计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(10):174−182 [doi: 10.15888/j.cnki.csa.009652]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

文本与关键点协同控制的人脸图像生成①

刘宇同, 王一丁

(北方工业大学 信息学院, 北京 100144)
通信作者: 王一丁, E-mail: wangyd@ncut.edu.cn

摘　要: 人脸图像生成对生成人脸的真实度和可控性有较高要求. 本文提出了一种由文本和脸部关键点协同控制的人脸图像生成算法. 其中文本主要是在语义层面对生成人脸进行约束; 脸部关键点使模型根据给定的脸部位置信息, 控制生成人脸的脸型、表情和细节等属性. 本文算法在现有的扩散模型基础上加以改进, 并额外引入了文本处理模块 (CM)、关键点控制网络 (KCN) 和自编码网络 (ACN). 其中, 扩散模型是一种基于扩散理论的噪声推理算法; CM 基于注意力机制设计, 可以对文本信息进行编码和存储; KCN 接收的是关键点的位置信息, 使生成人脸的可控性得以增强; ACN 缓解了扩散模型的生成压力, 减少生成样本所需的时间. 此外, 为了适配人脸图像这一生成任务, 我们构建一个包含 30 000 张人脸图像的数据集. 本文算法实现了: 给定一段先决条件文本和一张人脸关键点图, 模型可以提取出文本中的特征信息和关键点的位置信息, 生成高真实度和可控性强的目标人脸图像. 通过与目前主流方法进行对比, 本文算法的 *FID* 指标提高了约 5%–23%, *IS* 指标提高了约 3%–14%, 证明了算法的先进性和优越性.

关键词: 人脸生成; 扩散模型; 生成式人工智能; 文本编码; 自动编码器

引用格式: 刘宇同,王一丁.文本与关键点协同控制的人脸图像生成.计算机系统应用,2024,33(10):174–182. http://www.c-s-a.org.cn/1003-3254/9652.html

Facial Image Generation Based on Collaborative Control of Text and Key Points

LIU Yu-Tong, WANG Yi-Ding

(School of Information Science and Technology, North China University of Technology, Beijing 100144, China)

Abstract: Face image generation requires high realism and controllability. This study proposes an algorithm for face image generation that is jointly controlled by text and facial key points. The text constrains the generation of faces at a semantic level, while facial key points enable the model to control the generation of facial features, expressions, and details based on given facial information. The proposed algorithm improves the existing diffusion model and introduces additional components: text processing models (CM), keypoint control networks (KCN), and autoencoder networks (ACN). Specifically, the diffusion model is a noise inference algorithm based on the diffusion theory; CM is designed based on an attention mechanism to encode and store text information; KCN receives the location information of key points, enhancing the controllability of face generation; ACN alleviates the generation pressure of the diffusion model and reduces the time required to generate samples. In addition, to adapt to generating face images, this research constructs a dataset containing 30000 face images. In the proposed algorithm, given prerequisite text and a facial keypoint image, the model extracts feature information and keypoint information from the text, generating a highly realistic and controllable target face image. Compared with mainstream methods, the proposed algorithm improves the *FID* index by about 5%–23% and the *IS* index by about 3%–14%, which proves its superiority.

Key words: facial generation; diffusion model; generative artificial intelligence; text encoding; autoencoder

① 基金项目: 国家自然科学基金 (62276018)
收稿时间: 2024-04-06; 修改时间: 2024-05-06; 采用时间: 2024-05-14; csa 在线出版时间: 2024-08-28
CNKI 网络首发时间: 2024-08-30

人脸图像生成, 旨在给定先决条件 (如人脸特征的文本描述、面部关键点、掩膜分割图等) 的前提下, 生成符合约束条件的人脸图像. 目前, 研究者在该领域追求达到两个目标: 提高生成人脸的真实度和增强目标人脸的可控性. 其中, 真实度可以解释为生成人脸的质量和逼真感; 可控性体现在生成人脸对于条件的匹配程度.

