中国科学: 技术科学
SCIENTIA SINICA Technologica
techcn.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
论 文

基于Prompt与ControlNet协同的人脸姿态转向方法

刘汉青, 廖浩然, 肖敏华, 庞孟*

南昌大学数学与计算机学院, 南昌 330031
* E-mail: mengpang@ncu.edu.cn

收稿日期: 2025-07-07; 接受日期: 2025-08-29; 网络版发表日期: 2025-09-16
国家自然科学基金(批准号: 62466036)资助项目

摘要 随着跨姿态、跨角度的人脸理解与生成任务在身份识别、人机交互、虚拟数字人等领域的应用需求不断提升, 高保真、多视角的人脸合成已成为生成式视觉模型研究的重要方向之一. 虽然传统的基于生成对抗网络的方法(如DR-GAN)在二维图像空间实现了多角度人脸生成并在一定程度上保持了身份信息, 但其模型训练过程常伴随不稳定、模式坍塌以及生成图像细节不足等问题. 为此, 本文采用更加稳定快速收敛的Diffusion模型, 提出一种融合Stable Diffusion和ControlNet的人脸姿态重定向方法, 并且进一步设计了Prompt与ControlNet的协同优化策略, 使两者互补增强, 经过实验验证, 本文提出了一个在语义引导方面表现比ChatGPT生成的更加优秀的简短的Prompt, 并且通过本文方法能够实现不同姿态人脸正脸化. 在Multi-PIE与VIS混合数据集上的定量实验结果表明, 本文所提方法在身份保留度与姿态精度方面均显著优于现有方法, 本方法较现有方法在图片生成质量、ID保持度以及角度差异方面均有提升.

关键词 人脸姿态控制, Prompt设计, 扩散模型, 生成式模型

1 引言

人脸姿态转向是计算机视觉领域中的重要课题, 在跨姿态人脸识别、身份验证、人机交互以及虚拟现实等应用场景中发挥关键作用. 其核心目标是在保持人物身份一致的前提下, 将输入图像中的人脸转换为指定目标姿态.

早期的代表性工作之一是Tran等人[1]提出的DR-GAN(disentangled representation learning GAN)模型, 此方法以生成对抗网络架构(GAN)为基础, 借助解耦身份特征与姿态信息, 在生成器中采用编码器-解码器结构来提取身份表示, 并且引入姿态编码用以控制目标姿态生成. 判别器负责进行图像真假判断、身份识别以及姿态分类, 以此达成姿态不变的身份表示学习, 经由实验证实, DR-GAN在Multi-PIE和CASIA-WebFace等数据集上, 相较于传统CNN和基础GAN模型, 呈现出更为优良的识别性能, 在应对大角度姿态变化时, 依旧可生成保持身份一致且姿态合理的人脸图像.

近年来, 随着生成建模技术的不断进步, 越来越多的研究致力于在人脸姿态转向任务中实现更精细的姿

引用格式: 刘汉青, 廖浩然, 肖敏华, 等. 基于Prompt与ControlNet协同的人脸姿态转向方法. 中国科学: 技术科学, 2025, 55
Liu H Q, Liao H R, Xiao M H, et al. Pose-controllable face synthesis via Prompt and ControlNet collaboration (in Chinese). Sci Sin Tech, 2025, 55, doi: 10.1360/SST-2025-0177

© 2025 《中国科学》杂志社

态控制与身份保持. 例如, TP-GAN(two-pathway GAN)[2]采用局部与全局路径相结合的网络结构, 通过融合面部关键区域(如眼睛、鼻子、嘴部)的局部信息与整体结构特征, 有效提高了在大姿态变化条件下的生成质量和身份一致性. 此外, CAPG-GAN(couple-agent pose-guided GAN)[3]引入姿态引导模块与上下文感知机制, 使得生成网络能够充分利用姿态先验信息, 实现更加细致的姿态迁移.

除了完全基于GAN的方法外, 也有研究尝试引入三维人脸建模与渲染技术以增强几何结构建模能力, 例如FF-GAN(face frontalization GAN)[4]利用3DMM参数作为中间表示, 对极端角度下的人脸进行逼真还原.

尽管上述方法在姿态转向质量与身份保真度上取得一定进展, 但仍普遍受限于网络结构复杂、训练代价高, 对大规模姿态标注数据依赖性强, 难以实现连续精细的姿态控制. 与之相比, 扩散模型凭借逐步生成机制与天然可控性, 在处理复杂结构变形(如人脸旋转)时展现出更大潜力. 尤其是结合文本 Prompt 与控制图的“双通路”驱动方式, 不仅能增强生成图像的身份一致性, 还能显著提升姿态调节的精度与泛化能力.

为进一步提升该双通路机制的引导效果, 作者团队提出了一种面向姿态控制任务的Prompt优化策略. 作者团队发现, 设计一条“高效的Prompt”需要同时满足以下几个关键特性.

(1) 语义明确: 姿态描述应简洁清晰, 诸如“Look 30 degrees to the left”比“Turn left 30 degrees”更贴合模型的语言理解偏好;

(2) 结构紧凑: 过长的描述或包含低频词汇的Prompt会稀释关键信息, 引发注意力冲突, 降低图像细节的稳定性;

(3) 词频友好: 应优先使用高频且语义明确的词汇, 以提升CLIP模型生成的文本嵌入质量.

本文将在后续小节中通过系统化实验, 验证Prompt的长度、词频和结构对生成结果的具体影响, 并基于实验结果提出一种面向人脸姿态控制的高效Prompt优化策略.

如图1所示, 本文方法以Stable Diffusion与ControlNet的结合为架构以实现人脸姿态转向. 在该架构中, 用户可通过描述性文本(如“Look 30 degrees to the left”)输入姿态指令, 指令会首先被CLIP模型编码为Prompt embedding[5], 随后通过交叉注意力机制注入至

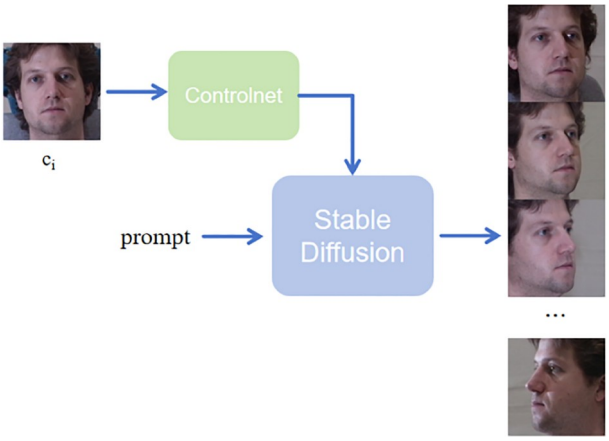

图 1 (网络版彩图)通过控制图和引导词的双重控制, 模型能同时兼顾身份保持和姿态转向

Figure 1 (Color online) By leveraging dual conditioning with control maps and textual Prompts, the model achieves both identity preservation and pose transformation simultaneously.

