doi:10.3969/j.issn.1001-4616.2025.02.011

一种适用于政务区块链的跨模态人脸生成模型

崔思颖[1],谭志杰[1],袁　想[1],李伟平[1],莫　同[1],乔秀全[2],吴中海[1]

(1.北京大学软件与微电子学院,北京 102600)
(2.北京邮电大学网络与交换技术国家重点实验室,北京 100876)

[**摘要**]　如今,区块链技术被应用到包含电子证照、人脸图像等政府数据共享领域,但当前的大型区块链系统普遍面临低带宽和高存储成本的问题. 本文提出了一种适用于政务区块链的跨模态人脸生成模型,将人脸图像转换为文本模态存储在链上,用户可使用文本与掩膜生成指定人的人脸图像. 首先利用多任务学习方法训练基于 ResNet-18 网络结构的人脸分类器,将人脸图像转换为身份代号文本存储在链上. 然后设计了区域感知码本和基于 Transformer 结构的混合专家采样器,采样器采用扩散模型的方法从码本中采样索引,采样结果由一个可学习的解码器转换成细粒度的人脸图像. 在进行数据增强后的 Casia Face V5 数据集上的实验表明,模型在人脸分类任务中准确率可达 95%以上,压缩效果达到了传统图像压缩方法 1/10 000 的持久化时间与 1/200 的文件大小,与其他先进人脸图像生成方法相比,此模型可以可控地生成高保真度的指定人的人脸图像,并以 1/20 的参数量达到与大型预训练模型相近的人脸生成效果.

[**关键词**]　区块链,跨模态人脸生成,可控图像生成,扩散模型,人脸识别
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2025)02-0102-10

Cross-Modal Face Generation Model for Government Blockchain

Cui Siying[1],Tan Zhijie[1],Yuan Xiang[1],Li Weiping[1],Mo Tong[1],Qiao Xiuquan[2],Wu Zhonghai[1]

(1.School of Software and Microelectronics,Peking University,Beijing 102600,China)
(2.State Key Laboratory of Networking and Switching Technology,Beijing University of Posts and Telecommunications,Beijing 100876,China)

Abstract:Blockchain technology is currently used in government data sharing,but faces challenges such as limited bandwidth and high storage costs. To address this,the study proposed a cross-modal face generation model for the government blockchain. This model converted face images into text modals and stored them on the chain,allowing users to generate face images of specific individuals using text and masks.To achieve this,the study trained a face classifier based on the ResNet-18 network structure using a multi-task learning method. The resulting identity code text is then stored on the blockchain. Additionally,the study constructed region-aware codebooks and designed a diffusion-based transformer sampler with mixture-of-experts. This sampler converts indexed from the codebooks into fine-grained face images using a learnable decoder. The experiments on the enhanced Casia Face V5 dataset demonstrated that the model achieved a face classification accuracy rate of 95%. Furthermore,it offered a persistence time of 1/10 000 and a file size of 1/200 compared to traditional image compression methods. Compared to other advanced face image generation methods,this model can generate high-fidelity face images of specific individuals while requiring only 1/20 of the parameters of large pre-trained models.

Key words:blockchain,cross-modal face generation,controllable image generation,diffusion model,face recognition

区块链(Blockchain)是一种去中心化的分布式计算技术,该技术可保证数据在区块链系统中不被篡改,得到安全存储. 在国家治理领域,区块链技术的应用体现在“政务数据上链”这一新兴概念,即将区块链技术应用到包含电子证照、人脸图像等政府数据共享领域,以满足行政系统内部保障数据安全与全过程留痕的工作要求,促进跨行政部门间的数据流通[1].

收稿日期:2023-06-28.
基金项目:辽宁省科学技术计划揭榜挂帅项目(2021JH1/10400010).
通讯作者:莫同,博士,副教授,研究方向:大数据分析及挖掘、知识图谱等. E-mail:motong@ss.pku.edu.cn

当前的区块链技术存在“不可能三角”,即无法同时达到可扩展、去中心化、安全性[2]. 政务区块链技术是一种追求去中心化与安全的技术组合,它的每一个节点都下载和储存所有的数据包,带来了巨大的存储空间损耗和校验成本,因此基于政务区块链的人脸图像的传输面临着带宽与存储的严格限制. 图像模态的数据具有数据量大、空间有序、冗余度高、相关性强等特点,为了提高传输效率,需要对其采用高压缩比进行压缩处理. 传统的图像编码技术(如 JPEG、WEBP 压缩编码算法)采用变换、量化及熵编码等方法进行图像压缩,以损失保真度为代价达到更高的压缩比,难以在存储成本、传输成本、保密程度上满足要求[3].

近年来文本与图像之间的跨模态转换受到越来越多的关注. 将人脸图像转换为文本模态进行传输不仅可以大幅降低数据维度、保证高压缩比,亦能起到加密作用[4]. 由人脸图像生成包含身份信息的文本需要对人脸进行分类,ResNet 网络结构因其良好的性能而被广泛应用于人脸分类的任务[5]. 由文本生成人脸图像的相关研究较少,因为人脸具有较多复杂的细粒度特征. 在人脸图像生成任务中,细粒度生成是指生成具有细微特征、清晰度高、质量优秀的人脸图像,包括面部表情、皮肤纹理、头发细节等. 这样的生成要求能够捕捉到人脸的微妙差异,使生成的人脸图像与真实人脸更加相似、更加逼真. 近年来,StyleGAN[6-7]在人脸生成的细粒度上取得了优秀的性能,许多模型[8-9]将其作为主干合成人脸,但由于噪声主导 StyleGAN 的生成结果,难以精确控制其生成指定身份的人脸. ControlNet[10]是一种通过添加额外条件输入来控制图像生成的神经网络结构,能精确引导生成结果,但模型较大,对网络资源有较高要求.