与人脸生成相关的算法很多. 早期的人脸生成是基于 3D 模型的方法, 通过人脸模板和变形技术生成新的人脸图像. Blanz 等人提出了 3DMM 模型[1], 该模型分解人脸为形状 S 和纹理 T, 再通过主成分分析法, 将不同的人脸归纳到同一个脸部模型中, 求得协方差矩阵, 最终可以通过修改插值参数来进行人脸的生成. 3D 模型的缺点在于生成的人脸真实度较低和多样性差. 随着深度神经网络的产生, GAN[2]模型被提出并得到不断改进, 基于 GAN 的人脸生成图像的质量与精度有了极大的提高. GAN 网络可以分为基于噪声输入的传统 GAN 网络和基于条件输入的 cGAN 网络. Radford 等人提出了 DCGAN[3]网络, 将深度卷积神经网络与生成对抗网络相结合, 输入随机噪声以生成人脸图像. 该算法学习了随机噪声和人脸图像的映射关系, 但是生成图像的质量较差. Isola 等人在 2017 年发表了 pix2pix 模型[4], 该模型基于 cGAN 的思想, 不同于使用随机向量作为输入, 而是通过特定图像与随机噪声的结合映射到向量中作为生成器的输入, 实现一种图到图的生成算法. 这里的图到图可以实现跨领域生成, 例如从标签图生成照片、边缘图重建以及物体上色任务. 虽然 pix2pix 网络应用广泛, 但在图像的生成精度以及效果上较为欠缺, 尤其是在人脸生成这类细节突出的生成领域不能适用. Zhu 等人提出了 CycleGAN[5]模型, 利用循环生成网络学习如何将图像从源域转换到目标域, 实现了跨域人脸转换任务. Park 等人提出了 SPADE (spatially-adaptive denormalization)[6]网络, 该网络使用一种空间自适应规范化模块, 实现了从掩模图到真实图像的转换, 将语义信息进行转换来调制标准化层的特征, 并允许用户自主控制语义和风格. 总体来说, 基于 GAN 的方法目前已经能生成真实度较高的人脸, 但是在对生成结果的可控性方面仍有待提高.

近年来, 在生成领域出现了一种新的框架, 即扩散模型[7], 并且一经出现就成为生成领域的主流[8–10]. 与 3D 模型方法和基于 GAN 的方法相比, 扩散模型适用于固定目标训练的前提条件, 没有需要优化复杂目标损失 (例如, 对抗性损失) 或平衡多个目标 (例如, 码本损失和重建损失). 扩散模型的重要特性是能接收各种形式的数据为条件, 如文本[11–17], 分段掩码[18]以及草图[19,20]. 以训练简单性为优点, 近年来, 基于扩散理论的方法越来越流行. VQ-Diffusion[12]解决了以往的生成模型存在的单项偏差问题, 并且使用掩蔽机制避免了推理过程中误差的累计. DALLE-2[21]提出了一种两阶段的图像生成方法, 第 1 阶段是 CLIP[22]图像和文本条件的嵌入先验模型, 第 2 阶段是基于扩散模型的解码器, 可以进行图像嵌入, 最终可以生成清晰度较高的图像. Imagen 模型[17]由一个用于文本序列的编码器和一个用于生成高分辨率图像的级联扩散模型组成, 除了改进扩散模型的输入外, 还改进了原有的 U-Net[23]来提高模型的工作效率.

在本文中, 我们提出了一种由文本和关键点协同控制的基于扩散模型的人脸图像生成算法. 该算法可以接收多模态的条件激励, 在可驱动的模型之间建立横向连接. 不同模态之间通过编码技术建立联系, 使用动态扩散机制自适应预测空间影响函数. 我们的框架具有的简单性和灵活性使其能使用很少的修改就能进行模型内部结构的编辑和参数的优化. 在结果评价方面, 我们选取了多种不同的评价指标进行定性分析, 充分证明了本文算法的先进性和优越性.

1 算法设计及理论

本文算法的总框架如图 1 所示. 从图中可以看出, 算法的主干网络是一个多次迭代的 Diffusion 过程. 在初始阶段, 模型会接收脸部关键点和文本的多重条件输入, 其中, 脸部关键点会通过关键点控制网络 (keypoint control network, KCN), 该网络会通过位置编码和动态扩散的方式提取关键点中的位置矢量信息; 而文本条件会通过 CLIP 模型 (CM) 生成一段规律的编码序列; 为了加速扩散过程的拟合速度, 我们添加了自编码器网络 (autoencoder network, ACN), 该模块的作用是将高维特征信息映射到隐空间, 减少扩散运算的计算量. 最后, 三者结合共同生成一张采样图像, 经过 n 次的迭代, 得到最终输出的人脸图像.

1.1 扩散模型

扩散模型的灵感来源于非平衡热力学, 可以分为两个阶段: 前向扩散加噪阶段和逆向去噪阶段. 此过程

满足以下的马尔可夫条件:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}\cdot x_{t-1}, \beta_t \cdot I\right) \tag{1}$$

其中, $t \in \{1,\cdot,T\}$, T为总的扩散步骤, 噪声的增长水平方差$\beta_t \in (0,1)$服从高斯分布, I代表单位矩阵, $N(x;\mu,\delta)$代表正态分布. 根据马尔可夫过程的概率特性, 当t服从均匀分布时, 可以得到公式:

$$q(x_t|x_0) = N\left(x_t; \sqrt{\hat{\beta}}\cdot x_0, \left(1-\hat{\beta}_t\right)\cdot I\right) \tag{2}$$

其中, $\hat{\beta} = \prod_{i=1}^{t}(1-\beta_t)$. 计算$q(x_t|x_0)$时采用了一种被称为重参数化技巧的方式.