U-Net的多个中间层, 从而对扩散过程进行引导. 用户也可输入控制图像, ControlNet将其作为条件输入, 提取与主网络对齐的空间特征, 并通过残差连接注入主干网络. Stable Diffusion本质上是在潜空间中进行图像建模, 其U-Net网络接收的输入为潜变量$z \in R^{4\times64\times64}$(包含4个通道、空间分辨率为64×64的张量), 并非直接的图像像素[6]. 为了在此空间内加入图像引导条件(如人脸结构或身份特征), 引入了ControlNet模型作为辅助分支, ControlNet接收姿态控制图或关键点图c_i作为输入条件[7]. ControlNet使用浅层复制的U-Net结构提取特征, 并与主扩散网络通过残差连接的方式融合, 从而有效保留输入图像的结构信息.

本文做出了以下贡献: (1) 提出将ControlNet与Stable Diffusion相结合用于人脸姿态转向任务; (2) 利用Prompt与控制图双重引导, 实现身份保持与姿态可控; (3) 设计出一种简洁合适的Prompt.

2 相关工作

2.1 扩散模型

近年来, 扩散模型(diffusion model)凭借其稳定的训练机制与优异的生成质量, 逐渐成为图像生成任务中的主流范式, 逐步替代传统的生成对抗网络(GAN). 扩散模型的基本思想最早由Sohl-Dickstein等人[8]提出, 其核心思想是将图像生成建模为一个逐步加噪与反向

去噪的马尔可夫链过程. 在前向过程中, 输入图像逐步添加高斯噪声直至完全变为噪声; 而在反向生成阶段, 模型学习如何从纯噪声中逐步还原出目标图像. 该方法通过最大化数据的似然函数, 在理论上能够逼近任意数据分布, 具有良好的建模能力.

Ho等人[9]随后提出了去噪扩散概率模型 (DDPM), 对扩散模型的去噪过程进行了简化与优化, 显著提升了生成图像的清晰度与多样性. DDPM采用均匀设置的时间步、固定的高斯加噪过程以及一个训练的去噪网络, 成功将扩散模型的生成质量推至当时SOTA水平.

在DDPM框架的基础上, Stable Diffusion[6]引入了跨模态的文本引导机制, 成为可控图像生成的重要里程碑. 该方法首先将文本Prompt通过CLIP文本编码器映射为语义嵌入向量, 再通过Cross-Attention的方式嵌入至U-Net模型的多个阶段, 从而实现对生成图像的精确语义引导. 同时, Stable Diffusion在潜空间中进行建模, 有效降低了计算开销, 使其在实际应用中更具可行性.

为了进一步增强模型对结构化条件的感知能力, ControlNet[7]被提出作为Stable Diffusion的结构增强模块. ControlNet在原有扩散模型结构之外添加了一个并行的残差网络分支, 专门用于接收和编码结构引导图(如边缘图、关键点图、深度图等). 该模块通过将结构条件信息以残差形式注入主干网络, 使得模型在保证结构一致性的同时仍能自由生成符合语义引导的图像, 特别适用于如人脸姿态转向、表情控制、草图重建等结构受限任务.

2.2 人脸姿态合成方法

人脸姿态合成技术大致可划分为三个发展阶段: 基于2D-GAN的显式姿态建模、基于3D表达的几何建模方法, 以及近年来兴起的基于扩散模型的概率生成方法. 每一阶段在身份保持、姿态精度、生成质量与训练成本之间做出不同权衡, 本文对其代表性工作进行简要综述, 并说明本文方法的定位与优势.

2.2.1 基于2D-GAN的显式姿态编码

早期人脸姿态合成主要依赖于生成对抗网络(GAN)框架. 代表性的DR-GAN方法[1]首次引入身份与姿态解耦机制, 通过姿态标签作为条件输入, 并在判别器中共同承担身份识别与姿态分类任务, 实现了条件生成的基本能力. 随后, TP-GAN[2]采用局部-全局双分支网络结构, 分别对五官区域与整体轮廓建模, 提升了大角度下的保真度. CAPG-GAN[3]进一步引入上下文感知模块, 以适应复杂姿态变化下的光照与背景建模问题.

此外, 如G2-GAN(geometry-guided GAN)[10]和HF-PIM(high fidelity pose invariant model)[11]等方法也在姿态条件编码方式、面部区域增强等方面做出探索, 逐步提高了身份保持与细节保真能力. 然而, 由于此类方法缺乏几何约束, 对姿态标签质量依赖较高, 导致在无标签或自然场景中泛化能力较弱, 且姿态调控往往呈离散化、分类化, 缺乏连续可控能力.

2.2.2 基于3D先验的生成方法

为克服2D-GAN方法对标签依赖与几何理解不足的问题, 近年来的研究逐渐转向将3D建模技术引入生成过程. 其中, pi-GAN(periodic implicit GAN)[12]首次将神经体渲染与生成网络结合, 通过隐式场景表示实现了高质量、连续可控的视角变换. EG3D(efficient geometry-aware 3D)[13]在此基础上引入三维显式网格结构, 结合StyleGAN架构, 显著提升了训练效率与分辨率.

与此同时, StyleRig[14]和GRAF(generative radiance fields)[15]等方法探索了控制变量与可微渲染器的联合建模, 从而实现姿态、表情、光照等多因素的独立调节. 尽管这些方法具有较强的可解释性与连续控制能力, 但由于其高度依赖三维模型、参数拟合过程复杂, 通常对算力与数据结构提出较高要求, 训练成本显著高于传统GAN.

2.2.3 基于扩散模型的条件生成方法

扩散模型因其稳定的训练机制与优越的生成质量, 在图像合成领域逐步取代GAN成为主流. FaceDiffuser[16]结合姿态热图与去噪采样轨迹控制, 有效提升了跨姿态合成的平滑性与连贯性. PoseDiffusion[17]则以姿态关键点图为条件输入, 在U-Net结构中引入pose-aware attention模块, 提升了姿态控制的响应度.

近期, DiffFace[18]和PhotoVerse[19]等方法进一步融合了ID向量、关键点与语义信息, 提升了身份保持与局部区域还原的能力. 这些方法多数基于Stable Diffusion框架, 在潜空间进行建模, 辅以条件嵌入模块(如

cross-attention或ControlNet)实现多模态控制.

2.2.4 本文方法的定位与优势

本文方法基于Stable Diffusion, 并在其去噪-采样链路中引入ControlNet, 实现“三元组”输入(x, x_t, p): 原图x保留身份特征, 结构控制图x_t明确指定目标姿态, 而文本Prompt p细化语义指令. 这种“语义和结构双通路”设计兼顾了以下几点.

