计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(6):12−21 [doi: 10.15888/j.cnki.csa.009138]

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

多属性无监督人脸风格翻译①

朱剑锋[1], 郑　熠[1], 廖聪慧[2], 李孝杰[1], 梁梦娇[1]

[1](成都信息工程大学 计算机学院, 成都 610225)
[2](成都信息工程大学 通信工程学院, 成都 610225)
通信作者: 李孝杰, E-mail: lixj@cuit.edu.cn

摘　要: 针对现有人脸图像翻译模型不能实现多个视觉属性之间的翻译及翻译后的人脸图像不清晰自然的问题, 提出了基于人脸识别方法的人脸多属性图像翻译模型. 模型主要由内容和风格编码器、AdaIN 解码器以及人脸识别模块构成. 首先, 两个编码器提取内容和风格图像的潜在编码, 然后将编码送入到 AdaIN 层中仿射变换, 最后解码器还原翻译后的图像. 该方法设计并训练了一个准确率 90.282% 的人脸识别模型并提出了一种联合人脸属性损失函数, 增强了模型对风格人脸的属性的关注程度, 解决了模型不能准确提取到人脸的属性信息以及摒弃了无关信息, 使得模型能够生成清晰的、多属性的, 多样的人脸翻译图像. 该方法在公开的数据集 CelebA-HQ 实验并在定量和定性指标上都高于基线方法, 在不同的人脸朝向时也表现出良好的鲁棒性. 模型生成的图像还能应用于人脸图像生成领域, 解决数据集匮乏等问题.

关键词: 人脸图像翻译; 人脸识别; 图像生成; 人脸属性; 无监督学习; 风格翻译

引用格式: 朱剑锋,郑熠,廖聪慧,李孝杰,梁梦娇.多属性无监督人脸风格翻译.计算机系统应用,2023,32(6):12–21. http://www.c-s-a.org.cn/1003-3254/9138.html

Multi-attribute Unsupervised Face Style Translation

ZHU Jian-Feng[1], ZHENG Yi[1], LIAO Cong-Hui[2], LI Xiao-Jie[1], LIANG Meng-Jiao[1]

[1](School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)
[2](College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: To tackle the problem that the existing face image translation models cannot realize the translation among multiple visual attributes and the translated face images are not clear and natural, this study proposes a multi-attribute face image translation model based on the face recognition method. The model is mainly composed of the content and style encoder, AdaIN decoder, and face recognition module. First, the two encoders extract the potential encoding of the content and style image and then send the encoding into the AdaIN layer for affine transformation, and finally the decoder restores the translated image. A face recognition model is designed and trained using this method with an accuracy rate of 90.282%. A joint face attribute loss function is proposed, which enhances the model's attention to the attributes of the style face, solves the problem that the model cannot accurately extract the attribute information of the face, and discards irrelevant information so that the model can generate clear, multi-attribute, and diverse face translation images. This method is tested on the open dataset CelebA-HQ, whose results are higher than the baselines in terms of both quantitative and qualitative indicators. It also shows good robustness in different face orientations. The image generated by the model can also be used in the field of face image generation to address dataset shortage.

Key words: face image translation; face recognition; image generation; face attribute; unsupervised learning; style translation

① 基金项目: 四川省科技厅重点研发计划 (2021YFQ0053, 2022YFG0152); 四川省科技成果转移转化示范项目 (2023ZHCG0018); 四川省高等教育人才培养质量和教学改革项目 (JG2021-1015); 成都信息工程大学本科教育教学研究与改革项目暨本科教学工程 (JYJG2022131)
收稿时间: 2022-12-07; 修改时间: 2023-01-06; 采用时间: 2023-01-19; csa 在线出版时间: 2023-04-20
CNKI 网络首发时间: 2023-04-23

人类思维在泛化方面的能力很强. 因见过其他动物奔跑, 当我们看到一只卧着的狗就可以想象出它奔跑的样子, 认为狗在奔跑方面也应该具有与其他动物奔跑特性相一致、相符合、相统一的特点. 早期机器学习及计算机视觉在很多任务上都已取得了媲美人类的表现, 但其泛化能力仍远不及人类. 随着神经网络和生成对抗网络[1] 的快速发展, 可以利用图像翻译技术使得计算机拥有这样的能力. 图像到图像的翻译旨在学习不同视觉域之间的映射关系[2]. 域是指不同的类别, 比如每个人、每种野生动物我们可以归为一个域. 风格翻译是保持一张图的内容换为另一张图的风格, 将人脸图像的头发、眼睛的颜色, 肤色以及是否有胡须这些属性称之为风格. 将人脸的轮廓、姿态和表情称为内容. 由于人脸的数据集的特殊性和唯一性, 这导致风格是多样的, 使得设计和学习人脸编辑的模型变得复杂.

早期的图像翻译使用成对的数据[2], 但数据集很难获取. 而 CycleGAN[3] 利用循环一致性损失来强化内容的一致性, 可以使用未配对的数据来进行图像翻译. 但是 CycleGAN[3] 和其他模型[4–6] 只考虑了两个域之间的翻译, 如果要想得到 N 个域之间的转换, 就需要训练 $N\times(N-1)$ 个生成器来生成每个域之间的翻译. DRIT[7] 能扩展到 3 个域, 但是不能将人脸完成转换. 如图 1 所示用人脸数据集训练 CycleGAN, CycleGAN 并不能将源类的图片转换为具有目标图片的风格的图像.