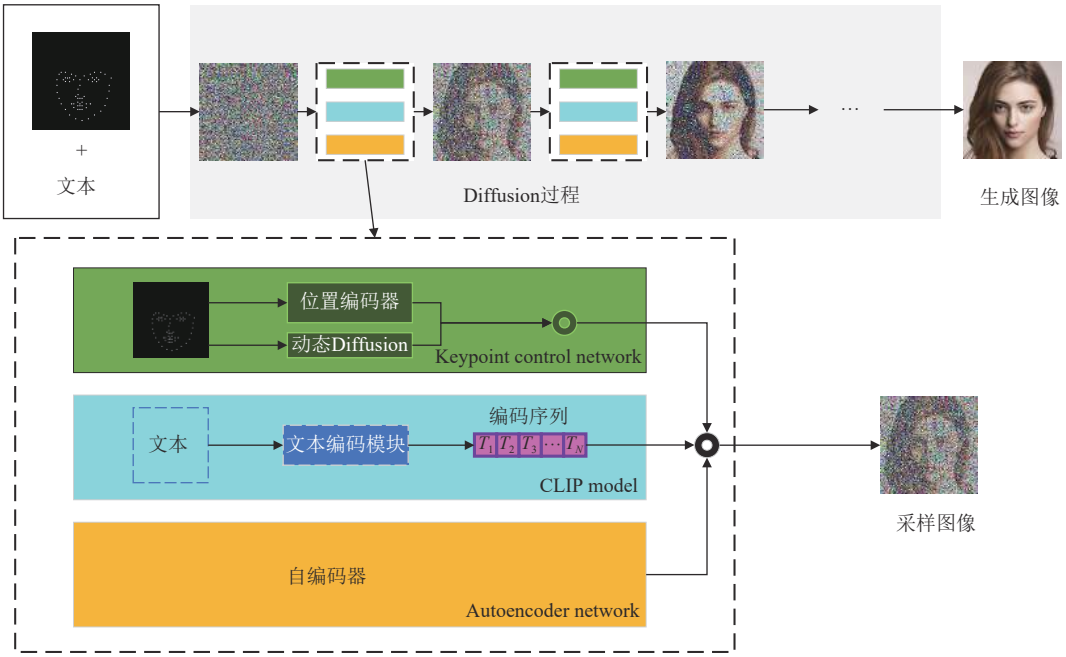

图 1 算法总框架

为了使服从正态分布的样本x标准化, 需要减去平均值μ并除以标准差σ, 进而得到服从标准正态分布的z, 由此可以通过标准逆变换推出x_t的递推公式为:

$$x_t = \sqrt{\hat{\beta}_t}\cdot x_0 + \sqrt{1-\hat{\beta}_t}\cdot z_t \tag{3}$$

其中, 噪声$z_t \sim N(0,I)$.

逆向去噪阶段的关键在于训练一个神经网络, 不断地调整网络的参数使反向过程的联合分布逐渐接近正向过程. 这里使用的方法是 Sohl-Dickstein 等人[9]所提出的最小化负对数似然的变分下界.

1.2 关键点控制网络 (KCN)

我们的算法采用了一种 ControlNet[24]框架注入人脸关键点信息. ControlNet 是一种常见的为神经网络注入附加条件的方式, 其原理通过锁定大型预训练模型的参数并复制其解码层, 保留了原始模型的拟合能力. 其复制原理如图 2 所示.

基于 ControlNet 的这种特性, 我们将原始模型视为学习了各种条件控制的强大主干网络. 在 ControlNet 中, 可训练的副本和原始锁定的模型通过零卷积层连接, 并且这些卷积层的权重初始值为 0, 在之后的训练中修改参数. ControlNet 能确保在训练开始时不会向大型扩散模型的深层特征添加有害噪声, 并保护可训练副本中的大型预训练主干网络免受此类噪声的破坏.

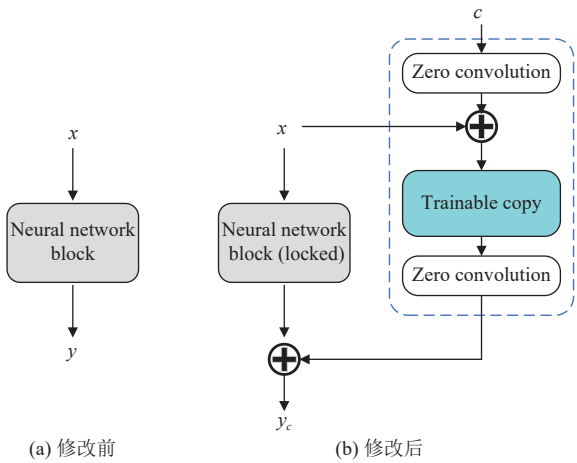

图 2 ControlNet 原理

本文使用的基于 stable diffusion 的 ControlNet 结构如图 3 所示.

