渲染成 2D 脸部时被遮挡的部分即是拟合偏差的部分，这部分的人脸，我们做掩码处理。特别的，如果这样的脸部是正面状态的，其欧拉角系数 $\mathbf{F}_{front}$。

$$\begin{aligned} T_{oc} &= T \cdot mask \\ \mathbf{F}_{front}^{mask} &= Render(S, T_{oc}, R_{front}) \end{aligned} \tag{3}$$

那么最后的产生的结果，我们认为这样掩码处理后的脸部代表了一个人的侧脸，同时我们也可以认为这样的脸部是有待补全的正面脸部 $\mathbf{F}_{front}^{mask}$。

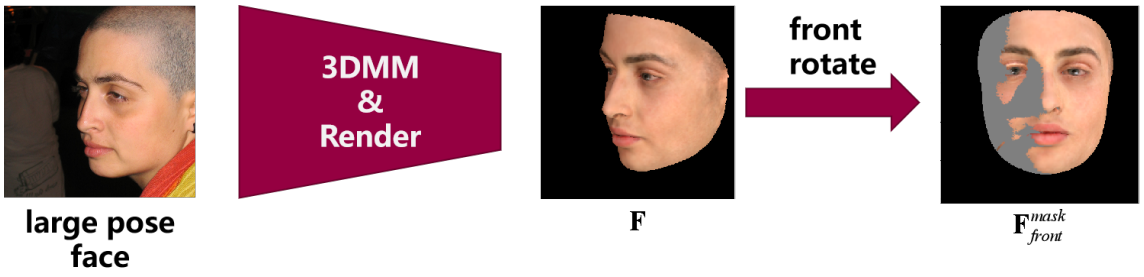

图 1. 3DMM 中任意的侧脸构造掩码处理后的正脸

基于对欧拉角人为编辑的思路，我们能够从任意一张正脸的 RGB 图像 I_{front} 出发，去构造其任意角度下的一张侧面脸部 $\mathbf{F}_{front}^{mask}$，其流程如图所示。具体来说，我们使用 CNN 网络来拟合的 3DMM 系数，与上述侧面-正面步骤相似，我们随机化其欧拉角系数得到视为 R_{rand} 旋转其脸部角度，那么通过渲染，我们可以得到：