综合考虑政务区块链网络的带宽与容量的限制以及人脸图像的保真度等因素,本文提出了一种适用于政务区块链的跨模态人脸生成模型. 为了尽可能压缩图像文件的体积,节省带宽,将同一个人的人脸图像转换为唯一的身份代号文本进行传输和存储,不仅具有强区分性,也避免了长文本描述存在的计算复杂与安全风险的问题. 在进行人脸重建时,基于人脸图像掩码和身份代号文本生成细粒度的人脸图像. 考虑到人脸的高度多样性,将码本的概念[11]引入到框架中,用来学习人脸图像的离散神经表示. 训练采用基于扩散模型的 Transformer[12-14]采样器从码本中选择适当的表示,在空间域中将其重新排列,从而生成人脸图像. 在采样器中,引入了混合专家机制,由多个专家头来预测人脸图像不同区域的索引.

本文的具体贡献如下:

(1)采用 ResNet-18 结构与多任务学习机制进行人脸识别,将人脸图像数据跨模态转换为身份代号文本进行区块链上的存储和传输,极大降低了资源耗费;

(2)提出一种跨模态可控人脸生成模型,根据身份代号文本即可将掩膜复原为指定人的人脸图像,保证了人脸具有的身份特异性;

(3)设计了区域感知码本,引入混合专家的概念,提出了一种基于扩散模型的 Transformer 采样器,实现人脸图像的细粒度合成;

(4)人脸特征被离散化后存储于模型的码本表示之中,避免了使用传统图像编码方法存在的人脸数据隐私容易泄露的问题.

1 相关工作

自生成对抗网络(GAN)出现以来,图像生成取得了快速进展[15],被广泛用于人脸图像处理的各类任务. VAE[16]是一种图像生成方法,它将图像嵌入到潜在分布中,并从先验分布中采样向量来合成图像. 基于 VAE 的结构进行图像生成可以改善图像的质量[17].

从区块链上的应用和交互的角度看,除了生成高保真的人脸图像,还有使用文本控制合成人脸图像的需求. 由文本合成人脸图像的相关研究较少,因为人脸具有较多复杂的细粒度特征,且具有身份信息. 基于 GAN 提出的 Pix2PixHD[18]实现了图生图的应用,可将图像掩膜丰富为实际图像,极大提升了 AI 生成的实用性. Chen 等[19]利用文本编码器、图像解码器和 3 个鉴别器合成 256×256×3 的人脸图像,但是这些方法生成的人脸保真度较低、五官不清晰. 许多模型[8-9]将 StyleGAN[6-7]作为主干合成高保真度人脸,但由于噪声主导 StyleGAN 的生成结果,难以精确控制其生成指定身份的人脸. 最近,ControlNet[10]在可控生成领域实现一大进步,它可以通过输入简单的掩膜图生成满足条件的人脸,但模型体积巨大,不适用于政务区块链应用场景.

2　适用于政务区块链的跨模态人脸生成模型

2.1　人脸图像转换为文本

为了达到高压缩比,同时保证由人脸图像转换得到具有身份信息的文本,先使用 ResNet 进行人脸分类,再将人脸图像转换为身份代号文本. 在人脸类别数过多时,ResNet 的性能可能会下降,往往需要更长的训练时间和更大的数据集来避免过拟合,甚至可能出现梯度消失或梯度爆炸等问题. 为了应对这个问题,使用 K-Means 聚类算法将人脸进行 PCA 降维后划分为多个簇,在 ResNet-18 网络最后一层设置两个全连接层,采用共享特征[20]同时完成分簇和分类的多任务学习.

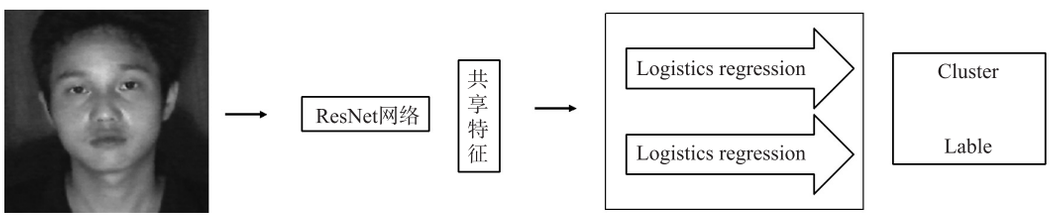


图 1　基于多任务学习的 ResNet 人脸识别方法框图

Fig. 1　Block diagram of ResNet face recognition method based on multi-task learning

定义 N 维向量 $\boldsymbol{Y}$ 和 $\boldsymbol{C}$,其中 $\boldsymbol{Y}$ 每个元素 y_i 表示第 i 个人脸图像样本的身份编号,$\boldsymbol{C}$ 的每个元素 c_i 表示第 i 个人所属簇号. 为了在训练过程中优化分类和聚类的结果,定义了一个联合损失函数:

$$\mathcal{L}_{f2t}=\omega_1\cdot CE(\boldsymbol{Y},Y')+\omega_2\cdot CE(\boldsymbol{C},C'), \tag{1}$$

式中,Y'和 C'分别表示模型预测的身份类号和簇号,ω_1 和 ω_2 分别是分类任务与聚类任务的权重,$CE(\boldsymbol{Y},Y')$和 $CE(\boldsymbol{C},C')$分别为人脸分类任务和分簇任务的交叉熵损失函数,计算公式如下:

$$CE(y,\hat{y})=-\sum_{i=1}^{n} y_i\log(\hat{y}_i). \tag{2}$$

在训练过程中,使用反向传播算法来更新模型参数,使得联合损失函数最小化,最终实现人脸图像到身份代号文本的预测.