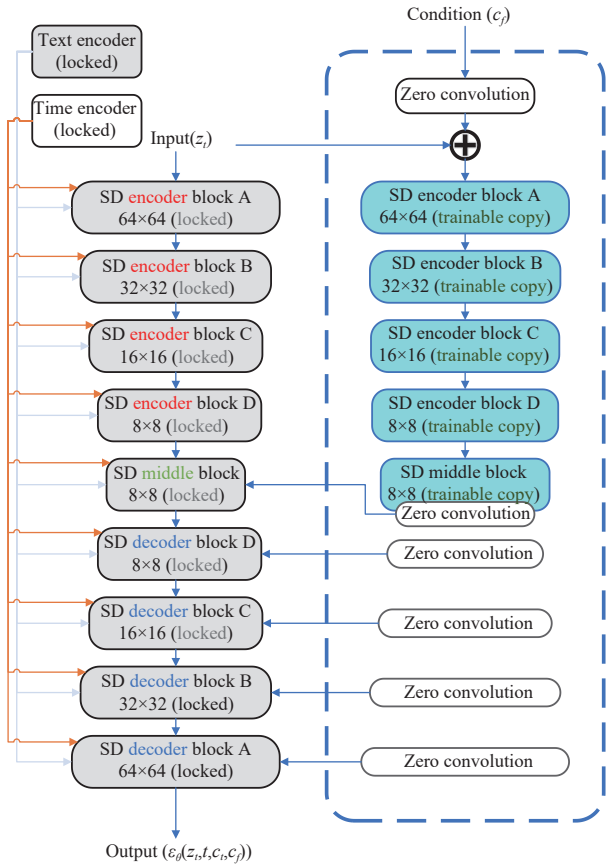

图 3 本文 ControlNet 结构

由图 3 可知, 我们使用的是一个基于 stable diffusion 大模型训练的具有特定任务条件的 ControlNet. 模型本质上是一个带有编解码器的 U-Net 结构. 在训练 ControlNet 的过程中, 我们控制原始 U-Net 的每个层级块, 由于训练不会修改原始权重, 所以只需要用到原始训练时间的 30% 左右就能训练好一个具有额外条件控制能力的 ControlNet.

1.3 文本处理模块 (CM)

本文算法使用的文本处理模块, 基于语言与图像对比预训练模型 (contrastive language-image pre-training, CLIP) 设计, 该模型是 OpenAI 团队建立的一个零映射 (zero-shot) 大规模自然语言和多模态理解任务预训练模型, CLIP 模型的主要任务是计算文本与图像的特征向量 (embedding) 的缩放成对余弦相似度, 以对比图像与文本的语义联系性. CLIP 的主要结构由一个用于编码文本的 Text-Transformer 和一个用于处理图像信息编码的 ViT (Vision-Transformer) 组成, 如图 4 所示.

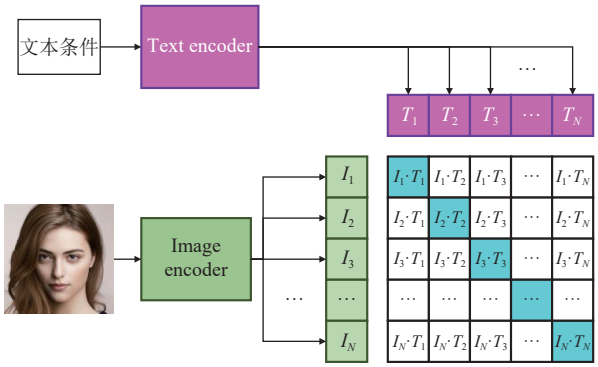

图 4 CLIP 文本处理结构

该结构首先将文本与图像的特征向量归一化, 然后求缩放成对余弦相似度s_{ij}, 如式 (6) 所示:

$$s_{ij} = \cos\theta_{ij} = \frac{I_{ij} \cdot T_{ij}^{\mathrm{T}} \cdot e^{t}}{|I_{ij}| \cdot |T_{ij}|},\ i,j = 1,2,3,\cdots,n \tag{4}$$

其中, I_{ij}表示图像的特征向量; T_{ij}表示文本的特征向量; t表示可以学习的温度参数, 该参数能用来缩放余弦相似度, 使n对匹配的文本图片对相似度最大; i, j表示向量编号.

CLIP 中 ViT 的结构如图 5 所示. 由图 5 可知, ViT 首先会将输入图像分成固定大小的 n×n 个子块, 然后将每一个图像块通过位置编码转换为一维的特征向量, 这样就能得到一个包含 n×n 个模相等向量的向量组, 这里的每一个向量可以记为一个 token.