(1) 身份保持: 不依赖显式身份向量, 而是通过噪声重构与skip-connection直接注入原图特征;

(2) 姿态精度: 关键点/边缘图在ControlNet中以残差方式注入, 对局部几何具备“硬约束”;

(3) 训练代价: 无需3D网格或渲染器, 也无需大规模姿态标签, 适合野外非结构化数据;

(4) 扩展性: 仅需替换控制图即可迁移到表情、光照等其他属性编辑任务.

3 本研究的方法

3.1 Stable Diffusion

传统扩散模型在采样过程中, 需要在每一步中将完整图像输入到U-Net中进行处理. 当总扩散步数T较多、图像分辨率较高时, 这一过程会变得非常缓慢. Stable Diffusion针对这一问题, 提出了一种更高效的解决方案. Stable Diffusion的核心思想是将扩散过程从图像空间转移到潜在空间中. 该方法使用一个训练好的编码器E将图像压缩成低维的潜在表示; 解码器D负责将潜在表示还原为图像. 这一结构大大减少了计算量, 同时保留了图像的主要特征.

此外, Stable Diffusion在条件控制方面也做出了重要改进. 它引入了交叉注意力机制, 对去噪网络中的U-Net进行增强. 通过这种方式, 模型可以将外部信息作为条件输入, 在生成过程中进行更精细的控制. Stable Diffusion成为一种条件图像生成器.

在训练方面, Stable Diffusion的目标函数与传统扩散模型类似. 不同之处在于它的输入不再是原始图像x_t, 而是潜在空间中的表示z_t. 同时模型还接收经过处理的条件输入$\tau_\theta(y)$. 最终的损失函数可以表示为

$$L_{\mathrm{LDE}} = E_{t, z_0, \varepsilon, y}\left[\left\|\varepsilon - \varepsilon_\theta(z_t, t, \tau_\theta(y))\right\|^2\right]. \tag{1}$$

3.2 ControlNet

ControlNet是一种用于控制图像扩散模型的神经网络结构. 它的主要作用是接收结构化输入(如边缘图、深度图或姿态图), 并将其作为条件信息, 引导图像生成的内容与结构. ControlNet通过这种方式实现对图像生成过程的精确操控.

ControlNet并不会直接修改原始的Stable Diffusion模型参数. 相反, 它会复制一份U-Net的参数作为一个独立的可训练分支. 原始参数保持冻结状态, 避免在微调时被改动. 这种做法的一个优势是防止在小规模数据集上训练时出现过拟合, 同时保留原始模型在大规模训练中获得的泛化能力.

ControlNet的训练依赖于网络中的“零卷积层”. Zero Convolution在初始化时其权重与偏置被设置为零. 在训练的初始阶段, 这些层不会对网络的前向传播结果产生影响, 因此模型的行为与未引入ControlNet时保持一致. 但在反向传播之后, 梯度会更新这些参数, 从而逐渐引导模型利用结构条件信息参与生成过程. 这种设计机制可以确保训练过程的稳定性. 虽然在初始状态下卷积层权重为零, 导致输入特征x对输出y的梯度为零(即$\partial y / \partial x = 0$), 但权重本身的梯度$\partial y / \partial \omega = 0$仍为非零. 因此在一次梯度下降后, 权重$\omega$将不再为零. 之后, ControlNet开始逐步影响生成结果.

3.3 通过Prompt实现的条件引导生成

3.3.1 问题陈述

在人脸姿态转向任务中, 模型的输入为一张正面人脸图像x, 以及其对应的标签$y = \hat{x}, y_p$, 式中$\hat{x}$为目标姿态下的真实图像, y_p表示目标的转向角度或姿态描述. 本任务的核心目标是在保持人物身份不变的前提下, 利用引导词将输入图像转换为目标姿态的人脸图像.

3.3.2 网络架构

为了实现这一目标, 本文提出一种基于Stable Diffusion和ControlNet的联合框架. 该方法的目标为使用一个更优的Prompt进行图像姿态控制和实现身份保持. 其中, Stable Diffusion模型负责图像生成与姿态控制, 而ControlNet用于在整个生成过程中保持身份特征

一致.

具体来说，本文将目标姿态标签y_p以自然语言提示(Prompt)的形式输入至Stable Diffusion的文本编码器$\tau_\theta(\cdot)$, 生成相应的语义嵌入表示$eprompt = \tau_\theta(y_p)$. 该提示嵌入通过交叉注意力机制注入至U-Net的多个中间层, 以控制图像生成过程中人脸的转向方向.

3.3.3 训练过程

为了训练具有跨姿态合成能力的模型，本文将Multi-PIE和VIS两个数据集进行了混合训练. 对于这两个数据集，统一采用正脸图像作为模型输入，将其余角度的图像作为模型的生成目标.

具体来说，在Multi-PIE数据集中，本文选择051号摄像机的图像作为原始输入图像. 其他摄像机视角对应的图像则作为目标图像，用于训练模型生成指定姿态下的人脸. 在VIS数据集中，本文采用a类角度图像作为输入图像，将e类和f类角度图像作为模型需要生成的目标图像.

为了更有效地引导Stable Diffusion生成出符合目标姿态的人脸图像，本文为每一对输入图像与目标图像构建了一个对应的文本提示(Prompt). 例如“Look 30 degrees to the left. ”或“Look left. ”. 所有Prompt被保存在.txt文件中，并按数据集分别管理，Multi-PIE和VIS有各自独立的文本文件.

此外，本文将输入图像、目标图像和对应的Prompt三者进行配对，构成模型训练所需的三元组(x, x_t, p), 式中x表示原始正脸图像, x_t为目标姿态图像, p为姿态描述的Prompt. 以这种三元组的方式，将x传入ControlNet, 将p传入Stable Diffusion, 并将x_t作为目标图像进行训练. 基于此, 本文构建了一种结构清晰的训练数据格式，使得模型能够在保持身份一致的前提下, 实现从正脸到目标姿态的人脸图像生成.

3.3.4 采样过程

本文的推理流程设计受到了Zhang等人[20]在GitHub上开源项目的启发. 为实现图像生成的可控性, 引入了两个关键参数: 控制强度(strength，记为λ)和引导尺度(scale, 记为γ). 它们分别用于调节控制图与文本条件在生成过程中的影响力, 其数学定义如式(2)和(3)所示:

$$\lambda_i = \lambda \cdot (0.825)^{12-i}, \tag{2}$$

$$\epsilon_\theta = \epsilon_\theta^{cond} + \gamma(\epsilon_\theta^{cond} - \epsilon_\theta^{uncond}), \tag{3}$$

式中，λ_i控制每一层U-Net中控制图的作用强度. 浅层侧重保留图像的几何结构(如轮廓、视角)，深层则细化局部内容(如纹理、边缘). γ控制语义引导的幅度: 其值越大，图像越贴合文本提示，但可能削弱几何一致性; 反之则优先保留控制图结构，牺牲语义完整度. 采样流程如下.