图 1 CycleGAN 模型使用打乱数据集训练后的结果

为了解决人脸编辑中不同属性域之间的编辑, StarGAN 方法[8] 将人脸图片和属性标签同时输入. 但 StarGAN[8] 学习每个属性的固定翻译, 即 StarGAN 只针对某一个属性做特定的翻译, 比如更改发色、眼睛、鼻子.

为了解决多属性人脸翻译问题, 且生成多样的清晰的人脸图像. 本文受到 FUNIT 模型[9] 分别对内容图片和风格图片编码的启发提出了一种基于人脸识别模型的多属性无监督的人脸编辑方法. 该模型使用少样本进行训练, 将不同人的身份 ID 表示为不同的类别. 这样使得模型能够完成任意性别之间的编辑, 增强了模型输出结果的多样性. 模型主要由内容编码器、风格编码器、ArcFace[10] 人脸识别分类器、AdaIN 解码器和 Patch 判别器人脸识别分类器组成. 内容编码器主要编码人脸的内容特征, 比如姿态、面部表情等. 风格编码器主要编码人脸的风格属性, 比如眼睛、胡须、头发等. 然后利用 AdaIN[11] 的仿射变换得到翻译后的人脸图. 人脸识别模型使用 ArcFace 损失函数进行训练. 人脸识别模型摒弃了一些例如人脸图像背景等对模型的影响, 增强了人脸风格属性编码器对于人脸属性的提取和不同属性的翻译能力. 本文提出了一种联合人脸属性损失函数, 它是由人脸属性损失和三元组损失组成. 损失函数能够强化模型学习到更多有价值的信息. 实验结果表明本文模型在公开 CelebA-HQ[12] 数据集在定量指标和定性指标上都取得了比基线方法更好的结果.

1 相关工作

1.1 神经风格迁移

受到神经网络 CNN[13] 在图像视觉领域的成功应用, Gatys 等人[14] 提出了一个基于 VGG (visual geometry group) 模型[15] 的神经风格迁移模型. 但是 Gats 提出的模型不能到快速优化, 一些方法使用前馈风格模型[16–18] 来对图像进行风格化. 然而这些模型都存在着生成的图像的风格比较固定的问题. Hang 等人[11] 提出了自适应实例归一化 (adaptive instance normalization, AdaIN), AdaIN 通过将内容图片的均值和标准差与风格图片对齐, 从而实现快速的任意的风格迁移.

1.2 基于生成对抗网络的人脸风格编辑

2014 年, 生成对抗网络 GAN 由 Goodfellow 等人[1] 首次提出, 该网络主要包括生成器和判别器, 利用生成器和判别器之间的博弈来对抗学习, 最后使得生成器生成能欺骗鉴别器的图片.

人脸编辑应用于人脸的卡通化[19]、人脸补全[20,21]、人脸去模糊[22]、人脸的妆容迁移[23] 以及人脸的属性修改[8] 等. Zhang 等人[20,21] 使用生成对抗网络完成了对缺失人脸图像的补全, Zhang 等人[22] 利用多尺度渐进式网络完成了对模糊人脸的高清重建. APDrawing[13] 利用 CycleGAN 的模型来生成肖像画, 用质量损失来

指导网络生成更好的肖像画. StyleGAN[24] 实验发现不同分辨率的噪声影响着人脸的属性, 可以通过控制噪声来影响生成人脸的不同的属性. 然而这种方法在转换真实的人脸图像不是很成功, 因此它被设计成一个生成器模型. AdaAttN[25] 提出了一种新的注意力和规范化模块来执行样式转换并可以保持良好的内容, 但在风格上无法生成自然清晰的图片. StyTr^2 [26] 使用 Transformer 模型对于上下文的信息和位置联系的属性来对风格和内容图像进行编码, 使得 StyTr^2 模型能够输出内容完整的翻译图像. 但 Transformer 编码器不能够在编码阶段将人脸里面的眼睛、发色、鼻子等属性分开, 而更多的关注于风格图像属性的整体.

StarGAN 将人脸图片和对应的域标签都送入到生成器中并使用域监督的方式让生成器学习域的确定性映射. 然而它不能学习多种域之间的映射, 且在翻译的时候需要给定标签数据, 且不能提升模型的泛化能力. 对此 StarGAN v2[27] 使用了风格编码网络, 多个域输入对应多个输出的分支. 但是生成器和判别器都输出了多个域的结果, 但是这会使得模型的负担加重. MUNIT[28] 也采用了类似于 StarGAN v2 的多个域的输出方式, 但是对于多域多属性没有很好的泛化能力.

