计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009812] [CSTR: 32024.14.csa.009812]

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

基于改进 StarGAN-V2 的多域面部表情转移①

王春东[1,2], 张浩龙[1,2]

[1](天津理工大学 计算机科学与工程学院, 天津 300384)
[2](计算机病毒防治技术国家工程实验室, 天津 300384)
通信作者: 王春东, E-mail: michael3769@163.com

摘　要: 多域面部表情转移涉及不同图像之间的相互转换, 目的是生成具有源面部表情和目标面部身份特征的高逼真度面部图像, 解决传统方法生成图像相似度高和图像真实性低的问题. 本文提出了一种基于改进 StarGAN-V2 的多域面部表情转移模型, 该模型由生成器、鉴别器、映射网络和风格编码器组成, 引入了空间注意力机制, 并将循环一致性损失改进为对抗性循环一致性损失, 在生成器后增加了一个新的域反馈鉴别器. 该改进后的 StarGAN-V2 模型能够基于源图像和目标图像, 生成具有源面部表情和目标面部身份特征的高逼真度面部图像. 实验结果表明, 改进后的模型潜在引导合成和参考引导合成 FID 值为 11.9 与 17.4, LPIPS 值为 0.491 与 0.426, 均优于对照模型, 改进后的模型解决了图像相似度高的问题, 生成的图像也更加真实.

关键词: 面部表情转移; StarGAN-V2; 多域风格; 网络安全

引用格式: 王春东,张浩龙.基于改进 StarGAN-V2 的多域面部表情转移.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9812.html

Multi-domain Facial Expression Transfer Based on Improved StarGAN-V2

WANG Chun-Dong[1,2], ZHANG Hao-Long[1,2]

[1](School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China)
[2](National Engineering Laboratory for Computer Virus Prevention and Control, Tianjin 300384, China)

Abstract: Multi-domain facial expression transfer entails the mutual transformation between different images to generate high-fidelity facial images with source facial expressions and target facial identity features, solve the problem of high similarity and low image authenticity of images generated by traditional methods. This study proposes a multi-domain facial expression transfer model based on the improved StarGAN-V2. The model consists of a generator, a discriminator, a mapping network, and a style encoder. The spatial attention mechanism is introduced, and the cycle consistency loss is upgraded to an adversarial cycle consistency loss. A new domain feedback discriminator is appended after the generator. The improved StarGAN-V2 model can generate high-fidelity facial images with source facial expressions and target facial identity features based on the source and target images. Experimental results show that for the improved model, the FID values of latent guided synthesis and reference guided synthesis are 11.9 and 17.4 respectively, and the LPIPS values are 0.491 and 0.426 respectively. These values are better than those of the control model. The improved model solves the problem of high image similarity and generates more realistic images.

Key words: facial expression transfer; StarGAN-V2; multi-domain style; cybersecurity

面部表情转移[1]是计算机视觉和安全领域近年来备受瞩目的研究课题之一, 其核心技术是将一个人的面部表情特征应用到另一个人的面部图像上. 这项技术不仅限于简单的面部表情变化, 还可以扩展到更多

① 基金项目: 国家重点研发计划"区块链"重点专项 (2023YFB2703900); 天津市科委重大专项 (15ZXDSGX00030)
收稿时间: 2024-09-17; 修改时间: 2024-10-30; 采用时间: 2024-11-07; csa 在线出版时间: 2025-02-25

形式的图像转换任务, 例如将某人的黑色头发通过图像处理转换为金色头发, 或者将一只猫的图像转换为狗的图像[2]. 这类图像转换任务的研究成果在多个实际应用领域中得到了广泛的应用, 包括电影特效制作、视频游戏开发、虚拟化妆、虚拟聊天系统等[3]. 另一方面, 在网络安全领域也具有重要的意义, 通过生成具有误导性的面部表情图像, 我们可以用于研究和加强对抗生成网络攻击, 提高网络安全系统的鲁棒性. 这些生成的图像可以模拟各种可能的攻击场景, 帮助开发者更好地理解和应对潜在的威胁, 通过改变面部表情减少面部识别系统对用户的跟踪和识别, 有效降低隐私泄露风险[4].

其中生成对抗网络 (generative adversarial network, GAN)[5]模型在这一领域取得了显著进展, 尤其是 StarGAN-V2 模型, 已经成为多领域图像生成任务中的一种主流工具. StarGAN-V2 模型最早提出时的设计初衷是实现多领域之间的图像转换任务, 例如将一个人的面部表情转换为另一个人的, 同时还能实现发型、肤色等其他领域特征的同步变化. 通过在多任务学习框架下来共享生成器网络, 而且 StarGAN-V2 模型能够同时处理多个图像转换任务, 从而显著提升了模型的训练效率和生成效果.