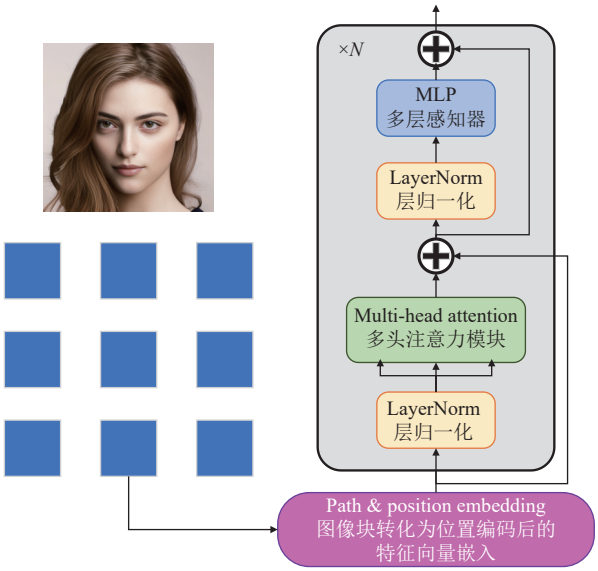

图 5 ViT 结构

1.4 自编码网络 (ACN)

由于扩散模型在推理过程中每一次采样都会生成

一张完整分辨率的样本图像, 当需要生成较大分辨率的人脸图像时, 推理时间会大大加长. 因此, 在扩散模型中加入编码器-解码器的结构能一定程度上缓解生成的压力. 在该条件下, 扩散模型只需要处理经过编码器处理过后的图像的隐空间信息. 本文采用的编码器-解码器模块结构如图 6 所示.

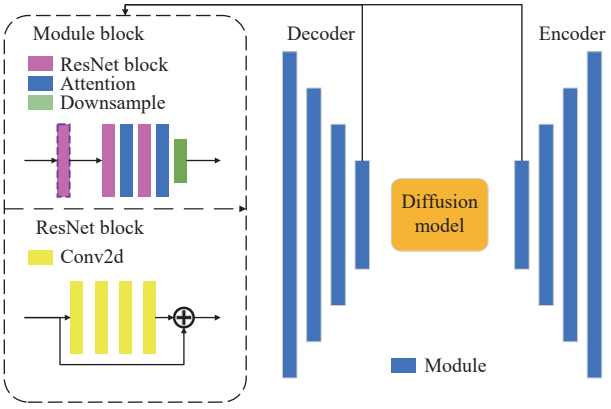

图 6 自编码器

由图 6 可知, 本文设计的编码器-解码器结构遵循传统的对称结构. 其中, 编码器有 4 个模块, 每个模块包含 2 个残差层、2 个注意力层和 1 个下采样层; 解码器也有 4 个模块, 每个模块包含 3 个残差层、2 个注意力层和 1 个上采样层.

在每个模块中, 残差层会将输入的特征图经过 4 个卷积层, 然后与原始特征图相加. 经过若干个残差层后, 神经网络的拟合能力能得到显著提高. 注意力层的作用是使模型对人脸图像的细节特征关注度提高.

2 实验结果与分析

2.1 数据集

本实验所使用到的数据集基于 CelebA (celebrities attributes) 制作, 该数据集由香港中文大学制作, 共包含超过 10 000 名社会知名人士的 20 万张以上的人脸图像. 在这些图像中, 每张人脸都划分了属性标签, 不仅包括脸部区域 (如胡须、眼镜), 也包括一些对于脸部的文字描述 (如微笑着的、面无表情的). 为了适应本文算法, 我们对原始的 CelebA 数据集进行了升级, 从中选取 30 000 张人脸图像, 同时对选取的人脸图像进行分辨率的统一, 均为 256×256. 之后, 人为地提取每一张人脸图像的脸部关键点, 用于基础模型的训练, 部分数据集样例如图 7 所示.

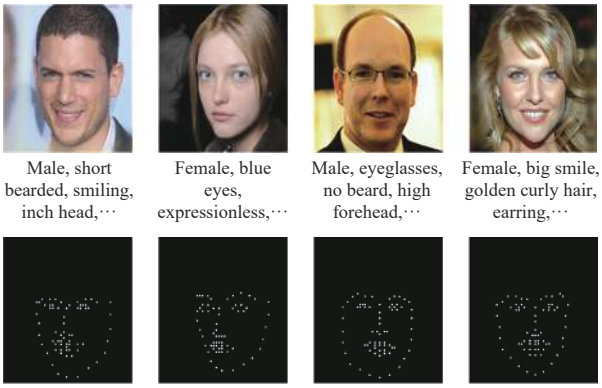

图 7 数据集 (部分)

2.2 实验步骤

本文的实验步骤可以分为 3 个阶段: 包括模型训练、人脸图像推理和生成图像质量评估.

(1) 模型训练阶段需要使用到的是图像数据集、通过 CLIP 方式得到的标注特征以及和每一张人脸图像的关键点信息, 然后运行训练脚本, 这里设置的超参数选取如下: 为了既方便查看采样效果, 同时节省训练时间, 设置样本分辨率 resolution 为 256×256, 最大训练步数 max_train_steps 为 15 000, 由于我们是在大模型上进行微调, 故学习率不宜过高, 设置 learning_rate 为 1E–5, 根据训练经验, 设置混合精度 mix_precision 为 fp16, 训练结果保存为 ckpt 检查点文件.