2.2　文本转换为人脸图像

本研究的目标是根据身份代号文本生成人脸图像,给定人脸掩膜编码、身份代号文本,输出应当是文本指定的人脸图像,整个算法流程如图 2 所示. 将人脸掩膜编码信息转换为区域属性嵌入,根据身份代号文本合成最终图像. 本研究建立了一个区域感知码本来表征人脸图像各个区域的纹理,如图 3 所示,并提出了一种混合专家机制的采样器对码本索引进行采样,其中不同区域的特征由不同的专家头接收来预测

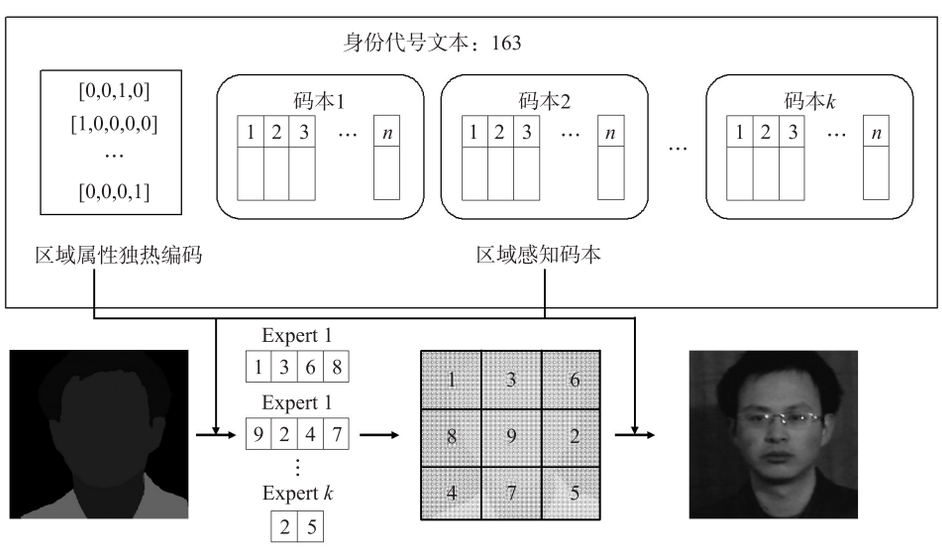


图 2　身份代号文本生成人脸图像的算法流程示意图

Fig. 2　Overview of the text-to-face algorithm

索引,进一步提高了生成图像的质量.

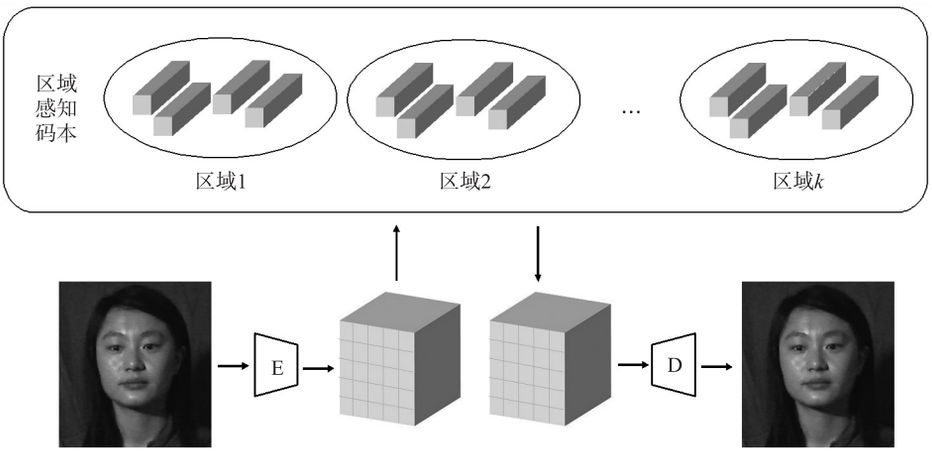


图 3　VQVAE 与区域感知码本图解

Fig. 3　Illustration of VQVAE and section-aware codebooks

2.2.1　VQVAE 与区域感知码本

考虑到人脸图像的高度多样性,引入码本的概念,它被广泛使用于基于矢量量化变分自动编码器(VQVAE)[11]的图像生成方法,能够学习到图像的离散向量表示. VQVAE 的目标是通过重建图像来学习离散码本. VQVAE 由一个编码器 E、一个解码器 D 和一个可学习的码本 $Z=\{z_k \mid z_k \in \mathbf{R}^{1\times 1\times c_z}\}_{k=1}^{K}$ 组成. 首先将图像 $I \in \mathbf{R}^{H\times W\times 3}$ 输入要训练的编码器 E,提取其中的连续神经表示 $\hat{z}$,即 $\hat{z}=E(I) \in \mathbf{R}^{H/16\times W/16\times c_z}$. 然后为 $\hat{z}$ 建立一个码本 Z,采用量化器 Quant 将连续的 $\hat{z}$ 离散化,该操作定义如下:

$$z_q=\mathrm{Quant}(\hat{z}):=\arg\min_{z_k\in Z}\|\hat{z}_{ij}-z_k\| \in \mathbf{R}^{h\times w\times c_z}, \tag{3}$$

将大小为 $H/16\times W/16$ 的图像量化表示输入解码器 D 来重建图像 $\hat{I}=D(z_q)$. 编码器、解码器以及码本通过感知损失、鉴别器损失以及下面的损失函数来进行训练:

$$\mathcal{L}=\|I-\hat{I}\|+\|sg(\hat{z})-z_q\|_2^2+\|sg(z_q)-\hat{z}\|_2^2. \tag{4}$$

式中,$sg(\cdot)$表示停止梯度的操作.