首先，输入图像input_image被统一调整为预设尺寸image_resolution，并转换为三通道格式. 随后将图像归一化为[0, 1]范围, 生成控制图detected_map. 该控制图被复制num_samples次，并重排为四维张量$control \in R^{B\times3\times H\times W}$, 作为模型的条件输入.

为能够复现采样结果, 采样过程设定随机种子. 随后，模型使用文本编码器分别对正向提示prompt +a_prompt和负向提示n_prompt进行编码，得到对应的交叉注意力嵌入向量c_crossattn, 并结合图像控制条件c_concat构建条件向量字典. 同时构造条件和无条件输入, 供DDIM采样器使用.

潜空间张量的初始形状为(4, H//8, W//8), 对应生成器的隐空间特征维度. 13层控制权重统一设为λ. DDIM采样器输出的潜变量samples被解码为图像空间表示. 解码过程通过decode实现，最终生成RGB图像$x_sample \in R^{B\times3\times H\times W}$. 图像经缩放和裁剪处理，转换为uint8类型, 构成最终输出.

3.3.5 Prompt设计过程

(1) 初始方案与问题暴露

本研究旨在实现基于文本语义控制的人脸姿态转向图像生成，同时尽可能保持身份一致性与细节质量. 因此, Prompt的设计不仅要准确表达“转向角度”, 还需突出“身份保持” “图像清晰度”以及“语义一致的风格指令”. 在初步尝试中，本文通过向ChatGPT提供任务指令，生成了如下Prompt示例: “Turn+方向+角度. Extremely detailed portrait of the same people, set against the same background, with no blurry, pixelated, or distorted details. The portrait should have a realistic style with no exaggerated cartoonishfeatures. Ensure that all textures on smooth surfaces are consistent, realistic,

and refined. Avoid overly simplified, abstract, or cartoonish elements."

然而, 在多轮实验中发现, 以上Prompt虽然语义描述详细丰富, 但实际生成效果并不理想. 如图2和表1所示, 图像中出现了重叠人像、局部模糊和纹理不一致等问题. 为此, 本文深入分析了Prompt与图像生成之间的关联机制, 并结合文献[21]的观点认为, CLIP文本编码器最多处理75~77个有效Token. 超长Prompt会触发Token截断或平均加权, 导致后部关键信息(如"same people"或"refined texture")被稀释. 由此展开多阶段Prompt优化.

(2) 阶段一: 移除negative_prompt

由于负面Prompt并非直接跳过无关内容, 而是在后期diffusion步骤对相应视觉特征做"相位抵消", 泛化词更可能误伤细节[22]. Stable Diffusion的文本引导基于classifier-free guidance技术, 其采样过程遵循以下

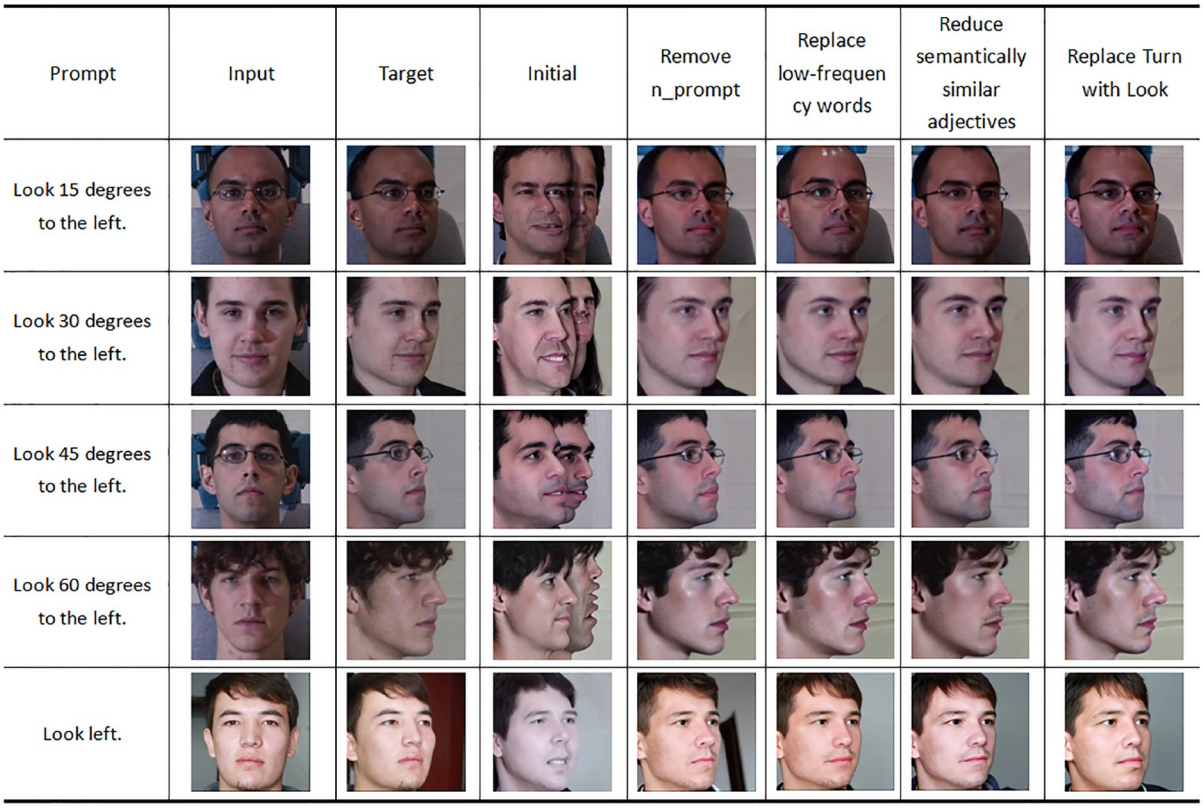

图 2 (网络版彩图)各版本Prompt输出结果

Figure 2 (Color online) Output results of Prompts across versions.

表 1 各版本Prompt指标评估结果[a)]

Table 1 Evaluation results of Prompt metrics across versions

Prompt	PSNR↑	SSIM↑	LPIPS↓	APD↓	cos(θ)↑
Prompt 0	15.01	0.505	0.406	7.428°	0.9367
Prompt 1	23.03	0.797	0.127	0.600°	0.9589
Prompt 2	24.08	0.819	0.106	0.648°	0.9525
Prompt 3	25.80	0.825	0.108	0.600°	0.9597
Prompt 4	24.54	0.819	0.110	0.602°	0.9597

a) ↑表示数值越大越好, ↓表示数值越小越好(表3同理). 各数值是模型在不同 Prompt下的实际性能表现.