(2) 将训练的模型路径写入 inference 文件, 由于推理时需要对样本的质量进行分析和对比, 故此处增大了图像的分辨率, 设置为 512×512, 推理步数初步设定为 50, 在之后的实验中, 我们还针对推理步数的变化进行了消融实验. 采样方法为 DDIM[25], 该采样方法极大加快了采样速度; 为了便于之后对样本质量进行评估, 将样本批次设置为 1 250, 每批图像数量设置为 8, 即一共生成 10 000 张图像.

(3) 对生成图像进行定性分析, 选取合适的指标评估其真实度和可控性.

2.3 结果分析指标

为了检验本文算法在生成人脸图像的真实度和可控性两方面的性能, 选取以下指标进行评估.

FID (Frechet inception distance): 是一项评价图像真实度和逼真度的指标, 它计算的是生成图像与真实图像之间的特征空间距离, 其值越小表示生成图像与真实图像的特征相似度越高, *FID* 的计算公式如下:

$$FID = \left\|\mu_r - \mu_g\right\|^2 + \mathrm{Tr}(T_m) \tag{5}$$

其中, μ_r 表示真实图片的特征均值, μ_g 表示生成图片的特征均值, Tr(·)表示求迹操作.

IS (inception score): 用来衡量生成图像与真实图像的质量差异, 其值越高表示生成图像的质量越高, *IS* 的计算公式如下:

$$IS\ (G) = \mathrm{e}^{E_{x\sim p_g} D_{KL}(p(y|x))\|p(y)} \tag{6}$$

其中, $p(y|x)$表示基于生成图像的条件分布, $p(y)$是获得特定分布的边缘分布, D_{KL}表示相对熵.

2.4 对比试验

我们选取了多种不同算法进行生成结果的比较, 包括: stable diffusion, 该框架是基于文字生成图像的典范, 由文献[16]的成果逐渐演变而来; TediGAN, 源于文献[26], 是一种文本与图像一起操作, 允许进行外观属性编辑的一种 GAN 网络; collaborative diffusion, 源于文献[27], 是一种使不同扩散模型协同合作的架构, 该架构能使生成人脸保持与控制条件的高度一致性; 以及本文算法. 我们使不同算法生成 10 000 张图片, 分别计算其 *FID* 指标和 *IS* 指标, 得到的对比实验结果如表 1 所示.

表 1 对比试验结果

模型	*FID*	*IS*
Stable diffusion	45.69	5.89±0.19
TediGAN	52.61	6.17±0.23
Collaborative diffusion	37.12	6.53±0.15
Ours	**35.32**	**6.77±0.09**

从表 1 中可以看出: 对于 *FID* 指标, 我们算法的最终得分为 35.32, 相较 stable diffusion 的 45.69, 提升了约 30%; 对于得分比较接近的 collaborative diffusion (其得分为 37.12), 我们算法也有 5% 左右的提升. 说明了本算法生成图像与真实图像的特征相似度较高. 从 *IS* 指标分析, TediGAN 的波动程度较高, 其他 3 种模型的波动程度均较高, 分别为 0.19, 0.23, 0.15, 而本文算法只有 0.09, 并且最终的 *IS* 值可达到 6.86, 说明本算法的生成结果质量是较好的. 综合以上结果表明, 相较于其余方法, 本文方法的 *FID* 指标和 *IS* 指标更加优秀, 也就是本文模型生成的人脸图像真实度更强, 图像内容更加丰富, 质量也有所提高.

图 8 展现了不同模型的生成结果.

2.5 真实度探究实验

为了对本模型生成图像的真实度进行进一步的探究, 我们设置了如下实验: 以 *FID* 分数作为定性分析指标, 从原始数据库中分别提取 10 000, 15 000, 20 000, 30 000 张人脸图像, 并标记每一张图像的文本标签和关键点, 以该标签和关键点作为输入生成相同数量的虚拟人脸, 部分选取的真实数据和虚拟数据的样本如图 9 所示.