本文提出一种基于改进的 StarGAN-V2 面部表情转移模型, 改进后的模型在模型生成器增加了空间注意力机制和优化后的对抗性循环一致性损失函数, 加入了新的域鉴别器, 使生成器能够更加准确地捕捉和再现源图像的细节, 同时保留目标域的特征, 不仅提升了图像生成的稳定性和细节清晰度, 还大大增强了模型在实际应用中的可靠性和实用性. 改进的模型不仅在视觉效果上有所突破, 还为网络安全提供了新的工具和方法, 不仅能推动图像生成技术的创新, 也为安全防御技术和网络安全研究带来了新的思路和应用场景, 又丰富了生成对抗网络在多领域的潜力.

1 相关方法

目前国内外关于面部表情转移的研究颇为广泛, 近年来, 面部表情转移技术在深度学习的推动下取得了显著进展, 尤其是生成对抗网络的应用使得这一领域取得了突破性成果[6]. GAN 通过对抗性训练, 能够生成高质量、逼真的图像, 成为面部表情迁移的核心技术框架. 与传统方法相比, 基于 GAN 的技术不仅能够保留目标人脸的身份特征, 还能精准地将源人脸的表情迁移到目标人脸上.

Isola 等人[7]提出使用 CGAN 以监督的方式训练数据, 结合了对抗损失和 L1 损失, 但是需要成对的数据样本, 导致生成的图像包括面部表情图像多样性不够, 容易出现模式坍塌现象, 比较过度依赖数据集. Kim 等人[8]和 Liu 等人[9]提出了不配对的面部表情图像到面部表情图像的框架, 暂时缓解了数据对的问题, 但是仍会遇到不能很好保留原始数据特征的问题. Zhu 等人[10]提出使用 CycleGAN 进行两个图像之间的转移, 利用循环一致性损失来保留源图像与目标图像的关键属性, 适用于没有配对训练数据的情况, 但是比较依赖全局图像特征转换, 缺乏图像的精细控制, 导致生成质量差别大. Wu 等人[11]提出了一种利用关键特征识别面部表情的多任务深度学习框架, 可以精确的识别面部的关键特征点, 细粒度高, 多任务学习又能提高训练效率, 但是需要大量的高质量标注数据集, 包括详细的面部关键特征点和多样化的表情数据, 成本较高且对质量有要求, 联合训练容易出现不稳定的问题. Tomar 等人[12]采取一个分类器, 将传统合作表达和逆线性回归结合起来, 传递图像的对称性构建一个新的面部表情样本, 这种方法增强了样本的多样性和对称性, 能够降低过拟合的风险, 但是依赖于图像的对称性假设, 现实中的图像并不总是完全对称的, 可能导致某些情况下生成的图像不够自然或存在伪影. Huang 等人[13]提出使用 MUNIT 进行面部图像转移, 通过内容特征和风格特征解耦, 使用两个生成器和两个编码器分别提前图像的内容特征和风格特征, 生成器将两个特征混合生成不同风格的面部图像, 但其是单对单领域转换, 扩展性较差, 风格控制精度不足. Lee 等人[14]提出 DRIT 的方法, 将输入图像的内容和风格特征解耦为两个独立的潜在变量空间, 引入了新颖的跨循环一致性损失, 允许生成具有不同风格特征的图像, 同样是单对单领域转换, 生成的图像在细节和一致性上较差, 尤其在复杂任务的表现. Mao 等人[15]提出了 MSGAN, 通过最大化生成图像之间相对于其对应潜在代码的距离比率, 显式地鼓励生成器在训练过程中探索更多的次要模式. 该模式探索正则化项可以直接应用于各种条件生成任务, 且不会增加训练开销或修改原始网络结构. 在多模态下表现地更好, 增强了多样性, 避免了模式崩溃. 但是在跨领域任务中仍然适用性较低. 在 Choi 等人[16]提出的

StarGAN-V2 模型中, 引入了映射网络和风格编码器, 不再依赖离散的域标签, 使用了 AdaIN 自适应实例归一化解决了传统的域迁移需要在两个不同域之间进行特征提取的问题, 更好地保留了内容特征, 同时更适合应用于风格变化. 刘鹤[17]对 StarGAN-V2 模型进行了修改 (以下称 StarGAN-V2Conv), 其生成器中嵌入具有多尺度融合功能的简化版 PSConv 模块, 提高了生成器的特征提取能力, 加入了多维注意力方法 MDconv, 提高了鉴别能力.