噪声引导公式:

$$z_{guided} = (1+\omega)\cdot\varepsilon_{null} - \omega\cdot\varepsilon_{neg}, \quad (4)$$

式中, ε_{neg}表示模型对负向Prompt的预测结果. 理论上, 负向Prompt是模型应规避的图像特征; 但若其中含有blurry、pixelated等词, 模型将倾向于抑制所有图像中的“模糊区域”, 包括图像本应存在的纹理高频信号, 进而影响生成图像的清晰度和锐利度.

因此, 本文完全移除negative_prompt, 仅保留正向内容作为输入, 得到优化后的Prompt 1. 从图2可见, 生成图像的细节质量显著提升, 由原先的人脸重叠现象转变为结构清晰、单一的人像图像.

(3) 阶段二: 替换低频词

进一步分析Prompt 1后发现, 其中包含诸如“extremely” “refined” “textures on smooth surfaces”等长尾表达. 这些词在CLIP的语料训练集中出现频率较低, 导致其嵌入向量偏离主干语义空间, 不仅语义不稳定, 还可能引发语义偏移[23].

本文借助BPE(Byte-Pair Encoding)分词器将Prompt 1拆分为Token, 并观察其Token ID分布: [49406, 2105, 1823, 274, 271, 8000, 269, 6519, 12609, 5352, …].

高编号且被拆分的Token往往对应低频或组合词. 表2总结了主要低频片段. 将extremely, refined, textures on smooth surfaces等替换为高频表达(如photorealistic, sharp, clean skin), 得到Prompt 2. 实验评估发现客观指标(PSNR/SSIM/LPIPS)略有提升, 但表情上仍与目标图像有所差别.

(4) 阶段三: 合并冲突形容词

Prompt 2中使用了多个风格形容词(如realistic, sharp, soft, refined), 这可能导致Cross-Attention层产生冲突信号. 根据prompt-to-prompt[24]与Dynamic Prompt Optimizing[25]的研究, 多个具有高引导权重的风格词将激活U-Net编解码器的不同注意力方向, 可能出现attention overlap, 进而破坏局部一致性, 使图像生成结果“上下不一致” “半糊半锐”.

为此, 本文在Prompt 3中将所有的具有重叠含义的形容词替换成高频且单一的单词same, 如portrait of the same person, set against the same background, 以保证语义聚焦. 结果表明, 尽管该Prompt在各项评估指标上表现最优, 但在定性观察上不符合人类视觉系统.

(5) 阶段四: 调整姿态表达

Prompt 3中使用“Turn left 30 degrees”来表达姿态指令, 但“Turn”一词在CLIP中可能引起歧义: 该词常见于全身/躯干旋转相关语义中, 易使模型生成肩膀或上半身扭转的图像[5], 从而偏离ControlNet所提供的head-pose控制图. 因此本文将其替换为“Look 30 degrees to the left”, 该表达更自然、直观且频率更高, 语义聚焦在面部方向调整上, 形成最终的Prompt 4. 这也符合human perception更倾向于“Look”来描述头部转动, 而非“Turn”.

Prompt 1: Turn left 30 degrees. Extremely detailed portrait of the same people, set against the same background. Realistic style. All textures on smooth surfaces are consistent, realistic, and refined.

Prompt 2: Turn left 30 degrees. Detailed portrait of the same person, with the same background. Photo-realistic. Skin texture is clear, clean, and sharp.

Prompt 3: Turn left 30 degrees. Portrait of the same person, set against the same background.

Prompt 4: Look 30 degrees to the left. Portrait of the same person, set against the same background.

为确保转向后生成图像中人物的身份一致性, 本实验以源图像(受试者的正脸图像)作为身份保持的基

表 2 低频Token片段分析示例

Table 2 Example analysis of low-frequency Token segments

Token ID	Original Token	Characteristics
12609	Extremely	Long-tail word, highly abstract, not core to image description
7965	Detailed	Medium frequency, strongly descriptive, but less common than realistic or sharp
28063	Textures	Particular nouns appear less frequently in the training set than face, skin, or light
20746	Refined	Abstract, literary-style vocabulary, generally a low-frequency Token in CLIP
26098	Surfaces	Technical term, low frequency

础. 同时, 通过精心设计的提示词(Prompt)引导模型生成具有指定方向和角度的人脸图像, 从而实现对姿态的精准控制.

各个Prompt中, 红色的提示词负责姿态控制, 明确头部姿态的方向和角度; 蓝色的提示词负责细节控制, 强调图像的清晰度、质感、纹理等高质量生成特征; 绿色的提示词负责身份保持控制, 强调保持"同一个人"的身份一致性, 避免面部结构变化. 若想要生成受试者其他方向与角度的侧脸, 可按需修改提示词中的方向与角度.

(6) 结果分析

表1和图2展示了各组Prompt的采样效果. 可以观察到:

(i) 从Prompt 0到Prompt 1: 移除负向描述后, 高频细节显著恢复, 生成图像由原先的人脸重叠现象转变为结构清晰、单一的人像图像;

(ii) 从Prompt 1到Prompt 2: 替换低频词带来轻微清晰度提升, 但表情上仍与目标图像有所差别;

(iii) 从Prompt 2到Prompt 3: 删除冗余形容词后, 姿态一致性和纹理锐度均达到最佳, 但定性观察上不符合人类视觉系统;

(iv) 从Prompt 3到Prompt 4: 将转向描述替换成常见表达后, 定量评估上差异不大, 在定性观察上更符合人类的视觉系统.

最终, 本文采用Prompt 4作为本文方法的默认文本条件, 既确保了姿态控制信号的集中, 又避免了语义冲突与频率噪声, 显著提升了生成质量.

3.3.6 与先前方法的比较

本方法旨在在保持身份一致性的前提下, 实现高质量的人脸姿态转变. 与近年来提出的先进方法相比, 本文的方案在建模方式、控制粒度以及生成质量上具有显著优势. 具体的定量比较结果可见表3.

基于StyleGAN的方法(如StyleFlow[26], InterFaceGAN[27])通过在潜空间中插值或寻找方向向量, 实现属性控制, 但这类方法对姿态建模不够直接, 且缺乏对目标姿态的精确对齐控制. 一些3D-aware GAN(如pi-GAN[12], EG3D[13])引入隐式体积渲染或3D先验以增强视角一致性, 但通常需要大规模数据和复杂渲染机制, 训练成本较高.

此外, Diffusion-based方法(如DiffFace[18])在高质量人脸图像生成任务中展现出强大潜力, 其通过逐步采样建模图像分布, 在图像保真度方面取得显著提升. 然而, 当前多数扩散模型在姿态或结构的显式控制方面仍存在不足, 难以同时实现身份保持与目标姿态的准确对齐.