由于人脸图像中不同区域具有不同的纹理材质,进一步设计了区域感知码本. 在原始尺度的人脸图像上,具有不同外观的区域(人脸、头发、衣领、背景等)在下采样的尺度上可能看起来相似,如果为所有区域构建一个单一的码本,则会导致歧义问题. 因此,分别为不同的区域属性构建不同的码本,每个码本存储一种区域纹理的向量表示. 根据区域属性对编码器提取的特征进行分类,将不同区域的特征输入不同的码本中进行离散化.

2.2.2　混合专家采样器

为了从学习的码本中采样图像,文献[21]使用自回归模型来预测码本索引的排序. 在近年提出的 VQGAN[17]中,采用了 Transformer[22]的结构,因为其能够捕获码本索引之间的长期相关性(在 Transformer 中,码本索引被视作“tokens”). 最近,一些工作[12-14,23]基于以下两个优点提出使用扩散模型来取代自回归模型:(1)基于全局和双向上下文预测索引,从而获得更连贯的采样图像;(2)使用并行预测指数来提升采样速度. 具体而言,在基于扩散模型的 Transformer 中,从完全由噪声掩膜的索引张量 k_0 开始,Transformer 通过 T 步采样来预测最终的索引张量 k_T. 第 t 步的索引 k_t 按照以下分布进行采样得到:

$$k_t \sim q_\theta(k_t \mid k_{t-1}), \tag{5}$$

式中,θ 是 Transformer 的参数. 每一步采样后,噪声索引会被随机部分替换为新采样的索引.

为了与区域感知码本相结合,将基于扩散模型的 Transformer 也调整为区域感知的. 一个简单的想法是为不同的区域训练专门的采样器,这种方法的缺点是:(1)整幅图像的上下文信息对于码本索引的采样至关重要,而训练单一区域的采样器无法让网络获知这样完整的信息;(2)多个 Transformer 需要高 GPU 资源配置,且模型过大,不适用于政务区块链的存储和分发.

因此,将专家混合的想法[24]引入到基于扩散模型的 Transformer 中. 混合专家采样器的输入由三部分

组成:(1)码本索引 T_{code};(2)人脸图像分割掩码 T_{seg};(3)身份代号文本 T_{label}. 利用 Transformer 的多头注意力机制来进行计算:

$$f=MHA(\boldsymbol{Emb}_{\text{code}}(T_{\text{code}})+\boldsymbol{Emb}_{\text{seg}}(T_{\text{seg}})+Enc(T_{\text{label}})). \tag{6}$$

式中,$\boldsymbol{Emb}_{\text{code}}$和 $\boldsymbol{Emb}_{\text{seg}}$是可学习的嵌入向量,$Enc(T_{\text{label}})$表示对身份代号进行独热编码.

利用多头注意力机制提取的特征 f 被传递到不同的专家头,基于 T_{seg} 提供的区域属性信息,每个专家预测单个区域的索引. 索引的预测可视为分类任务,类的数量就是码本的大小. 最终的码本索引由所有专家的输出组成.

在训练期间,码本索引 T_{code} 由 VQVAE 获得. 在采样期间,T_{code} 由噪声索引掩膜初始化,以新采样的索引迭代填充完全.

3　实验结果及分析

3.1　实验数据

本文采用的数据集为 CASIA-FaceV5 亚洲人脸公开数据集,该数据集包含 500 人的上半身图片,每个人 5 张图片,共 2 500 张图片,图片原始大小为 640 * 480,格式为 BMP 无损文件.

为了使模型专注于人脸的特征捕捉,使用 Github 上的主流人脸识别工具包 face_recognition 来实现人脸位置检测,将上半身图像裁剪为 256 * 256 的人脸图像. 此外,为了增强模型的泛化能力,对数据集做了数据增强处理从而扩增训练集样本. 对人脸图片进行翻转、去雾、改变亮度、改变饱和度、改变对比度的处理,从而将每个人的人脸图片扩增到 25 张,图片总数扩增为 12 500 张,图像处理结果如图 4 所示.



图 4　图像预处理结果

Fig. 4　Image preproceeding sessing results

3.2　实验环境

使用的操作系统是 Ubuntu 20.04,CPU 是 Intel Xeon(R)E5-2603v4@ 1.7GHz,内存大小为 64 GB,使用 1 个 NVIDIA GeForce RTX 3090 GPU 来进行模型的训练与测试.

专家混合采样器需要人脸图像分割掩码 T_{seg} 和身份代号文本 T_{label} 作为输入,T_{seg} 是使用文献[25]的自监督结构敏感损失函数及卷积网络来生成,T_{label} 是每个人唯一的编号标识,对于数据集中的 500 个人脸,其编号标识由“000”递增到“499”,在实际应用于区块链上时,指的则是人的身份证号文本.