$$\mathbf{F}_{rand} = Render(S, T, R_{rand}) \tag{4}$$

接下来我们重复前文所说的掩码部分的操作，就能从一张正面的 RGB 人脸照得到 $\mathbf{F}_{front}^{mask}$。

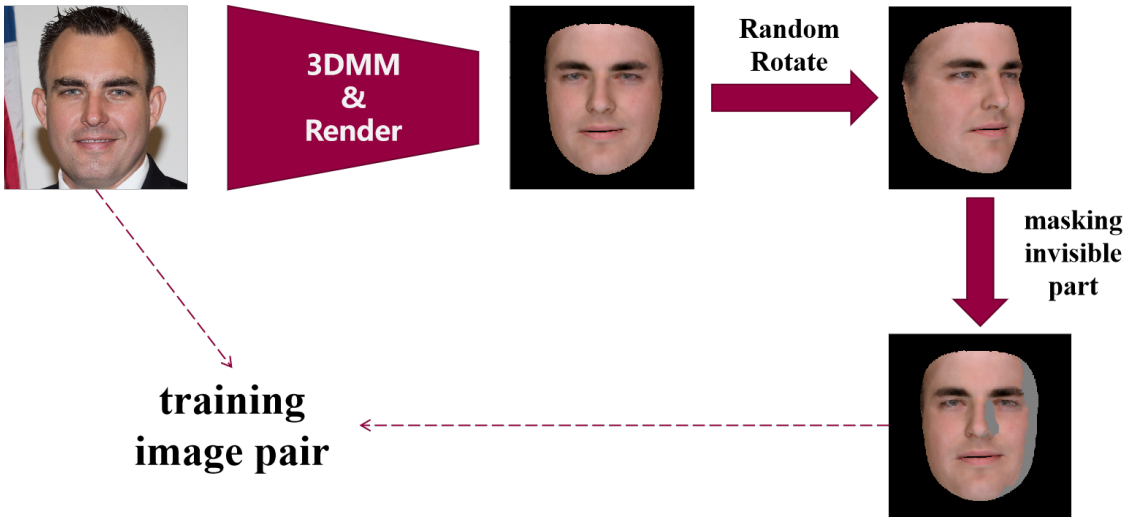

图 2. 正脸脸数据对的构造

正侧人脸数据对，尽管这样的数据对不全是现实场景 RGB 的，但是其仍然有它的价值所在，具体如何使用，在后面生成任务的框架上会详细说明。对于 FFHQ，CelebA 等高清人脸数据集，其包含了大量的正面化人脸图像，所以理论上我们能够使用众多高清的人脸数据集构造我们的训练数据集，我们的模型在训练数据上不再受拘束于有监督的框架下的缺点，包括但不限于训练数据有限导致的泛化能力差，在单一数据集上过拟合，而在跨域数据集，或者是野外人脸上表现差。

3.2 3DMM 面部驱动的正面化生成

在本文中，我们基于大规模文本到图像的潜在扩散模型——稳定扩散 (Stable Diffusion) 的预训练模型，并且进一步微调以适配我们的任务。Stable Diffusion 通过一个充分训练的 U-Net 网络估计数据分布的去噪序列来生成数据样本。为了能够有更高的效率和更为稳定的训练过程，Stable Diffusion 采用了一个预训练的自动编码器 Auto Encoder[29]，ε 将图像 x 转换成潜在的 z，并用解码器 $\mathcal{D}$ 解码，从隐空间映射到 RGB 色彩空间。这个潜在的表示是通过使用 VAE，Patch-GAN 和 LPIPS 的混合目标来训练得到的。扩散和去噪过程在隐空间中进行，在扩散过程中，在 t 时刻方差为 $\beta_t \in (0,1)$ 的高斯噪声被添加到编码隐变量 $z = \varepsilon(x)$ 中，们可以推导出该时间步下的含噪隐空间特征为：

$$z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{5}$$

其中 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$，$\alpha_t = 1 - \beta_t$，$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$。当 t 足够大时（一般我们在模型里面设为 1000），我们认为近似为一个高斯噪声。一个网络是通过预测随机选取的时间步长 t 在 c (在 stable diffusion 中为文本) 条件下的噪声 ϵ 来学习的。隐扩散模型的优化的损失函数定义如下：

$$\mathcal{L}_{ldm} = \mathbb{E}_{z,c,t,\epsilon}[\|\epsilon - \epsilon_\theta(z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t)\|_2^2] \tag{6}$$

其中 x，c 是从数据集中采样的，$\mathbf{F}_{front}^{mask}$，$z = \mathcal{E}(x)$，$t$ 是从均匀分布中采样的，ϵ 是从标准高斯分布采样。在第一阶段中，我们通过人脸重建技术 3DMM，并对编辑系数中控制欧拉角的部分，渲染得到 $\mathbf{F}_{front}^{mask}$，这样具有掩码的脸部仍然不是真实场景下的图片。我们接下来的任务是，$\mathbf{F}_{front}^{mask}$ 以作为条件 c，生成真实场景下的人脸图片。为此，我们的策略是在已经预训练的 Stable Diffusion 的基础上，加入额外模块感知条件 c，使之能够作为控制条件调控 Diffusion 去噪网络 U-Net 的去噪过程。具体流程如图所示：