2 提出的模型

现有的模型不能学习到人脸多种属性且生成清晰自然的人脸图像, 为了解决这一问题, 本文受 FUNIT 和人脸识别模型启发, 将人脸识别模型中模型能够很好地学习到人脸的多种属性域的能力应用到人脸编辑的模型中. 本文的目标是使得模型具有泛化能力且生成清晰自然和具有多属性的人脸图像. 即给定两张人脸图像, 一张为内容图像, 一张为风格图像, 翻译图像保留着内容图像的姿势, 而具有风格图像的属性以及相似度. 本文的模型主要由人脸内容编码器、人脸属性编码器、人脸翻译解码器、人脸识别模块以及多任务对抗判别器构成.

2.1 人脸风格翻译生成器

本文的人脸风格翻译生成器如图 2 所示. 具体的, 真实内容人脸图像 I_C 先通过卷积神经网络编码后得到内容编码 Z_C, Z_C 中主要包含着人脸图像的姿态、面部宽度、轮廓等信息, 卷积神经网络由 4 层卷积层和两层的残差网络组成. 真实的风格图像 I_S 通过人脸风格编码器后得到人脸属性的潜在代码 Z_S, 里面包含了胡须、发色、眼睛、鼻子等信息. 解码器首先计算 Z_S 的均值和方差, 然后将这些矢量用作是 AdaIN 残差层的仿射变换参数, AdaIN 残差层将 Z_C 的均值和方差对齐来完成风格上的迁移. 然后通过上采样层将图片还原为输入图像的大小. 模型中输入输出的图片大小都是 128×128, 得到转换后的图片 I_T 和 I_S 被送到人脸识别模块中, 人脸识别模块提取两张人脸图像的特征, 并通过联合人脸属性损失函数计算损失, 并反向作用于人脸风格编码器. 提升了人脸风格编码器对人脸属性的学习, 如胡须、眼睛、发色, 从而使得模型更加关注于人脸图像 I_S 的属性特征, 且摒弃了人脸图片无关的信息, 无关的信息会影响生成图片的清晰度, 由此完成多域多属性的人脸编辑. 这样机器和计算机能够像人一样具有泛化的能力.

2.2 人脸识别模块

图 3 展示了本文的人脸识别模块的设计结构图. 特征提取层主要由 10 个卷积块、1 个展平块和 1 个全连接层构成. 具体的特征提取部分被设计为:

$$\begin{aligned} &Conv2BnPReLU \rightarrow Conv2Bn \rightarrow DepthWise \\ &\rightarrow DepthWiseResNet \rightarrow DepthWise \\ &\rightarrow DepthWiseResNet \rightarrow DepthWise \\ &\rightarrow DepthWiseResNet \rightarrow Conv2Bn \\ &\rightarrow Conv2BnPReLU \end{aligned}$$

Conv2Bn 是卷积操作加一个 *Bn* 层, *Conv2BnPReLU* 是在 *Conv2Bn* 加上一个 *PReLU* 激活层, *DepthWise* 卷积里面的一个卷积核只和一个通道进行卷积计算, 它实现了更高效率的计算. *DepthWiseResNet* 是在 *DepthWise* 卷积后加上了一个原始输入, 即是利用了 ResNet[29] 的思想, 这样网络可以有效地利用前面的信息. *Flatten* 是铺平操作, 最后接一个 *Linear* 层. 本文的人脸识别和直接利用 VGG 模型[15] 模型的人脸识别不同的是, 本文的模型没有很多的层数, 比如 VGG19 有 19 层, ResNet50 有 50 层. 这样设计不仅加快了模型的计算速度, 也使得网络模型更加小. 在训练时, 人脸图片经过特征提取网络得到人脸的特征, 特征经过 ArcFace[10] 计算当前特征和目标权重之间的角度, 然后在目标角上加上一个附加的角度. 最后通过 Softmax 来得到分类结果. ArcFace[10]损失函数能够使得人脸类间的距离变小使得类间的距离更大.

该人脸识别模型在 CelebA 数据集[30] 上训练, 并在 LFW (labeled faces in the wild)[31] 数据集测试, 模型在 LFW 数据集上测试成绩正确率为 90.282%.