在实验一中,优化器为 Adam,学习率设置为 1×10^{-3},批次大小为 5,训练 10 个时期. 每张人脸图像经过 PCA 方法降为 120 维的向量,之后聚类成 100 个簇. 分类和分簇任务权重和分别设置为 0.8 和 0.2. 为了加快模型收敛速度,利用 Fine-tuning 方法对 ResNet-18 的预训练模型进行微调. 使用 PIL 9.2.0 for Python 进行图像压缩,使用 PlatONE 2 来模拟区块链系统.

在实验二中,优化器为 Adam,学习率设置为 1×10^{-4}. 先训练码本 100 个时期,批次大小为 5. 训练的损失函数为交叉熵损失. 训练采样器 300 个时期,批次大小为 5.

3.3　消融实验

3.3.1　ResNet 结构与分类学习

对于模型人脸分类的效果,在训练集和测试集中随意抽取 100 张人脸图片进行预测,结果如表 1 所

示. 相比于单任务分类学习,可以看出多任务分簇学习的加入使准确率有所提高,这说明本文所引入的多任务学习机制十分有效. ResNet50+多任务分类学习网络测试集准确率比 ResNet18+多任务学习仅高出 0.44%,而 ResNet18 训练时间较短,有效地降低了模型对存储资源的消耗,测试集准确率达到了 95.29%,说明在多任务机制的辅助下,ResNet18 足以高效完成一定规模的人脸识别任务.

表 1　人脸预测效果

Table 1　Face prediction results

模型	测试集准确率	训练集准确率	模型大小/KB
ResNet50+单任务分类学习	0.944	0.975	102 097
ResNet50+多任务分簇和分类学习	0.957	0.989	102 489
ResNet18+单任务分类学习	0.928	0.949	48 002
ResNet18+多任务分簇和分类学习	0.953	0.986	48 086

3.3.2　区域感知码本与专家混合

为了评估区域感知码本和专家混合设计的有效性,训练了单个存储图像上所有信息的码本以及一个采样器从该码本上采样生成图像. 生成结果的案例如图 5 所示,没有专家和区域感知码本混合的采样器,模型无法生成清晰的皮肤、头发、衣服等纹理,而区域感知码本的设计使采样器更好地区分图像的各个部分,在生成图像的细节上有更好的表现.

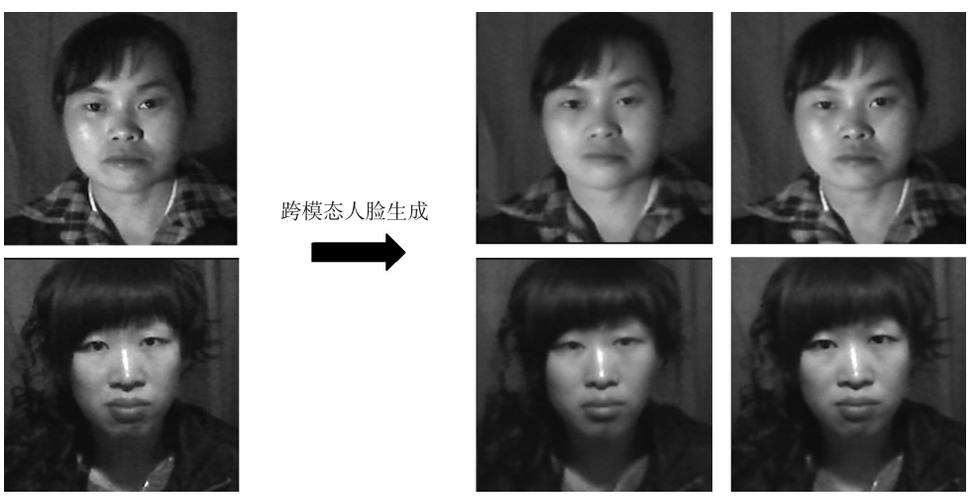


图 5　由单个码本生成人脸图像的案例

Fig. 5　Cases for generating a face image from a single codebook

3.4　敏感性分析

在人脸图像转换为文本的过程中,聚类任务与分类任务的损失权重 ω_1、ω_2 将对模型识别人脸的准确率造成影响,因此本研究对这两个参数进行了敏感性分析. ω_1 和 ω_2 的和为 1.0,测试了 ω_1 为 0.1,0.2,…,0.9 时 ResNet-18 对于测试集人脸图像识别的准确率,结果如图 6 所示. 从结果可以观察到,ω_1 为 0.7、ω_2 为 0.3 时,分类和分簇任务的准确率都较高,分类任务的准确率达到所有权重组合中的最高,此时,分簇任务可以很好地促进模型对于分类任务的学习.

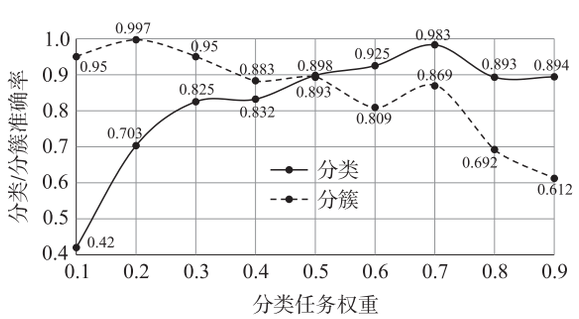


图 6　不同任务权重下的分类/分簇准确率

Fig. 6　Classification/clustering accuracy with different task weights

3.5　实验结果

3.5.1　人脸图像转换为文本

对于压缩效果,研究使用压缩后的文件大小与 TPT(Total Persistence Time)作为度量标准. TPT 指的是所有文件传输到区块链并成功在链上持久化后的总耗时. 开始时间是发送命令以开始上载图像的时间,结束时间是包含图像数据的最后一个块的生成时间.