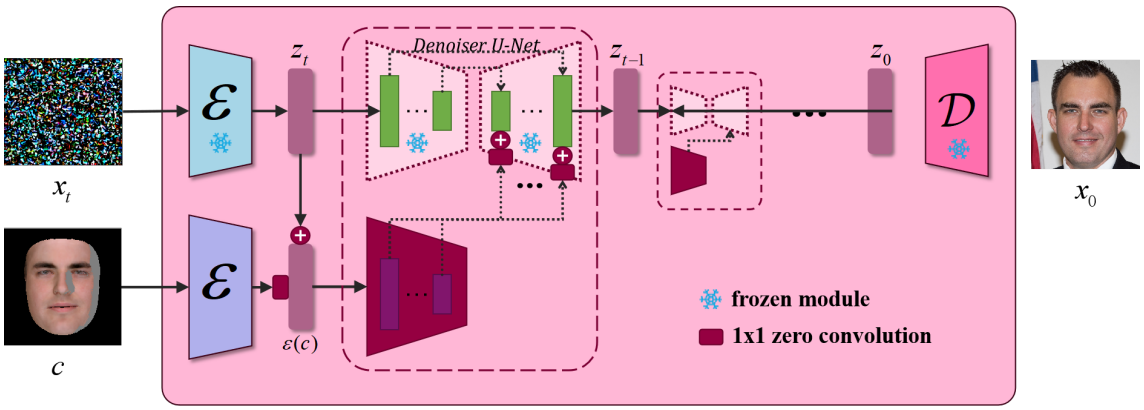

图 3. $c = \mathbf{F}_{front}^{mask}$ 控制下的 Diffusion 去噪网络

原有的 U-Net 网络包括一个下采样模块，中间模块，一个上采样模块，上采样模块在不同尺度的特征中融合下采样对应尺度的特征。我们设计了一个结构类似于原有 U-Net 网络的并行模块来感知控制条件，首先编码器将控制条件 C 映射到隐空间，得到隐空间的控制条件 $\mathcal{E}(c)$。再将 $\mathcal{E}(c)$ 与尺度相匹配的 z_t 相加输入到继承自原有 U-Net 网络下采样的并行模块中，并行模块各尺度的结果进一步通过相加的方式与原有 U-Net 网络上采样对应尺度的特征融合。需要注意的是，我们会在各层特征融合之前，加入一个 1x1 卷积核的零卷积层，该卷积层能够在初始的训练中保护并行模块直接输入较大有害噪声从而使得训练更加困难。由于我们任务不需要任何文本条件，而 Stable Diffusion 中会需要文本条件的输入，考虑到用需要使用原有预训练模型，尽可能不修改原有的模块，所以我们把文本条件设为空，即输入的文本固定为空文本。综上所述，我们扩散模型的优化目标转化为如下：

$$\mathcal{L}_{FDiff} = \mathbb{E}_{z_t,c,\epsilon,\mathcal{E}(c)}[||\epsilon - \epsilon_\theta(z_t, c, \mathcal{E}(c))||_2^2] \tag{7}$$

总的来说，只有 U-Net 去噪器中的跳跃连接特征针对我们的特定任务进行了调整。我们训练的目标得到一个编码控制条件的编码器和一个能够把编码后的条件融合到原有去噪 U-Net 网络的并行模块，通过微调的方式允许我们在原有 Diffusion 的基础上，通过额外的控制条件进一步挖掘其生成能力。

4 复现细节

本文所提出的方法涉及 2 个模型的源代码，即包括对一个 3DMM 人脸重建工作[1]的复现，以及在另外一个关于可控图像生成 diffusion[2]的改进。由于过去实现人脸转正的工作比较老旧，如 2017 年 [17]，2020 年 [30] 等等，这些方法要么开源代码上的文件已经失效，要么依赖于卡耐基梅隆大学的具有正侧人脸数据对标注的数据集 Multi-PIE [13]，而这个数据集并不公开且需要花费 5000 美元购买。所以本文的对比实验无法使用现有的成果，只能选取现有方法具备这个功能的相关图像生成方面的工作。因此，两个对比方法的代码[3][4]也需要复现，并在我们数据集上进行测试。

[1] https://github.com/sicxu/Deep3DFaceRecon_pytorch

[2] https://github.com/lllyasviel/ControlNet

[3] https://github.com/adobe-research/diffusion-rig

[4] https://github.com/huggingface/diffusers/tree/main/examples/textual_inversion

4.1 与已有开源代码对比

由于对比方法的代码只是简单依据 Github 上的源代码进行复现而并没有改进之处，所以接下来主要介绍本文所提出方法涉及的源代码相关的对比。

构建正侧脸的数据对需要用到 3DMM 人脸重建工作。如图4所示，源代码虽然能够将重建的人脸渲染到 2D 平面，但是无法实现来回旋转过程中，遮挡部分进行掩码，需要进行相关改进实现该功能。

(a) (b) (c) (d)

图 4. (a) 为原 RGB 图像，(b) 为源代码所依据人脸图像所重建得到的 2D 渲染图，(c) 为旋转一定角度后所渲染得到的 2D 图，(d) 为掩码 (c) 中面部遮挡区域后再取消所旋转角度后所得到的 2D 渲染图