相比之下, 本文提出的方法构建于Stable Diffusion框架之上, 借助ControlNet有效结合了结构控制图与文本语义提示(Prompt), 从而实现对姿态的显式建模与细粒度控制. 采用的三元组(x, x_t, p)输入形式, 使得模型在不显式编码身份向量的前提下, 仍可保持高身份一致性. 整体框架同时具备强大的生成能力与结构对

表 3 不同模型实现人脸姿态转向的评估结果

Table 3 Evaluation results of face pose transformation across different models

Model	PSNR↑	SSIM↑	LPIPS↓	APD↓
DR-GAN	11.3	0.36	0.50	16.0°
TP-GAN	13.2	0.56	0.42	4.7°
DaGAN[32,33]	16.9	0.73	0.24	1.5°
Rome[32,34]	7.5	0.72	0.43	1.5°
Face2Face[32,35]	171	0.76	0.27	1.5°
DPE[32,36]	14.7	0.69	0.38	4.2°
StyleHeat[32,37]	17.5	0.79	0.28	3.4°
HyperReenact[32,38]	18.9	0.80	0.25	0.5°
DiffusionRig[32,39]	15.6	0.76	0.38	1.3°
Ours	24.5	0.82	0.11	0.6°

齐能力, 在多个定量指标上均优于现有方法.

4 实验结果

4.1 实验设置

本研究旨在探究文本提示(Prompt)在受控图像生成任务中的作用, 特别是在多姿态、多光照与表情变化条件下, 模型对人脸身份特征的保持能力. 为此, 本实验选用了结构严谨的多视角人脸图像数据库Multi-PIE作为核心评估数据集.

鉴于本实验获得的Multi-PIE数据量庞大, 且部分视角下存在图像缺失, 为控制变量、提升实验的可重复性与可比性, 最终仅选择光照编号为10的图像(最接近自然光照条件). 该策略有助于降低非自然光源对生成质量的影响, 从而提高实验结果的稳定性和可信度.

在数据筛选过程中, 本实验剔除了所有包含非中性表情(如微笑、张嘴等)以及大角度侧脸(75°与90°)的图像. 该决策的依据在于, 尽管扩散模型具备较强的语义理解和图像合成能力, 但在极端表情或大角度姿态下保持面部身份一致性仍具有挑战. 当输入图像中存在明显的表情变化或遮挡时, 模型在重建过程中容易偏离原始身份, 甚至出现“变脸”现象[28]. 此外, 大角度侧脸通常伴随面部关键区域(如眼睛、鼻梁、嘴部等)的严重遮挡, 信息缺失使得生成图像更易出现模糊、结构扭曲等问题[29].

本实验使用的图像来自Multi-PIE与VIS两个公开数据集. Multi-PIE中选取了249位被试者, 每位被试包含9个拍摄角度的图像, 其中051号摄像机的正面图像作为输入, 其余8个角度(041、050、080、090、130、140、190、200)图像作为目标图像; VIS数据集中共选取564位被试者, 每人包含5个视角图像, 实验中使用a类图像作为正面输入, e与f类图像分别作为左右目标图像. 两个数据集合并后, 共计使用3933张图像样本.

在文本提示语(Prompt)设计方面, 初期本实验尝试使用较为复杂且带有显式质量约束的语句, 例如: “Turn left 75 degrees. Extremely detailed portrait of the same people, set against the same background, with no blurry, pixelated, or distorted details. The portrait should have a realistic style with no exaggerated cartoonish features.” 尽管该类Prompt在语义层面较为完整, 但实验发现, 其实际生成效果反而较差, 图像常出现模糊、伪影等问题. 分析认为, 这类过度细化的约束可能限制了模型生成空间, 从而导致图像缺乏自然感与细节表现. 最终本实验采用了结构更为简洁的表达方式, 例如: “Look 75 degrees to the left; portrait of the same person, set against the same background. ” 这种简化的提示语在保持语义清晰的同时, 允许模型有更大的生成自由度, 整体生成效果更为稳定和自然.

图像预处理阶段, 所有输入与目标图像均统一调整为256×256分辨率. 输入图像被转换为RGB色彩空间, 并归一化至[0, 1]; 目标图像归一化至[−1, 1]区间. 在早期实验中, 本实验曾尝试在采样阶段使用512×512分辨率, 但生成图像中频繁出现多张人脸、背景错乱等异常现象. 进一步分析发现, 这并非模型本身空间建模能力不足所致, 而是由于训练过程中模型仅见过256×256分辨率的图像, 而采样阶段却强行提升至更高分辨率, 导致模型在未见数据分布下泛化失败. 因此本实验为确保训练与生成阶段一致性, 提升图像质量与稳定性, 最终统一采用256×256分辨率进行训练与推理.

在训练策略方面, 数据集按8:2的比例划分为训练集与测试集. 对扩散模型中的关键超参数进行了系统性调优. 对于strength参数, 实验发现当其值高于0.7时, 生成图像与输入图像过于相似, 缺乏姿态或内容变化; 而当其值低于0.5时, 则易造成面部结构缺失、身份模糊等问题. 最终将strength设置为0.65, 以实现图像结构保真性与语义变化之间的平衡.

对于scale参数的调试结果显示: 当其值高于5.2时, 生成图像常出现过曝、高亮区域或边缘重影等不自然现象; 当其值低于1.0时, 则常出现图像模糊、结构不完整的问题. 综合考虑后, 本实验将scale参数设置为3.0, 以兼顾语义控制强度与生成图像的质量稳定性.

为确保实验结果的可复现性, 所有图像生成过程均设定统一随机种子3407. 该设定参考了Picard和Torch [30]关于随机性对生成一致性影响的研究建议, 有助于提高模型生成结果的一致性与对比实验的可信度.

4.2 实验评估

4.2.1 图片质量

本实验采用PSNR、SSIM以及LPIPS三种指标对生成图像与目标图像之间的质量进行定量评估与比

较. 具体计算结果如表1和表3所示.

4.2.2 角度识别

在人脸姿态估计任务中, 通常使用MAE评估三个角度(偏航角、俯仰角、翻滚角)的平均估计误差. 一般而言, 若平均误差控制在5°以内, 表明模型具备较高精度. 以HopeNet为代表的方法在数据集上的误差通常为5°~6°, 被认为是性能良好的基线方法[31]. 对于每个样本, 分别预测三个姿态角度: 偏航角(yaw)、俯仰角(pitch)和翻滚角(roll), 并与真实值进行比较, 计算每个角度的绝对误差:

$$error_{yaw}=\left|\hat{\theta}_{yaw}-\theta_{yaw}\right|, \tag{5}$$

$$error_{pitch}=\left|\hat{\theta}_{pitch}-\theta_{pitch}\right|, \tag{6}$$

$$error_{roll}=\left|\hat{\theta}_{roll}-\theta_{roll}\right|. \tag{7}$$

对于单个样本, 其MAE定义为三个角度误差的平均值:

$$MAE_{sample}=\frac{(error_{yaw}+error_{pitch}+error_{roll})}{3}. \tag{8}$$

在整个测试集上, 总体MAE计算如下:

$$MAE_{overall}=\frac{1}{N}\sum_{i=1}^{N}MAE_{sample_i}, \tag{9}$$

式中, N表示测试集中的样本数.