将本文的跨模态人脸图像压缩方法与最流行的图像有损压缩方法(.jpg/.webp)进行了比较. 表 2 中的 W.75/J.75 方法意味着数据集中的原始图像(.bmp)使用默认存储质量 75 保存为.webp/.jpg 格式. 表 2 中的 W.1/J.1 表示数据集中的原始图像(.bmp)使用存储质量 1 保存为.webp/.jpg. 默认质量参数 75 可以很好地保存图像细节,并以 30~60 倍的压缩比节省存储空间. 当质量参数为 1 时,图像会丢失很多细节,并且会出现明显的模糊.

表 2　人脸图像压缩效果

Table 2　Face image compression results

压缩方法	压缩后文件大小/KB	TPT/s	压缩方法	压缩后文件大小/KB	TPT/s
无	96 026.37	609.966	W.75	1 493.67	171.432
J.75	3 351.88	172.934	W.1	395.86	125.569
J.1	869.87	146.531	本文	5.91	0.027

从表 2 中很容易看出,与图像压缩方法相比,本文提出的框架在持久化速度和文件大小方面具有无可比拟的优势. 此方法可以将大量的图像模态的数据映射为一个 2.44 KB 大小的文本文件与一张 3.47 KB 大小的掩膜图片,以存储在一个区块链块中. 相对于原始的人脸图像,本文方法达到的压缩后图像大小优于.jpg/.webp 压缩方法的 10^3 倍,在文件大小和持久化时间上分别优化为 W.1 的 1/200 和 1/10 000,这将大大降低整个区块链系统的通信和同步压力. 实验二将证明此同时不会造成人脸图像细节的严重损失.

对人脸图像转换为文本的程序运行时间进行了 10 次测试,并计算平均值,得到每张 256×256 大小的人脸图像的平均转换时间为 0.544 s,能够很好地满足政务应用场景中用户对于压缩算法响应时间的需求.

3.5.2　文本转换为人脸图像

从数据集中选取 500 张不同人的人脸图片进行重建,并与以下先进模型进行人脸图像生成效果上的比较,定性比较结果如图 7 所示.

(1)Pix2PixHD[18]:语义分割图引导图像合成的条件 GAN. 这里使用人脸图像分割掩码作为输入.

(2)DINO[26]:通过捕获语义分割图与真实图像两个域之间的共享信息,实现图像翻译功能的条件 GAN. 本文使用人脸图像分割掩码和真实人脸图像作为输入的两个域.

(3)GPEN[27]:一种 GAN 先验嵌入式网络,将训练好的 GAN 嵌入到 U 型 DNN 中作为先验编码器,再用一组严重退化的低质图像对其进行微调,可用于将人脸掩膜转换为人脸图像. 使用真实图像训练 GAN 嵌入结构,用人脸图像分割掩码作为退化图像输入.

(4)ControlNet[10]:一种针对预训练的大型扩散模型的增强控制网络,可以端到端地在小数据集上实现扩散模型的训练,并支持分割映射作为条件输入. 与本文相同,使用人脸图像分割掩码和身份代号文本作为输入来训练扩散模型,从而进行公平比较.

由图 7 可以看出,Pix2PixHD 和 DINO 生成的人脸在不同程度上出现了五官模糊的问题,无法生成高保真的细粒度人脸图像;GPEN 模型在一定程度上可以改善人脸的生成精度,但是出现了眼镜的伪影,模型无法对不同的人脸进行区分;ControlNet 与本文模型因为由人脸分割掩膜图像、身份代号文本驱动,可以

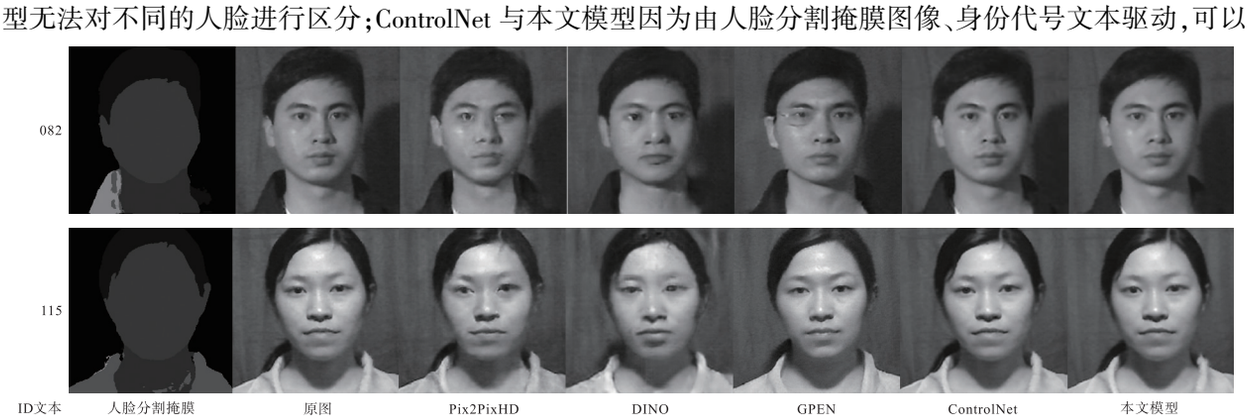


图 7　人脸生成模型性能定性比较

Fig. 7　Qualitative comparison of face generation models

很好地确定五官特征,有效地避免生成偏离原身份的人脸. 然而 ControlNet 生成的五官及面部阴影有时会出现锯齿状边界,而本文模型生成的人眼细节更为自然逼真,面部轮廓较为流畅,并在相对较小的数据集下能很好地学习到人脸的身份信息特征,达到大模型的人脸生成质量.