所以我的工作所增加的代码分为 **1) 修改重建系数关于人脸角度的部分，使其的渲染结果为随机角度的侧脸。2) 获取该侧脸 3D mesh 数据结构中，会因遮挡而不产生着色的顶点，修改其所在三角片面的着色**。由于 2D 平面所着色的像素点于 3D 平面中的点并不一一对应，所以在 **2)** 这部分的实现中存在难度。具体操作为，在渲染4(c) 前，对其 3D mesh 数据执行光栅化，根据管线渲染所使用到的三角片面坐标，找到对应的顶点，则另外的顶点即为被遮挡的顶点，将这些遮挡顶点的 RGB 颜色值赋为特定值，即可完成遮挡点的掩码操作。接下来执行 **1)** 中修改人脸角度系数的逆操作，最后渲染得到的人脸即为具有遮挡部分掩码的 2D 渲染图。主体部分的代码如图所示5

```
# 获取光栅化操作用于着色的三角片面
visible_tri_index = rast_out[0][...,-1].reshape(-1).long()
visible_tri_index = visible_tri_index[visible_tri_index != 0]-1
visible_vertex_index = tri[visible_tri_index,:]
visible_vertex_index = visible_vertex_index.reshape(-1).unique()


# 剔除不可见的三角点
bool_index = torch.ones(vertex.shape[1],dtype=torch.int32)
bool_index[visible_vertex_index.long()] = False
invisible_vertex_index = torch.nonzero(bool_index == 1).reshape(-1)

return mask, depth, image, invisible_vertex_index
```

图 5. 构建正侧脸数据对的主体部分代码

接下来根据4以 (a) 作为原图，(d) 作为 condition 控制条件的数据集代码，替代原有 ControlNet 源代码 2 中的数据集代码后既可以进行扩散模型相关的训练。

4.2 实验环境搭建

因为只需要一张正脸，就能通过我们的方法构造出训练人脸数据对，所以理论上我们拥有无限大的数据集作为我们的训练集，但是考虑到许多人脸数据集存在域单一的问题，并且受到训练成本与实验室条件的限制，我们选定 FFHQ [18]，包含 70000 张高清人脸的野外人脸数据集作为我们的训练集，并且缩放到 512×512 的大小以适配我们的预训练模型 Stable Diffusion 的输入。在这样的数据集上训练，一方面有效规避了上述问题，另一方面也能说明我们的模型更具有落地部署的价值。

用于拟合单张图片的 3DMM 人脸系数的模型，我们采用的是被广泛认可的 [6]，在该模型的估计种，输入的图片会再次下采样至 224×224 的大小，生成的 3DMM 系数包括面部形状，表情，纹理，光照球谐函数，欧拉角，相机位移。对于可控生成的网络，我们使用的是 Stable Diffusion 2.1-base 作为我们的预训练模型，并在此基础上以批次大小为 4 微调 20 轮次，设置的迭代器为 Adam [19]，学习率设置为 $\mathbf{10}^{-5}$。

本文所有的训练以及测试均在视觉所的搭配 NVIDIA GeForce RTX 3090 的服务器进行，在该服务器上搭建 pytorch lightening 环境。

4.3 创新点

本文的创新点集中在无监督的框架上，以往的人脸旋转工作依赖于配对的人脸数据对，相关的限制在本文的引言部分已经有所介绍。而本文能够依据 3DMM 可以修改人脸角度的这一角度出发，找到来回旋转中产生遮挡的区域并进行掩码从而构建出正侧人脸数据对，实现人脸旋转的自监督训练。

5 实验结果分析

5.1 现有生成模型的转正效果对比

我的结果展示在图8中，所使用的图片为随机选取的 3 张野外人脸图片，并且使用 MTCNN [28] 进行人脸对齐后剪裁到 512×512 的大小作为输入。由于我们的方法是基于无监督框架的，所以我们更倾向于与现有具备捕获特定概念的微调方法作比较，他们的方法需要对在给定图像上微调，对此，我选定单张侧脸作为微调的图片集。

可视化结果正如图8所示，对于野外图像中的人脸，现有基于 diffusion 的模型都能以精准捕获到特定概念，Textual Inversion 难以保持个体的身份信息，而 DiffusionRig 虽然在保持个体信息与外观条件略优于我们的方法，但是可以明显看到，在有其他角色出现在照片中的情况下，容易出现图片崩溃的现象。DiffusionRig 与我们的方法类似，都是基于 3DMM 先验条件的生成，进而说明加入 3D 条件作为可控生成的条件对于人脸可以更多地保持人脸的信息。