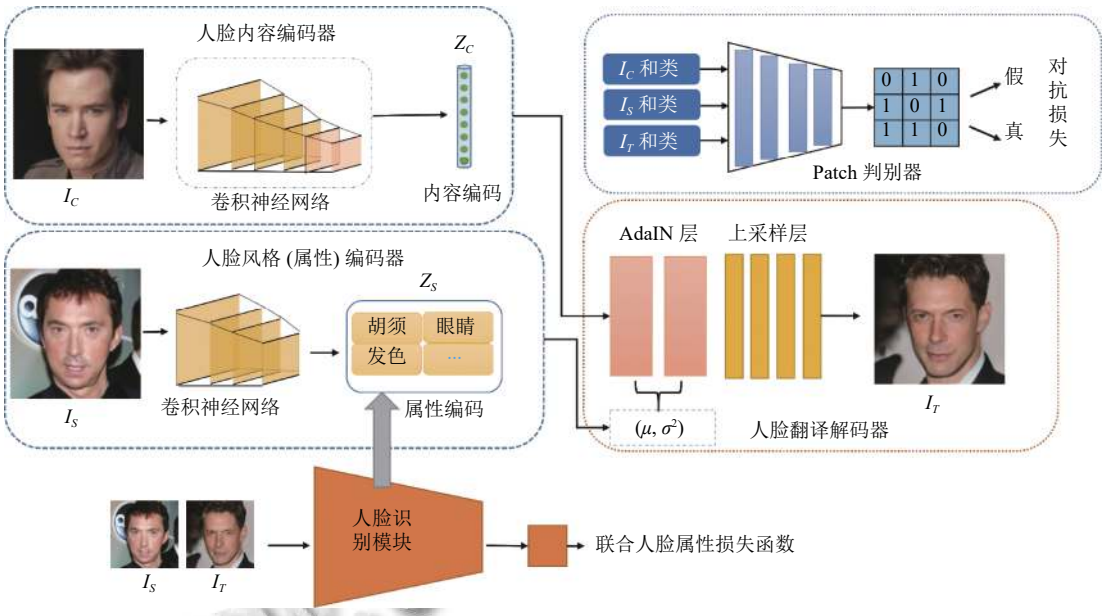


图 2 网络模型架构

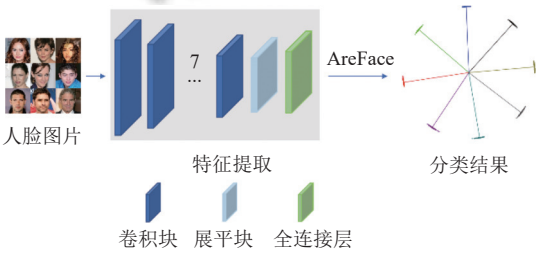


图 3 人脸识别模块的网络结构图

2.3 多任务对抗判别器

判别器的主要功能是使得生成器能够生成清晰自然且准确的人脸图像, 如图 2 所示. 该判别器来自 PatchGAN[32], Patch 判别器将输入的样本计算其为真样本概率判别器不同的是 Patch 判别器将输入的样本映射为 $N\times N$ 的矩阵, 即分块判别, 而最终的概率是矩阵元素的平均值. 由于是分块, 所以会更加关注细节的变化, 这更适合于人脸这样对细节要求更高的图.

判别器是通过求解多个对抗的二分类任务来进行训练的. 具体的, 判别器需要判别生成的图像是真实的图片还是模型生成的图片. 当更新判别器时, 如果判别器将输入的真实图像判别为假, 则会惩罚判别器, 如果对于转换后的图像, 判别器判定为真, 也会惩罚判别器. 当更新生成器的时候, 只有当判别器将生成的图片判别为假的时候, 才会惩罚生成器. 具体的判别器的结构设计为:

Conv-64 → *ResBlk*-128 → *ResBlk*-128 → *AvePool*
→ *ResBlk*-256 → *AvePool* → *ResBlk*-512
→ *ResBlk*-512 → *AvePool* → *ResBlk*-1024
→ *ResBlk*-1024 → *AvePool* → *ResBlk*-1024
→ *ResBlk*-1024 → *Conv*

Conv 是卷积层, *ResBlk* 是残差层, *AvePool* 是池化层.

2.4 损失函数

模型的总的损失函数被表示为:

$$\min_G \max_D L_{\text{GAN}}(D,G)+\lambda_R L_R(G) +\lambda_F L_F(G)+\lambda_{FA} L_{FA}(G) \tag{1}$$

其中, L_{GAN}、L_R、L_F、L_{FA}, 分别是 GAN 对抗损失、内容重建损失、特征匹配损失、联合人脸属性损失. 具体定义如下.

L_{GAN} 对抗损失函数采用下面的有条件的损失函数:

$$L_{\text{GAN}}(D,G)=E_C\left[-\log D^{CL_C}(C)\right] +E_{C,S}\left[\log(1-D^{CL_S}(I_T))\right] \tag{2}$$

其中, D 的上标表示的是对象类别, CL 表示类, C 和 S 是输入的内容图像和风格图像, I_T 表示的是模型生成的人脸图像. 本文计算损失使用的是其二分类的预测分数. 对抗损失函数能使得判别器和生成器对抗训练, 生成器生成的图像达到想要的结果.

L_R 重建损失可以帮助生成器 G 生成的图像 I_T 保持和内容图像 I_C 一样的内容. 在训练的时候将内容图

像 I_C 分别以内容图像和风格图像送入到模型中, 结果模型生成的图像应该是和输入的图像一样, 以这样的方式来保持内容的一致性, 然后计算其 $L1$ 范数来作为损失函数:

$$L_R(G) = E_{I_C}\left[\|I_C - G(I_C, \{I_C\})\|_1^1\right] \tag{3}$$

特征匹配损失函数能够有效地解决生成器和判别器不能训练的问题, 使得 GAN 模型训练正常化. 同时, 特征匹配损失作用于生成器, 能够约束生成器生成质量更加好的图片. 使用判别器 D 最后一层预测层前面的层构造图像的特征提取器 D_f. 然后提取转换输出的图像 I_T 和风格图像的 I_S 的特征, 并计算其 $L1$ 范数, 并使其最小化.