为了评估姿态角度预测的准确性, 本文对HopeNet模型在测试集上进行了MAE(平均绝对误差)评估, 分别计算了三个欧拉角方向上的误差结果. 如表1和表3[32~39]所示.

4.2.3 ID保持

余弦相似度常用于衡量两个向量之间的相似程度, 其取值范围为[–1, 1], 数值越接近1表示两个向量越相似. 在实验中, 本文将生成图像与对应真实图像的特征向量作为比较对象, 通过计算它们的余弦相似度来评估生成图像与真实图像在特征空间上的相似度.

具体地, 提取生成图像和真实图像的特征表示后, 计算它们的余弦相似度, 反映模型在保持身份特征一致性方面的表现. 余弦相似度的计算公式如下:

$$Similarity(f_1,f_2)=\frac{f_1\cdot f_2}{\|f_1\|\|f_2\|}, \tag{10}$$

式中, f_1和f_2分别代表生成图像和真实图像的特征向量.

4.3 实验一: 人脸多姿态生成

测试集的生成结果如图3所示. 从图中可以观察到, 生成的人脸图像整体保持了较高的身份一致性.

面部关键特征如眼睛、鼻子、嘴巴的位置与结构

图 3 (网络版彩图)人脸多姿态生成输出结果

Figure 3 (Color online) Multi-Pose face generation results.

在多角度条件下保持稳定. 同时, 背景内容在生成过程中保持不变, 提示语中“背景保持一致”的引导效果被成功学习.

此外, 生成图像中不同角度的人脸轮廓过渡自然, 未出现明显的几何扭曲或遮挡错误. 模型不仅学习了角度信息的表达, 还能够较好地重建遮挡区域的面部细节. 然而, 仍有少部分样本在大角度转向时出现模糊现象, 尤其在下巴或头发区域. 这可能与训练集中某些角度的数据分布不平衡有关, 也是后续改进方向之一.

根据表3所示的评估结果, 在多个主流指标(包括图像质量指标PSNR、SSIM、LPIPS以及姿态准确性指标APD)上, 本文提出的方法(Ours)均表现优越, 综合性能显著领先于现有方法. PSNR(24.5)和SSIM(0.82)表明生成图像具有更高的保真度与结构一致性; LPIPS(0.11)显著低于其他方法, 说明本文方法在感知一致性上更加贴近目标图像; APD(0.6°)代表姿态误差极小, 优于大多数现有扩散方法(如DiffusionRig)及重建方法(如Face2Face).

4.4 实验二: 多姿态人脸正脸化

上述任务实现了从正脸角度生成左右不同角度的转向图像. 为探究该方法是否适用于逆向过程(即从左右不同角度转向至正脸), 本实验将target图像与source图像进行互换, 并将所有生成提示(Prompt)统一设置为“Look straight into the lens. Portrait of the same people, set against the same background. ” 其余实验条件保持不变. 经过300个训练轮次(epoch)后, 其测试结果如图4所示.

5 结论

本文围绕基于Stable Diffusion和ControlNet的人脸姿态转向任务展开, 系统性探讨了Prompt设计对图像生成质量和姿态控制精度的影响. 本文以“Turn left 30 degrees”为目标指令, 从ChatGPT生成的超长Prompt出发, 逐步精简、改写并优化提示语, 最终归纳出一套具有代表性和普适性的Prompt精炼流程. 具体来说有以下几点.

(1) Prompt长度与Token稀释问题: 初始的Prompt (Prompt 0)由于过长, 导致Token注意力被稀释, 姿态和身份指令难以准确表达. 其PSNR仅为15.01, SSIM也较低, 仅为0.505, 表明图像质量不佳.

(2) 负面Prompt的副作用: Prompt 1移除negative_prompt后, PSNR从15.01提升至23.03, SSIM提高至0.797, LPIPS降低至0.127, 图像清晰度显著提升, 表明

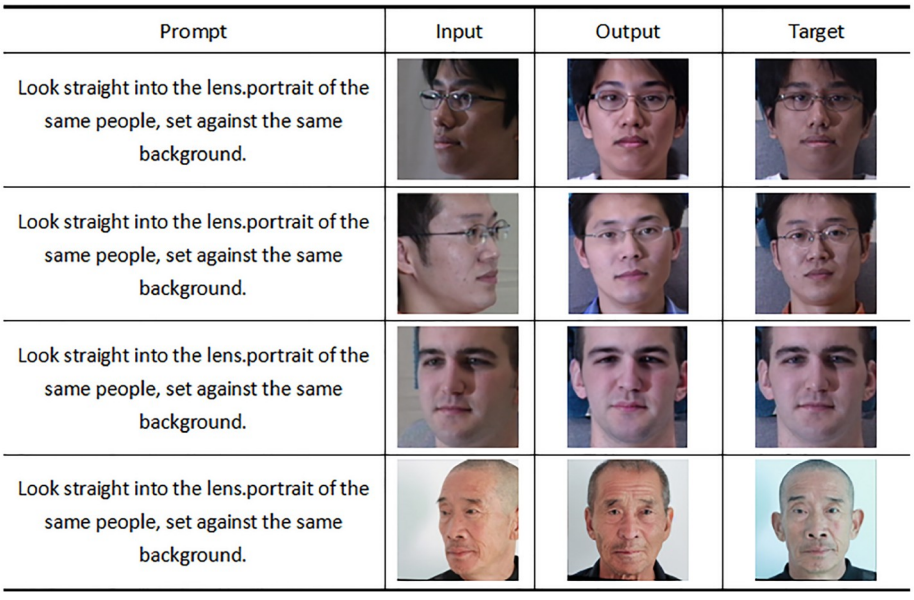

图 4 (网络版彩图)多姿态人脸正脸化输出结果

Figure 4 (Color online) Frontalization results of multi-pose faces.

过度抑制模糊词会牺牲有效细节.

(3) 替换低频词的微调提升: Prompt 2使用高频形容词替换了抽象、低频的表达, 使得SSIM进一步上升至0.819, LPIPS降至0.106, Prompt中Token embedding的词频影响在视觉质量上确有正向作用.

(4) 规避注意力冲突所带来的效益: Prompt 3去除冗余描述, 仅保留核心语义, 其PSNR达到最高25.8, SSIM为0.825, APD角度误差仅为0.600°, 各项指标综合表现最优, 说明减少语义冲突能提升模型聚焦能力.

(5) 语义指令调整的直观优化: Prompt 4将姿态描述由“Turn left”改为更自然的“Look to the left”, 虽然定量指标变化较小(如PSNR仍为24.54), 但在定性观察中, 生成图像的姿态方向与ControlNet输入图一致性更强.