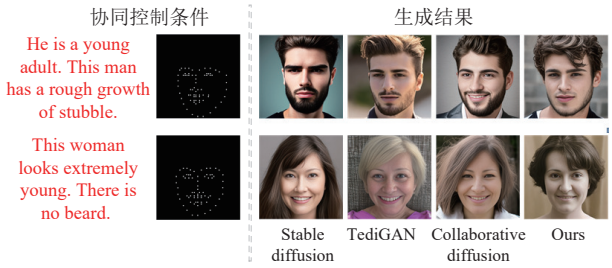

图 8 不同模型的生成结果展示

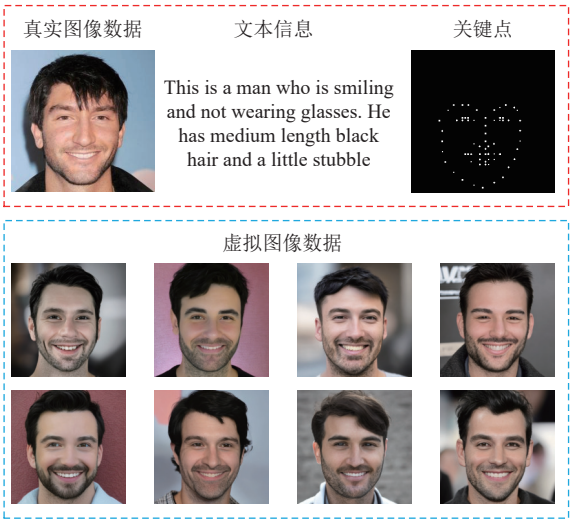

图 9 真实图像数据和虚拟图像数据

再分别计算真实图像数据和虚拟图像数据的 *FID* 指标, 由此评价生成人脸分布与真实分布的近似情况. 得到的实验数据记录如表 2.

表 2 可控性探究实验结果 (不同样本数)

数据量	*FID* (真实)	*FID* (虚拟)	匹配度 (%)
10 000	57.91	61.52	94.13
15 000	54.32	57.13	95.08
20 000	62.17	64.52	96.35
30 000	49.85	50.34	99.02

对于数据量为 10 000, 15 000, 20 000, 30 000 的情况, 可以看到虚拟数据分布与真实数据分布的匹配分别达到了 94.13%, 95.08%, 96.35%, 99.02%, 这充分说明了模型产生的虚拟人脸图像已经具有了很高的逼真度和真实性, 进一步验证了本文算法的优越性.

2.6 可控性探究实验

为了探究本模型能否有效提取关键点条件的位置信息, 我们绘制了生成人脸的关键点提取图像和条件图像的融合图, 该过程如图 10 所示. 首先, 将文本和关键点输入算法模型, 生成一张人脸, 然后提取人脸的关键点图像, 以原始关键点图像和虚拟人脸的关键点图像作为条件, 可绘制一张融合图. 每张融合图中, 白色区域表示两个对应的脸部关键点没有交集, 深色区域表示关键点互相重合, 黑色区域是无关信息区域.

图 10 条件图像和提取图像的融合

之后, 我们计算了分别生成 1 000、2 000、5 000 以及 10 000 张图像的关键点丢失率, 这里的关键点丢失率定义为: 目标关键点相对于条件关键点的偏移程度. 计算公式如下:

$$L_k = 100\% - \frac{I_k}{U_k} \tag{7}$$

其中, I_k表示关键点的重合区域, U_k表示两个点的并集区域. 由于我们的条件图像提供了 70 个脸部关键点, 对于每一张生成人脸的融合图, 计算其关键点丢失率, 在此基础上进行打分并统计不同分数段关键点的个数, 然后计算出不同分数段的百分比. 得到的实验数据记录如表 3 所示.

表 3 可控性探究实验结果 (不同样本数)

样本数	关键点占比 (%)			
	丢失率 (0, 5%]	丢失率 (5%, 10%]	丢失率 (10%, 20%]	丢失率 (20%, 1]
1 000	95.37	3.71	0.78	0.14
2 000	94.98	2.53	1.37	1.12
5 000	94.32	2.73	0.58	2.37
10 000	94.29	3.01	2.08	0.62

分析表 3 可以得出, 在低样本数和高样本数的情况下, 丢失率的差别并不大, 由于每一张人脸的生成具有随机性, 这是可以预见的结果. 从丢失率指标上来讲, 处于 0–5% 丢失率的关键点占比大约保持在 95%, 说明模型能较好地利用给出的关键点信息, 只有极少部分的提取关键点与条件关键点重合度低于 3%, 这是一个较为优秀的结果.

2.7 消融实验

为了对算法性能进行进一步地探索和优化, 设计了如下消融实验: 探究动态推理步数对实验结果的影响, 设置推理步数分别为 30、40、50、60; 探究模块之间的相互作用关系, 具体来说, 设置了扩散网络+CM, 扩散网络+CM+KCN, 扩散网络+CM+ACN, 扩散网络+CM+KCN+ACN 的对比消融实验. 实验结果如表 4 和表 5 所示.

表 4 消融实验结果 (不同推理步数)

Step	*FID*	*IS*
30	38.62	6.2±0.12
40	37.13	6.33±0.11
50	35.64	6.73±0.09
60	36.98	6.49±0.10

表 5 消融实验结果 (不同模块组合)

模型	*FID*	*IS*
Diffusion+CM	63.15	5.93±0.10
Diffusion+CM+ACN	53.17	6.15±0.09
Diffusion+CM+KCN	43.98	6.21±0.09
Diffusion+CM+KCN+ACN	35.64	6.73±0.09

根据表 4 可以看出, 在使用同样数据集的前提下, 采样步数对结果的影响较小, 比较 step 为 30, 40, 50, 60 时的 *FID* 指标, 大概在 50 步左右时能取得 *FID* 的最小值. 可以得出结论, 采样步数过小或者过大时均会影响生成图片的质量.