$$L_F(G) = E_{I_C, I_S}\left[\|D_f(I_T) - D_f(I_S)\|_1^1\right] \tag{4}$$

联合人脸属性损失函数约束模型提取到更好的人脸属性, 以及可以让模型生成的图像 I_T 更接近于 I_S 的域. 具体地讲, 模型生成的图像看起来应该是属于 I_S 的域, 即拥有 I_S 的属性, 但是其姿势更加接近 I_C, 将 3 张图片全部送入到 ArcFace 人脸识别模型中, 并提取其特征. 如式 (5) 所示, F_R 是指人脸识别模型提取图片的特征, 最小化 I_S 和 I_T 之间的距离以及最大化 I_T 和 I_C 之间的距离, 然后计算 I_S 和 I_T 的特征之间的余弦距离, 其最好的效果是余弦距离为 1, 故用 1 减去它们的余弦距离值. α 是 margin 参数, 其作用是拉大 $F_R(I_T)$ 和 $F_R(I_S)$ 图像对和 $F_R(I_T)$ 和 $F_R(I_C)$ 图像对之间的距离, 按照经验将 α 的值设定为 0.5.

$$\begin{aligned} L_{FA}(G) = \max(&\|F_R(I_T) - F_R(I_S)\|^2 - \|F_R(I_T) - F_R(I_C)\|^2 \\ &+ \alpha, 0) + (1 - \cos(F_R(I_T), F_R(I_S))) \end{aligned} \tag{5}$$

3 实验

3.1 实验设计

(1) 数据. 本文的数据来自公开数据集 CelebA-HQ[12], 里面包含了 7605 个人的人脸数据, 大多数是西方人. 本文将 CelebA-HQ[12] 中的图片按照每个人的身份 ID 进行整理, 我们没有发现网上有对该数据集的身份整理数据. 整理后的数据可以在 https://github.com/MOKEjf/CelebA-HQ-Classify-by-ID 查看和下载.

(2) 对比基线方法. 本文使用最近 5 年的模型 2018: MUNIT[28]、DRIT[8]; 2019: MSGAN[5]; 2020: StarGAN v2[27]; 2021: AdaAttN[25]; 2022: StyTr^2[26]. 作为实验的对比基线. 为了保证实验的公平性, 上述对比基线实验的设置都是用作者提供的开源版本以及作者提供的参数进行训练.

(3) 评估指标. 本文使用 *FID*[1]、*LPIPS*[33] 和 *IS*[34] 来评估生成模型结果图像的质量. *FID* 的分数衡量真实图像和生成图像的特征向量之间的距离的度量, 值越低, 代表生成的两组图像越相似. *LPIPS* 意为学习感知图像块的相似度, 用来度量两个图像之间的差别. *IS* 用来衡量生成的图像的清晰度和图像的多样性. *FID*、*LPIPS* 和 *IS* 的计算公式分别为:

$$FID = \|\mu_x - \mu_g\|_2^2 + Tr\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{1/2}\right) \tag{6}$$

其中, μ_x 和 μ_g 分别是真实图像和生成图像的特征均值, Σ_x 和 Σ_g 代表真实图像和生成图像的协方差矩阵, T_r 为矩阵理论中成为矩阵的迹. 使用来自图像预训练的 Inception-V3 模型的最后一层平均池化层的特征向量, 对真实图像和生成的图像计算其 *FID* 平均值.

$$LPIPS = d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}_{hw}^l - \hat{y}_{0hw}^l\right)\right\|_2^2 \tag{7}$$

其中, d 表示真实图像 x 和生成图像 x_0 之间的距离. 从 l 层提取特征堆并在通道维度中进行单位规格化, 最终计算 $L2$ 距离. 本文使用从图像网络预训练的 AlexNet 模型提取特征, 并计算真实图像和生成图像之间的距离, 最后报告其平均分数作为 *LPIPS* 分数.