此外,参与比较的前 3 种模型都不提供对人类身份的控制,而本文的方法可以显式控制身份属性,更适用于政务区块链的工作需要.

为了对本文模型由身份代号文本生成人脸图像的质量进行更加直观地评价,采用如下评价指标对模型进行评估:

(1)FID(Fréchet Inception Distance):评估生成图像和训练图像之间相似性的度量. FID 越低表示生成图像质量越高.

(2)PSNR(Peak Signal-to-Noise Ratio):峰值信噪比,用来衡量重建图像的质量. PSNR 越高表示生成图像质量越高.

(3)SSIM(Structural Similarity):结构相似指标,可以衡量图片的失真程度. SSIM 越高表示生成图像质量越高.

(4)LPIPS(Learned Perceptual Image Patch Similarity):学习感知图像块相似度,用于度量两张图像的差别,更符合人眼感知. LPIPS 越低表示生成图像质量越高.

结果如表 3 所示,每项指标的最优值已加粗表示,实验结果表明本文模型在 3 个指标上均优于其他模型,区域感知码本与专家混合的引入,能帮助模型更加准确地拟合人脸图像,生成更精细的、高保真度的人脸图片. 在 FID 指标上,ControlNet 的表现仅优于本文模型的 1.12%,而模型体积是本文模型的 17 倍. 相比之下,本文模型可以在小数据集上,以 1/20 的参数量达到与大型模型相近的人脸生成质量,节省大量存储空间资源,更符合政务区块链使用场景需要.

表 3　人脸生成模型性能与大小定量比较

Table 3　Quantitative comparison of face generation models

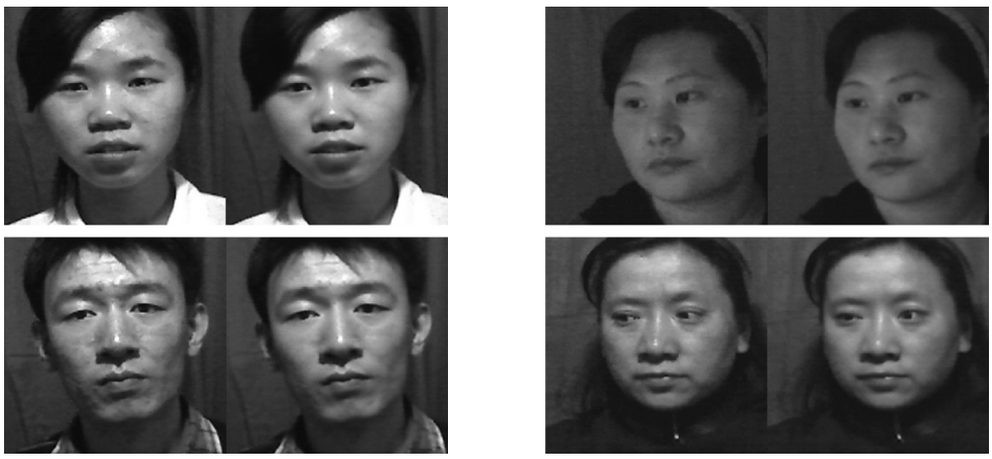
模型	FID	PSNR	SSIM	LPIPS	模型大小/MB
Pix2PixHD[18]	28.426	18.07	0.840	0.370	722.12
DINO[26]	46.422	27.85	0.942	0.292	855.35
GPEN[27]	32.762	22.53	0.871	0.356	1 405.52
ControlNet[10]	**16.942**	38.73	0.984	0.122	11 336.23
本文	17.142	**39.69**	**0.988**	**0.121**	**666.69**

对文本生成人脸图像的程序运行时间进行了 10 次测试,并计算平均值,得到生成每张 256×256 大小的人脸图像的平均时间为 2.432 s,能够在短时间内生成高保真度的指定人的人脸图像,以满足实际需求.

4　讨论

在本章中提出的人脸生成模型的 3 个常见限制和未来的研究方向:(1)皮肤平滑. 在 Casia Face V5 数据集中,包含较多细节(痣、斑、痘等)的人脸生成效果不佳. 图 8(a)显示了两个皮肤平滑的示例,面部的凹凸细节会丢失,这是目前生成模型都面临的常见问题,未来可以尝试设计多层 VQVAE、增加包含更多皮肤细节的数据来提升模型性能. (2)眼神偏移,如图 8(b)所示. 这归因于数据集的不平衡性,在训练图像中只有少数人半睁眼或眼神偏向一侧. 不过本文模型已经达到了生成细粒度的自然双眼的性能. 在未来的工作中,可以通过添加更多具有复杂神情的人脸数据来提高模型表现. (3)生成模型需要一定的采样时间. 模型生成人脸的响应时间关乎到政务区块链用户的使用体验,在本文模型生成一张人脸图片的时长里,扩散模型采样码本索引的步骤占据了绝大部分,未来可以引进采样速度更快的结构来对模型架构进行优化,提升生成人脸图像的速度.

此外,第三章实验的设置和设计是根据现有资源和时间等限制进行的. 在当前条件下,将 500 个身份的人脸图像转换为文本的准确率可达 95%以上,文本转换的所有人脸图像也达到了较高的质量与还原度,所采用的 2 500 张图片的数据集规模可以有效地验证所提出的跨模态人脸生成模型的性能,结果较为稳定,可以认为所提出的模型在更大数据集上也可能会有相似的表现. 未来,在更大数据集上进一步验证模型性能是一个重要的研究方向,这将为该模型在实际应用中的推广提供更多的支持.