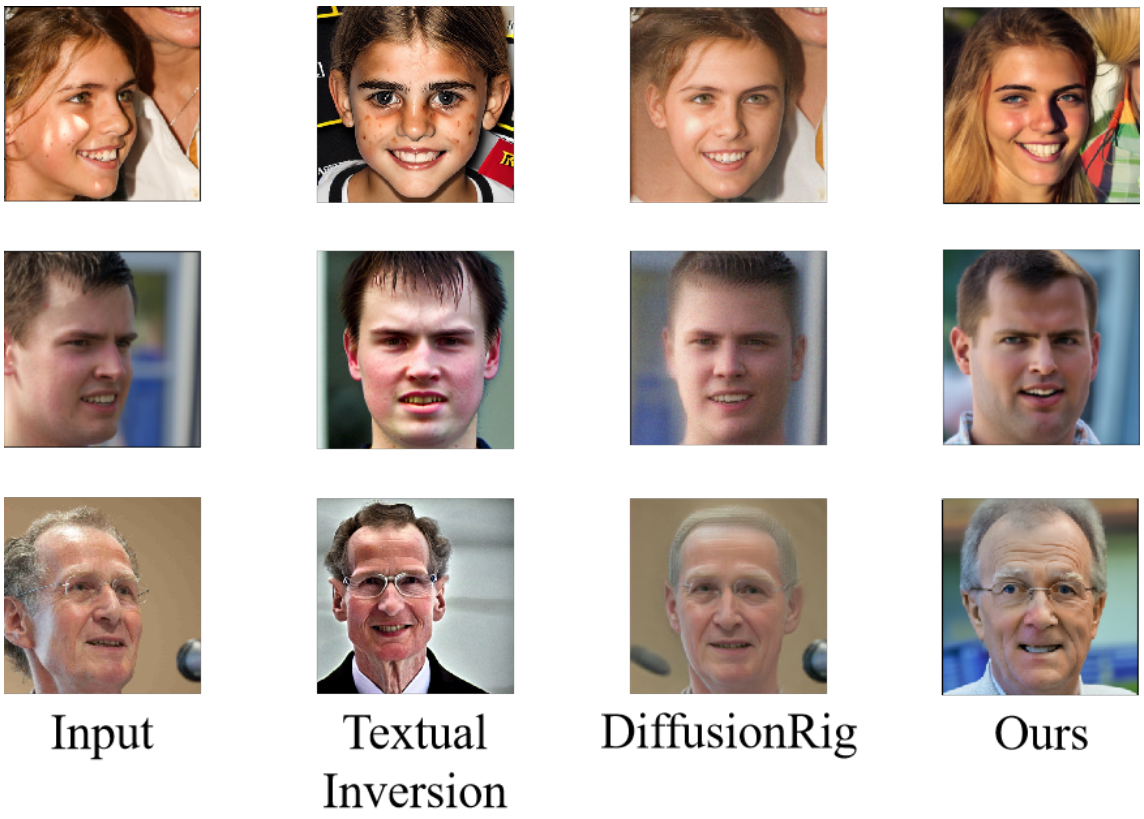

图 6. 与现有捕获特定身份概念的模型效果对比图

5.2 ID 身份保持效果

对于人脸旋转的任务，保持人脸身份是一个非常重要的前提，如果无法保持身份信息，那么生成的图片已经是背离了任务本身，因此身份信息保持度也是衡量这个任务的一个重要指标。为了衡量该任务的身份信息保持度，我们需要一个能够衡量旋转前后人脸的身份信息保持度的度量器。对此，我们使用在 MS1MV2 [14] 上训练的 Arcface [5] 作为感知损失的提取器，旋转前后人脸图片提取出的特征做余弦相似度衡量作为我们的相似度结果。对于 Stable Diffusion 而言，由于需要只能提供文本作为提示性指导，所以我们选取部分 LFW 数据集中的名人及其名字作为 Stable Diffusion 指示性文本。对于另外两种对比的方法，由于其需要微调，对于特定身份的人物选取了其最多 10 张图片作为微调的输入（部分超出 10 张的则随机选取 10 张）。