$$IS = \exp\left(E_{x\sim P_g} D_{KL}(p(y|x) \,\|\, p(y))\right) \tag{8}$$

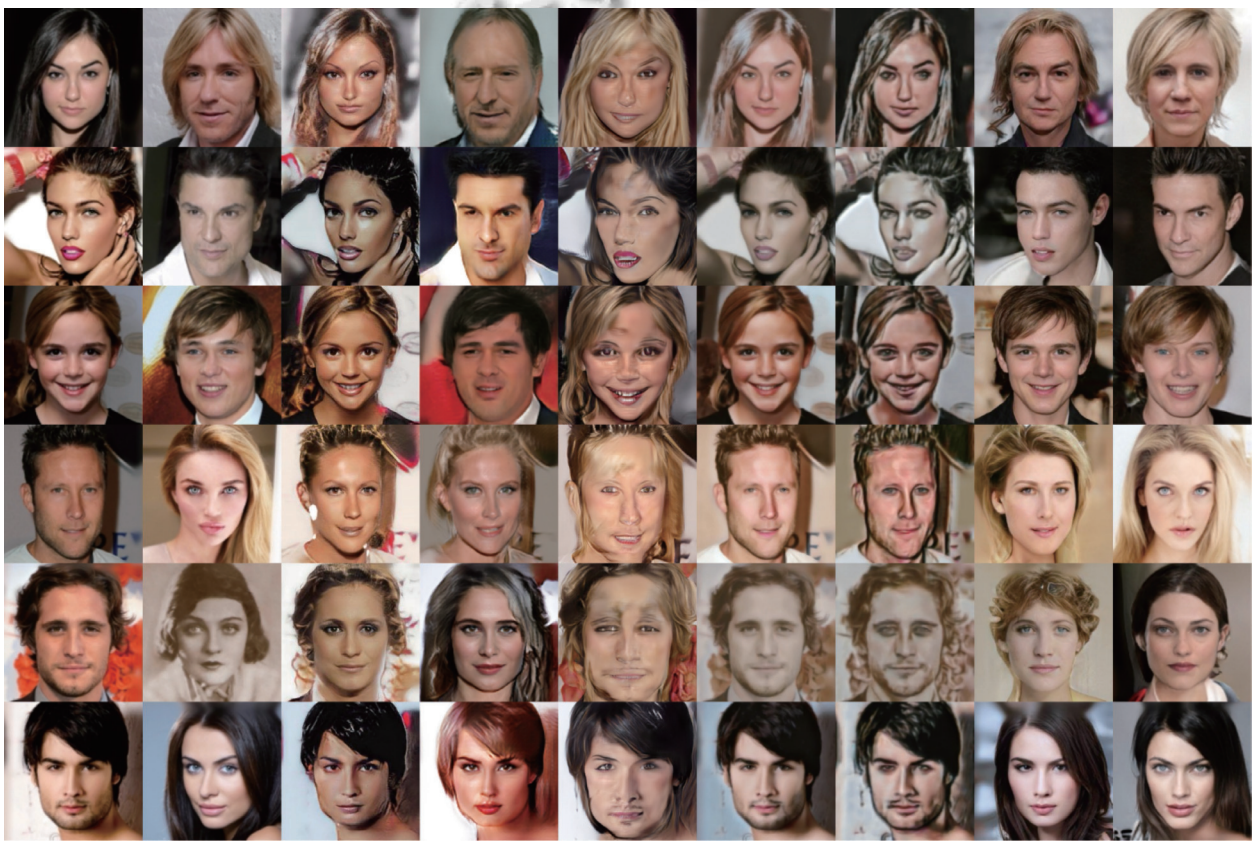
其中, g 表示生成器, x 表示生成的图像, y 表示 Inception-V3 模型[35] 预测的标签.

(4) 实验参数设置. 本文的实验中将 λ_R, λ_{FA} 和 λ_F 参数分别设置为 0.1、10、和 1. 使用 *RMSProp* 优化器来优化训练, 将学习率设置为 0.000 1, 最终使用 GAN 损失的铰链版本作为模型的结果, 最后的生成器是中间生成器的平均结果. 本文使用 Mescheder 等人[36] 提出的真实梯度惩罚正则化, 能够保证损失函数稳定下降. 每一批次训练同时给模型送入 8 张的内容图像和 8 张的风格图像, 设置训练迭代次数最大为 300 000 次. 实验在一张 2080TI 图像计算卡上完成, 环境为 Python 3.7.9、PyTorch 1.4.0、CUDA 10.0 以及 Ubuntu 16.04.

为了保证实验的公平性, 其他的对比方法也使用和本文的相同的硬件环境和软件环境, 对于对比方法中的参数使用其论文中的默认参数来进行实验.

3.2 定性比较

如图 4 所示本文模型和基线模型的翻译结果图像对比结果. MUNIT 模型几乎完全保留了内容图像的轮廓, 但是却不能反映每个风格图像的风格, 而且不能输出质量较高的人脸图像. DRIT 能够生成相对自然的人脸图像, 但是其风格大多是颜色上的匹配, 不能很好地捕获风格人脸图像的属性. MSGAN 能够捕获到风格图像的属性, 但是翻译的图像却不自然, 内容保持不好. StyTr^2 使用 transformer 来编码内容并且注重内容上的一致性, 但是其不能完成对人脸风格的翻译, 只有少量的图片完成的对头发颜色的翻译. AdaAttN 不能生成清晰自然且轮廓不连贯的人脸图, 只有倒数第 3 行的风格翻译明显. StarGAN v2 输出的图像反映了风格图像的风格特征, 也对于内容图像的姿态也有很好的保持, 和本文的模型不同的是 StarGAN v2 输出的图像在视觉上和内容图像相似, 而该模型生成的人脸更像风格图像. StarGAN v2 输出图像的背景和人脸部分没有很好的连贯起来, 不自然. 特别的, 图 4 的倒数第 2 行的风格图像是一副黑白的图像, 但是本文的模型以及基线的方法都赋予了彩色的风格, 这是因为训练数据集中很少有这种黑白的图像, 模型具有更好的泛化能力, 而对于这种少量的图像不能够匹配其风格. 本文的模型生成的图像在内容上、风格迁移的准确度以及人脸的细节都高于基线的方法.