(a) 皮肤平滑　　(b) 眼神偏移

图 8　人脸生成模型局限性案例

Fig. 8　Bad cases of the face generation models

5　结论

本文在公开 Casia Face V5 数据集上,提出了一种适用于政务区块链的跨模态人脸生成模型. 结合 ResNet-18 和多任务学习机制,进行高精确度的人脸分类,将人脸图像转化为身份代号文本后再上传至区块链,达到了传统图像压缩方法 1/10 000 的持久化时间与 1/200 的文件大小. 本文设计了区域感知码本来存储人脸图像不同区域的特征表示,然后提出了一个混合专家机制的采样器,以根据人脸分割掩膜对身份代号文本指示的人脸进行采样. 实验表明,相比其他先进的生成模型,本文提出的模型能够生成更细粒度和高逼真度的人脸图像,并以 1/20 的参数量达到与大型模型相近的人脸生成效果.

[参考文献]

[1] 王鹏,魏必,王聪. 区块链技术在政务数据共享中的应用[J]. 大数据,2020,6(4):105-114.

[2] NI Q Y,ZHANG L F,ZHU X R,et al. A novel design method of high throughput blockchain for 6g networks:performance analysis and optimization model[J]. IEEE internet of things journal,2022,9(24):25643-25659.

[3] SHAKYA K,PATLE D. A schematic review on image compression and encryption algorithm[J]. International journal for research in engineering and emerging trends(IJREET),2022,6(1):400-403.

[4] TAN Z J,YUAN X,MENG S W,et al. Building a modal-balanced blockchain with semantic reconstruction[J/OL]. arXiv preprint,arXiv:2303.02428,2023.

[5] MANDAL B,OKEUKWU A,THEIS Y. Masked face recognition using resnet-50[J/OL]. arXiv preprint,arXiv:2104.08997,2021.

[6] KARRAS T,LAINE S,AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2019:4401-4410.

[7] KARRAS T,LAINE S,AITTALA M,et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2020:8110-8119.

[8] STAP D,BLEEKER M,IBRAHIMI S,et al. Conditional image generation and manipulation for user-specified content[J/OL]. arXiv preprint,arXiv:2005.04909,2020.

[9] XIA W H,YANG Y J,XUE J H,et al. Tedigan:text-guided diverse face image generation and manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2021:2256-2265.

[10] ZHANG L,RAO A,AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). Paris,France:IEEE,2023:3836-3847.

[11] OORD A V D,VINYALS O,KAVUKCUOGLU K. Neural discrete representation learning[C]//Proceedings of the 31st International Conference on Neural Information Proceedingsessing Systems. New York:Curran Associates Inc,2017:6309-

6318.
[12] BOND-TAYLOR S,HESSEY P,SASAKI H,et al. Unleashing transformers:parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes[C]//Proceedings of the Computer Vision-ECCV 2022:17th European Conference. Berlin:Springer,2022:170-188.
[13] ESSER P,ROMBACH R,BLATTMANN A,et al. Imagebart:bidirectional context with multinomial diffusion for autoregressive image synthesis[C]//Advances in Neural Information Processing Systems. Montreal:MIT Press,2021,34:3518-3532.
[14] GU S Y,CHEN D,BAO J M,et al. Vector quantized diffusion model for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2022:10696-10706.
[15] GOODFELLOW I J,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Proceedingsessing Systems-Volume 2. Cambridge:MIT Press,2014:2672-2680.
[16] KINGMA D P,WELLING M. Auto-encoding variational bayes[J/OL]. arXiv preprint,arXiv:1312.6114,2013.
[17] ESSER P,ROMBACH R,OMMER B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2021:12873-12883.
[18] PARK T,LIU M Y,WANG T C,et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2019:2337-2346.
[19] CHEN X,QING L B,HE X H,et al. Ftgan:A fully-trained generative adversarial networks for text to face generation[J/OL]. arXiv preprint,arXiv:1904.05729,2019.
[20] CAMPOS-TABERNER M,ROMERO A,GATTA C,et al. Shared feature representations of lidar and optical images:trading sparsity for semantic discrimination[J]. 2015 IEEE international geoscience and remote sensing symposium(IGARSS),2015:4169-4172.
[21] SALIMANS T,KARPATHY A,CHEN X,et al. Pixelcnn++:improving the pixelcnn with discretized logistic mixture likelihood and other modifications[J/OL]. arXiv preprint,arXiv:1701.05517,2017.
[22] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. California:MIT Press,2017.
[23] CHANG H W,ZHANG H,JIANG L,et al. Maskgit:Masked generative image transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2022:11315-11325.
[24] SHAZEER N,MIRHOSEINI A,MAZIARZ K,et al. Outrageously large neural networks:The sparsely-gated mixture-of-experts layer[J/OL]. arXiv preprint,arXiv:1701.06538,2017.
[25] LIANG X D,GONG K,SHEN X H,et al. Look into person:Joint body parsing & pose estimation network and a new benchmark[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway,NJ:IEEE,2018,41(4):871-885.
[26] VOUGIOUKAS K,PETRIDIS S,PANTIC M. Dino:A conditional energy-based gan for domain translation[J/OL]. arXiv preprint,arXiv:2102.09281,2021.
[27] YANG T,REN P R,XIE X S,et al. Gan prior embedded network for blind face restoration in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2021:672-681.

[责任编辑:黄　敏]