根据表 5 的实验数据可以分析得出, 在原始扩散模型中加入 ACN 和 KCN 结构能有效地提高生成图像的质量. 在加入 ACN 结构的情况下, *FID* 指标由 63.15 下降到了 53.17, 可见 ACN 能有效提取图像数据集的概率分布信息; 在加入 KCN 结构的情况下, *IS* 指标从 5.93 上升到了 6.21, 并且波动幅度下降了 10%, 可见 KCN 结构对增强采样结果的稳定性具有较大优势, 同时对维持生成的可控性提供帮助; 在同时加入 ACN 结构和 KCN 结构的前提下, *FID* 指标相较原始模型下降了 27.51, *IS* 指标上升了 0.80, 这比单独加入 ACN 结构或 KCN 结构的提升均要明显, 可见 ACN 和 KCN 具有优势互补的影响, 二者结合能达到最优的效果.

3 结论与展望

本文提出了一种由文本和脸部关键点协同控制的人脸图像生成算法, 该算法以扩散模型为基础, 加入

CM 结构对本文信息进行编码, KCN 结构提取关键点位置信息, ACN 结构加快模型的采样速度. 在对比实验中, 我们与 3 种主流方法进行了对比, 通过结论证明本算法生成的人脸图像具有较高的清晰度和逼真性; 我们还设计了真实度与可控性两方面的探究实验, 深入探讨各模块与参数在算法框架中的潜在作用; 通过消融实验可以得出结论: 推理步数一定程度上能影响生成效果, 一般不宜设置过多与过少; KCN 结构可以有效提取位置矢量信息, 大大加强了对生成人脸的可控性, 同时 ACN 结构对图像的隐空间信息进行编解码, 缓解了扩散过程的生产压力, 在生成同等分辨率图像的条件下大大缩短了时间. 我们针对实验的不确定性和潜在限制进行了进一步讨论, 也到本文算法在某些方面具有的局限性, 例如, 由于数据集覆盖的不同年龄范围的人群不是均衡的, 算法生成青年人脸的效果普遍优于生成幼年人脸和老年人脸的效果, 在这一方面, 我们也将会继续深入研究. 综上, 本文算法在提高人脸生成模型的真实度和可控性方面, 具有一定的优越性, 对于行业内其他的多模态生成任务来说, 具有一定的借鉴意义.

参考文献

1 Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles: ACM, 1999. 187–194.

2 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144. [doi: 10.1145/3422622]

3 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. Proceedings of the 4th International Conference on Learning Representations. San Juan, 2016.

4 Isola P, Zhu JY, Zhou TH, *et al*. Image-to-image translation with conditional adversarial networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5967–5976.

5 Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2242–2251.

6 Park T, Liu MY, Wang TC, *et al*. Semantic image synthesis with spatially-adaptive normalization. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2332–2341.

7 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 574.

8 Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 672.

9 Sohl-Dickstein J, Weiss EA, Maheswaranathan N, *et al*. Deep unsupervised learning using nonequilibrium thermodynamics. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 2256–2265.

10 Song Y, Sohl-Dickstein J, Kingma DP, *et al*. Score-based generative modeling through stochastic differential equations. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

11 Avrahami O, Lischinski D, Fried O. Blended diffusion for text-driven editing of natural images. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 18187–18197.

12 Gu SY, Chen D, Bao JM, *et al*. Vector quantized diffusion model for text-to-image synthesis. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 10686–10696.

13 Kawar B, Zada S, Lang O, *et al*. Imagic: Text-based real image editing with diffusion models. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 6007–6017.

14 Kim G, Kwon T, Ye JC. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 2416–2425.

15 Nichol AQ, Dhariwal P, Ramesh A, *et al*. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022. 16784–16804.

16 Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 10674–10685.

17 Saharia C, Chan W, Saxena S, *et al*. Photorealistic text-to-image diffusion models with deep language understanding. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2643.

18 Wang WL, Bao JM, Zhou WG, *et al*. Semantic image synthesis via diffusion models. arXiv:2207.00050, 2022.

19 Wang TF, Zhang T, Zhang B, *et al*. Pretraining is all you need for image-to-image translation. arXiv:2205.12952, 2022.

20 Cheng SI, Chen YJ, Chiu WC, *et al*. Adaptively-realistic image generation from stroke and sketch with diffusion model. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2023. 4043–4051.

21 Ramesh A, Dhariwal P, Nichol A, *et al*. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125, 2022.

22 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763.

23 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

24 Zhang LM, Rao AY, Agrawala M. Adding conditional control to text-to-image diffusion models. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3813–3824.

25 Song JM, Meng CL, Ermon S. Denoising diffusion implicit models. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

26 Xia WH, Yang YJ, Xue JH, *et al*. TediGAN: Text-guided diverse face image generation and manipulation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2256–2265.

27 Huang ZQ, Chan KCK, Jiang YM, *et al*. Collaborative diffusion for multi-modal face generation and editing. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 6080–6090.

(校对责编: 孙君艳)