(a) I_C (b) I_S (c) MUNIT (d) DRIT (e) MSGAN (f) StyTr^2 (g) AdaAttN (h) StarGAN v2 (i) Ours

图 4 该方法生成的图像和其他基线方法的定性对比结果

图 5 更详细地展示了本文模型输出的翻译图像的结果. 可以看出, 本文的模型能够将风格人脸图的风格属性翻译到输出的图片中同时保留着内容图像的姿态和内容. 本文的模型输出了清晰且自然的人脸图像, 不论是计算机或者人类都很难辨别出这是由模型生成的图像. 同时本文的模型还能够完成同性别和不同性别的人脸之间的翻译, 对于图 5 最后一行的风格图像有着黑色的背景和人脸特征不是很容易提取的图来说, 本文翻译图像仍然能够保留着风格图像人物整体肤色和胡须等特征.

(a) I_C (b) I_S (c) I_T

图 5 模型的翻译生成结果图展示

3.3 定量对比

本文报告了该方法和基线方法的 *FID*[1]、*LPIPS*[33] 和 *IS* 分数[34], 如表 1 所示.

由表 1 可知在 *FID* 和 *IS* 两项指标上本文都取得了高于基线方法的结果, 在 *LPIPS* 评价指标上仅仅低于 StarGAN v2, 高于其他的基线方法. *FID* 主要判定两组图像相似程度, 使用真实的图像和生成的图像来计算 *FID* 的值, 数值越低越好, 本文取得了较低的分数. 这也和展示的图片结果一致, AdaAttN 的 *FID* 值较高, 这是由于 AdaAttN 输出的人脸图像质量不好, 和真实的图像相似性很低. *LPIPS* 衡量两个图像的差别, 实验中计算 I_C 和 I_T 图像对之间的 *LPIPS* 值, 并求平均值. *LPIPS* 感知损失指标比 StarGAN v2 低了 0.067, 这是因为 StarGAN v2 的转换输出结果在视觉效应上更加接近于内容图片, 所以它具有更好的 *LPIPS* 的结果. *IS* 主要衡量生成的图片是否多样和清晰, 本文的结果高于基线的方法, 较 StarGAN v2 提升了 0.471, MUNIT 和 StyTr^2 都取得了 2.0 以上的分数. 最后一行真实的图片的数值是全部使用真实的图像去计算, 由于不存在和真实图像一一对应的假图故不计算 *LPIPS* 值, 由于其他的实验都是算法合成的图像故在 *PID* 和 *IS* 指标上真实的图像取得最好的分数.

表 1 本文和基线方法的性能指标对比

方法名称	*FID* ↓	*LPIPS* ↑	*IS* ↑
MUNIT	48.1	0.176	2.211
DRIT	21.3	0.258	1.698
MSGAN	39.6	0.312	1.672
AdaAttN	89.3	0.313	1.584
StyTr^2	20.1	0.301	2.207
StarGAN v2	23.8	**0.388**	1.761
Ours	**18.8**	**0.321**	**2.232**
Real images	17.9	—	2.446

注: 箭头的方向表示指标的性能优的方向

3.4 消融实验

为了进一步证明人脸识别模块和联合人脸属性损失函数的能够使得模型具有很好的人脸属性翻译功能. 将人脸识别模块和联合人脸属性损失函数去除后的实验结果如图 6 所示, 可以观察到, 去除过后得到的结果和风格图片的相似度没有本实验的好, 而且没有将发色、胡须准确翻译. 本文的结果也取得了更好的视觉效果, 性能指标的结果如表 2 所示, 本文的模型在 *FID*、*LPIPS* 和 *IS* 指标上都取得了更好的结果.

(a) I_C (b) I_S (c) 消融 (d) Ours

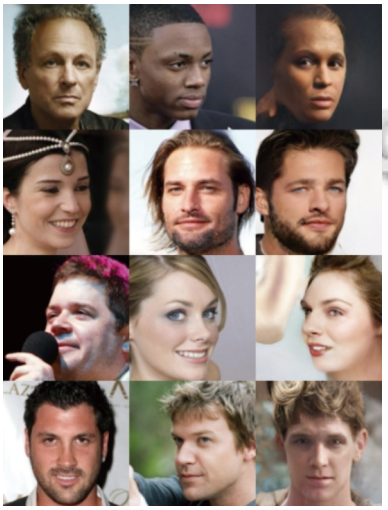
图 6 消融实验的结果图

表 2 消融实验的性能指标对比

实验名称	*FID* ↓	*LPIPS* ↑	*IS* ↑
消融	48.1	0.176	2.211
Ours	**21.3**	**0.258**	**1.698**
Real images	17.9	—	2.446

3.5 其他分析

实验中发现, 数据集中的大部分的人脸图像都是正脸的, 但是也有少部分的人脸图像是姿态偏移比较大, 这就要求模型具备很好的特征提取能力和更好的生成图片的能力. 无论是对于内容图像还是风格图像, 如果人脸的姿态有很大的偏移, 模型往往生成的图像人脸的五官不会很清晰以及不会很好地将两张图像融合. 实验证明该模型对于不是正脸的图像也能够得到很好的翻译结果, 如图 7 所示, 该模型输出的人脸图像都保留着风格图像的属性, 无论是发色、面部特征、胡须等.