表 1. 与现有方法身份保持度的对比

方法	身份相似度	（微调）图片数量
Stable Diffusion$_{\text{文本}}$	18.27%	0
Textual Inversion$_{\text{微调 + 文本}}$	51.93%	1~10
DiffusionRig$_{\text{微调 + 文本}}$	81.48%	1~10
本文方法	76.30%	0

在表1中可以看到，对比于现有的方法，我们的模型能够在免去微调这一耗时的步骤的前提下，依然能够保持较高的身份相似度。虽然相较于 DiffusionRig 身份相似度有所降低，其

中原因在于 DiffusionRig 中有全局特征编码器，在微调阶段能够识别图片的外观条件，这一定程度上也会影响 Arcface 对人脸相似度的判别。

6 总结与展望

本文的研究提出了一种基于 3DMM 驱动的人脸扩散生成模型 FrontDiff, 用以实现人脸从任意姿态转向正脸的能力。本文的方法利用 3DMM 的人脸重建技术从单张人脸图像自动构造大量的正侧脸训练数据对, 从而解决了基于 GAN 生成模型需要大量有监督图像对数据的问题。模型在保留原扩散模型优异生成能力的同时, 加入额外模块用以感知 3DMM 建模的人脸条件, 利用条件信息参与和调控扩散去噪过程, 从而实现人脸姿态的可控生成。实验结果表明, 我们的模型能够在无需任何微调的情况下, 只基于 3DMM 条件实现人脸旋转, 并且相对于其他基于 diffusion 微调的方法保持了更高的人脸身份一致性。该方法成功地将 3DMM 技术应用在扩散技术中, 有效结合了三维空间信息和现有 2D 扩散生成模型的优点, 实现了人脸旋转生成任务的自监督学习。模型还存在可以进一步优化的空间, 比如 3DMM 重建结果的准确性直接影响生成质量, 未来可以考虑 joint training 3DMM 和生成网络来提升整体效果。此外, 模型当前仅用于单张人脸图像的情况, 未来可以扩展到视频序列或多视角人脸图像上, 从而实现更广泛的应用前景。总体来说, 本文基于自监督的思想为人脸多视角生成任务提供了一种可行且有效的解决思路。

目前工作依然存在缺点，未来需要更多地对生成图像的细节进行把控。如下图所示，由于 3DMM 重建渲染得到的图片本身缺乏高频细节，因此生成的图像高频细节，如皱纹，雀斑等等一些脸部上的细节无法精准控制，未来希望能够从原图高频域中提取相关特征来进行相关改进。同时我们的图像也无法控制帽子，胡子，头发等等的外观条件，需要引入 face parsing 这样更为细腻度的人脸分割指导图像生成。

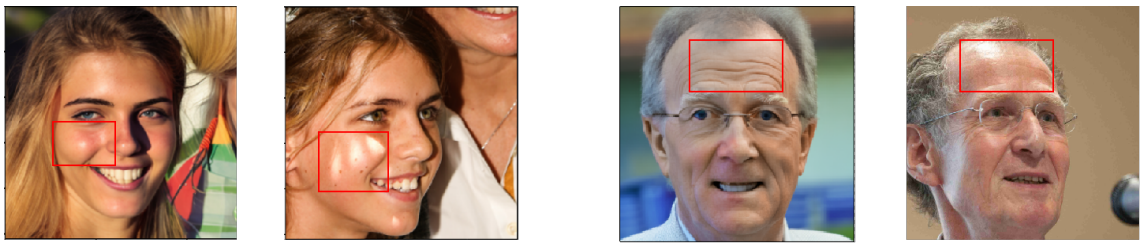

图 7. 高频细节的不准确

图 8. 外观条件不可控

参考文献

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[2] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 48–57, 2017.

[3] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[7] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023.

[8] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

[10] 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.

[11] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018.

[12] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 9224–9232, 2018.

[13] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[16] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[17] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 2439–2448, 2017.

[18] 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.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, et al. Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. *arXiv preprint arXiv:2312.03763*, 2023.

[21] Yuanzhen Luo, Qingyu Zhou, and Feng Zhou. Enhancing phrase representation by information bottleneck guided text diffusion process for keyphrase extraction. *arXiv preprint arXiv:2308.08739*, 2023.

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[23] 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.

[24] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pages 460–469. IEEE, 2016.

[25] 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.

[26] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020.

[27] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[28] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[30] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5911–5920, 2020.