综合来看, Prompt的构造在Stable Diffusion与ControlNet联合驱动的人脸生成任务中发挥着至关重要的作用. 通过对Prompt长度的合理控制、避免使用低频词汇、去除语义冲突的形容词, 以及谨慎处理负面词的使用方式, 能够有效提升生成图像的质量、身份保持的一致性以及姿态控制的精度. 本文所提出的Prompt优化流程在多个实验中均展现出显著优势, 验证了其在语义引导中的稳定性与泛化能力.

此外, 该优化策略不仅在标准姿态变换任务中表现出色, 在不同姿态条件下的人脸正脸化任务中亦展现出良好的鲁棒性与适应性. 这表明所设计的Prompt策略具有较强的通用性, 对后续扩散模型在姿态编辑、表情生成乃至多模态图像合成等任务中的文本控制能力提升, 具有广泛的参考意义和应用价值.

参考文献

1 Tran L, Yin X, Liu X. Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 1415–1424

2 Huang R, Zhang S, Li T, et al. 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. Venice, 2017. 2439–2448

3 Hu Y, Wu X, Yu B, et al. Pose-guided photorealistic face rotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 8398–8406

4 Yin X, Yu X, Sohn K, et al. Towards large-pose face frontalization in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 3990–3999

5 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. Vienna, 2021. 8748–8763

6 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022. 10684–10695

7 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 3836–3847

8 Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the International Conference on Machine Learning. Lille, 2015. 2256–2265

9 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst, 2020, 33: 6840–6851

10 Song L, Lu Z, He R, et al. Geometry guided adversarial facial expression synthesis. In: Proceedings of the 26th ACM International Conference on Multimedia. Seoul, 2018. 627–635

11 Cao J, Hu Y, Zhang H, et al. Learning a high fidelity pose invariant model for high-resolution face frontalization. Adv Neural Inf Process Syst, 2018, 31: 2867–2877

12 Chan E R, Monteiro M, Kellnhofer P, et al. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021. 5799–5809

13 Chan E R, Lin C Z, Chan M A, et al. Efficient geometry-aware 3D generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022. 16123–16133

14 Tewari A, Elgharib M, Bharaj G, et al. StyleRig: Rigging styleGAN for 3D control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020. 6142–6151

15 Schwarz K, Liao Y, Niemeyer M, et al. GRAF: Generative radiance fields for 3D-aware image synthesis. Adv Neural Inf Process Syst, 2020, 33: 20154–20166

16 Stan S, Haque K I, Yumak Z. FaceDiffuser: Speech-driven 3D facial animation synthesis using diffusion. In: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. Rennes, 2023. 1–11

17 Wang J, Rupprecht C, Novotny D. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 9773–9783

18 Kim K, Kim Y, Cho S, et al. DiffFace: Diffusion-based face swapping with facial guidance. Pattern Recogn, 2025, 163: 111451

19 Chen L, Zhao M, Liu Y, et al. PhotoVerse: Tuning-free image customization with text-to-image diffusion models. arXiv: 2309.05793

20 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 3836–3847

21 Zhang B, Zhang P, Dong X, et al. Long-CLIP: Unlocking the long-text capability of CLIP. In: European Conference on Computer Vision. Milan, 2024. 310–325

22 Ban Y, Wang R, Zhou T, et al. Understanding the impact of negative Prompts: When and how do they take effect? In: European Conference on Computer Vision. Milan, 2024. 190–206

23 Yu S, Song J, Kim H, et al. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. arXiv: 2109.03127

24 Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-Prompt image editing with cross attention control. arXiv: 2208.01626

25 Mo W, Zhang T, Bai Y, et al. Dynamic Prompt optimizing for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024. 26627–26636

26 Abdal R, Zhu P, Mitra N J, et al. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM Trans Graph, 2021, 40: 1–21

27 Shen Y, Gu J, Tang X, et al. Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020. 9243–9252

28 Bouzid H, Ballihi L. Facenhance: Facial expression enhancing with recurrent DDPMs. arXiv: 2406.09040

29 Duan Q, Zhang L. Boostgan for occlusive profile face frontalization and recognition. arXiv: 1902.09782

30 Picard D. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. arXiv: 2109.08203

31 Ruiz N, Chong E, Rehg J M. Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, 2018. 2074–2083

32 Bounareli S, Tzelepis C, Argyriou V, et al. DiffusionAct: Controllable diffusion autoencoder for one-shot face reenactment. arXiv: 2403.17217

33 Hong F T, Zhang L, Shen L, et al. Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022. 3397–3406

34 Khakhulin T, Sklyarova V, Lempitsky V, et al. Realistic one-shot mesh-based head avatars. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, 2022. 345–362

35 Yang K, Chen K, Guo D, et al. Face2face ρ: Real-time high-resolution one-shot face reenactment. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, 2022. 55–71

36 Pang Y, Zhang Y, Quan W, et al. DPE: Disentanglement of pose and expression for general video portrait editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023. 427–436

37 Yin F, Zhang Y, Cun X, et al. Styleheat: One-shot high-resolution editable talking face generation via pre-trained styleGAN. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, 2022. 85–101

38 Bounareli S, Tzelepis C, Argyriou V, et al. HyperReenact: One-shot reenactment via jointly learning to refine and retarget faces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, 2023. 7149–7159

39 Ding Z, Zhang X, Xia Z, et al. DiffusionRig: Learning personalized priors for facial appearance editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. Vancouver, 2023. 12736–12746

Pose-controllable face synthesis via Prompt and ControlNet collaboration

LIU HanQing, LIAO HaoRan, XIAO MinHua & PANG Meng

School of Mathematics and Computer Science, Nanchang University, Nanchang 330031, China

With the increasing demand for cross-pose and cross-angle face understanding and generation in applications such as identity recognition, human-computer interaction, and virtual avatars, high-fidelity and multi-view face synthesis has become a key research direction in generative vision models. Although traditional GAN-based methods (e.g., DR-GAN) enable multi-angle face generation in 2D image space while partially preserving identity, they often suffer from training instability, mode collapse, and a lack of fine details. To address these limitations, this paper proposes a face pose redirection method based on a more stable and fast-converging Diffusion model by integrating Stable Diffusion and ControlNet. Furthermore, a collaborative optimization strategy is introduced for Prompt and ControlNet, allowing the two to complement each other. Through extensive experiments, we identify a short, semantically effective Prompt that outperforms those generated by ChatGPT in guidance quality. The proposed method enables accurate frontalization across a wide range of poses. Quantitative results on the combined Multi-PIE and VIS datasets demonstrate that our approach achieves superior identity preservation and pose accuracy compared to existing methods, with consistent improvements in image quality, identity consistency, and angular alignment.

face pose manipulation, Prompt engineering, diffusion model, generative model

doi: 10.1360/SST-2025-0177