(a) I_C (b) I_S (c) I_T

图 7 面部偏移的部分展示图

该模型生成的图像也可以应用于其他研究, 比如用来做人脸生成, 用于扩充人脸数据集, 因为该模型生成的图像是不存在的人脸图像, 生成的人脸图像有着良好的视觉效果、清晰度和很高的质量. 结果如图 8 所示.

图 8 模型生成的人脸的部分结果展示图

4 结论和展望

为了生成保留内容图像姿态和朝向且具有风格图像风格属性的人脸图片. 本文设计了一种基于人脸识别模型的多属性翻译的生成模型, 它解决了现有的大多数模型不能翻译多种属性的问题以及能够生成更多样的图片. 人脸识别模块以及联合人脸损失函数能够使得模型学习到更多的人脸属性, 将每个人的身份 ID 以一个类来训练模型, 能够让模型生成更多的人脸图像, 且可以完成对任意一张人脸的风格翻译. 在人脸的公开数据集上的实验表明本文的模型无论是定性指标还是定量指标都好于基线的方法. 本文的模型在人脸面部偏移也表现出良好的鲁棒性.

现在人脸编辑也应用于很多方面, 同时也有很大的挑战. 如生成的更好更高清的人脸对模型以及对算力都有很高的要求. 因此人脸翻译模型中生成高分辨率的人脸图片将是后续研究的重点.

参考文献

1 Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks. arXiv:1701.00160, 2016.

2 Yang Z, Chen W, Wang F, *et al*. Improving neural machine translation with conditional sequence generative adversarial nets. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 1346–1355.

3 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.

4 Almahairi A, Rajeshwar S, Sordoni A, *et al*. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 195–204.

5 Mao Q, Lee HY, Tseng HY, *et al*. Mode seeking generative adversarial networks for diverse image synthesis. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1429–1437.

6 Na S, Yoo S, Choo J. MISO: Mutual information loss with stochastic style representations for multimodal image-to-image translation. arXiv:1902.03938, 2019.

7 Lee HY, Tseng HY, Huang JB, *et al*. Diverse image-to-image translation via disentangled representations. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 36–52.
8 Choi Y, Choi M, Kim M, *et al*. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8789–8797.
9 Liu MY, Huang X, Mallya A, *et al*. Few-shot unsupervised image-to-image translation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 10550–10559.
10 Deng JK, Guo J, Xue NN, *et al*. ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, IEEE: 2019. 4685–4694.
11 Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1510–1519.
12 Karras T, Aila T, Laine S, *et al*. Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196v3, 2017.
13 Yi R, Liu YJ, Lai YK, *et al*. Quality metric guided portrait line drawing generation from unpaired training data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 905–918. [doi: 10.1109/TPAMI.2022.3147570]
14 Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2414–2423.
15 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.
16 Park DY, Lee KH. Arbitrary style transfer with style-attentional networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5873–5881.
17 Zhang H, Dana K. Multi-style generative network for real-time transfer. Proceedings of European Conference on Computer Vision. Munich: Springer, 2018. 349–365.
18 Li YJ, Fang C, Yang JM, *et al*. Diversified texture synthesis with feed-forward networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 266–274.
19 Zhuang N, Yang C. Few-shot knowledge transfer for fine-grained cartoon face generation. Proceedings of the 2021 IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE, 2021. 1–6.
20 Zhang X, Wang X, Shi CH, *et al*. DE-GAN: Domain embedded GAN for high quality face image inpainting. Pattern Recognition, 2022, 124: 108415. [doi: 10.1016/j.patcog.2021.108415]
21 Zhang X, Shi CH, Wang X, *et al*. Face inpainting based on GAN by facial prediction and fusion as guidance information. Applied Soft Computing, 2021, 111: 107626. [doi: 10.1016/j.asoc.2021.107626]
22 Zhang H, Shi CH, Zhang X, *et al*. Multi-scale progressive blind face deblurring. Complex & Intelligent Systems, 2022. [doi: 10.1007/s40747-022-00865-9]
23 Nguyen T, Tran AT, Hoai M. Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13300–13309.
24 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4396–4405.
25 Liu SH, Lin TW, He DL, *et al*. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 6629–6638.
26 Deng YY, Tang F, Dong WM, *et al*. StyTr^2: Image style transfer with transformers. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11316–11326.
27 Choi Y, Uh Y, Yoo J, *et al*. StarGAN V2: Diverse image synthesis for multiple domains. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8185–8194.
28 Huang X, Liu MY, Belongie S, *et al*. Multimodal unsupervised image-to-image translation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 179–196.
29 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
30 Liu ZW, Luo P, Wang XG, *et al*. Deep learning face attributes in the wild. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago:

IEEE, 2014. 3730–3738.

31 Huang GB, Mattar M, Berg T, *et al*. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Proceedings of Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Marseille, 2008. 1–14.

32 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.

33 Zhang R, Isola P, Efros AA, *et al*. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595.

34 Salakhutdinov R, Tenenbaum J, Torralba A. One-shot learning with a hierarchical nonparametric bayesian model. Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop. Bellevue, WA: JMLR.org, 2011. 195–207.

35 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826.

36 Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge? Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 3481–3490.

(校对责编: 牛欣悦)