上述部分方法集中在单个域, 不能实现一个源图像向多个目标图像转换, 部分方法仍然存在清晰度不够、图像不自然等系列问题. 虽然 StarGAN-V2 及修改的模型相对于以上方法取得了更好的性能, 但是仍存在一些缺陷.

(1) 生成图像真实性低: 尽管 StarGAN-V2 在生成多模态图像和高质量图像方面表现出色, 但存在生成的图像仍然会不够自然或不够真实.

(2) 部分图像相似度高: StarGAN-V2 的目标是进行图像的风格转换, 因此在某些情况下, 生成的图像可能与源图像非常相似, 只是在特定风格上有所变化. 模型不仅需要保留源图像的主要特征, 还要关注图像的细节特征.

2 改进的 StarGAN-V2 模型介绍

2.1 改进的 StarGAN-V2 模型组成

本文提出的改进 StarGAN-V2 模型是在 StarGAN-V2 的模型基础上提出的用于解决多域图像转换清晰度和自然度问题的一个模型, 有效提高了生成图像的质量和多样性, 其在未来广泛的应用前景和突出的性能使其成为图像生成和转换领域的重要工具. 主要架构由生成器、映射网络、风格编码器以及两个功能不同的鉴别器 (域反馈鉴别器和图像反馈鉴别器) 组成.

2.1.1 生成器 (generator)

改进后的 StarGAN-V2 模型的生成器结合了 U-Net 的编码器-解码器架构和 ResNet 的残差块. 而空间注意力机制通常通过卷积层操作, 对输入特征图中的不同位置进行加权处理. 在结合 U-Net 时, 从编码器传递到解码器的特征图中, 加入空间注意力模块, 该模块首先对输入特征图进行一次 3×3 卷积以捕捉局部空间信息, 随后通过一个通道叠加 (最大池化和平均池化) 生成二维注意力图, 再通过 1×1 卷积和 *Sigmoid* 激活函数生成一个权重图, 这个权重图与原特征图逐元素相乘, 以增强特征图中重要位置的响应值. 这样可以保留面部关键细节, 减少背景的干扰. 在残差块内部, 空间注意力机制可以被嵌入到卷积层之间, 在残差块的第一个卷积层后添加空间注意力模块, 生成一张空间权重图, 然后将这个权重图应用到卷积输出上. 这样做的效果是提升特征图中关键区域的权重, 使生成图像在特定风格转换中能更加细腻地保留关键细节[18]. 改进后的生成器的组成包括编码器、解码器、residual block 残差块、AdaIN 模块等.

2.1.2 映射网络 (mapping network)

映射网络是一个从潜在空间到风格空间的映射函数, 由具有多个输出分支的 MLP 组成, 作用是将一个简单的输入, 通常是随机噪声向量, 映射到高维的风格空间, 这些风格向量随后用于控制生成器中自适应实例归一化层, 从而实现对生成图像风格的多样性控制[19]. 首先通过将随机噪声向量映射到风格空间, 可以生成不同的风格向量, 使得生成器能够生成具有多样性和风格变化的图像, 之后在训练过程中学习不同域的风格特征, 可以生成不同域的风格向量, 从而控制生成器生成特定域的图像, 将风格空间与输入图像的内容空间解耦, 使得生成器能够独立地控制图像的内容和风格, 它允许从潜在空间中采样不同的风格特征, 从而生成具有多样性风格的图像[20]. 主要由输入层、全连接层和输出层组成.

2.1.3 风格编码器 (style encoder)

风格编码器的主要任务是从目标域图像中提取风格特征, 目标图像中特定的域被风格编码器提取出来后转换成一个低维的向量, 即是风格向量. 风格编码器在模型中发挥了重要的作用, 其所提取的风格向量包含目标域的风格特征, 如纹理、细节等, 之后根据风格向量指导生成器在生成图像时调整风格. 通过改变不同的目标域图像输入到风格编码器, 来生成不同的风格向量, 根据不同的风格向量, 生成器可以生成不同的符合不同风格的面部表情图像. 风格编码器增强了生成器的灵活性和多样性, 它的存在使得生成器更好地适应不同风格需求, 从而生成高质量的多样化图像. 风格编码器由特征提取层、风格特征提取层、全连接层等组成.

2.1.4 域反馈鉴别器 (domin feedback discriminator)

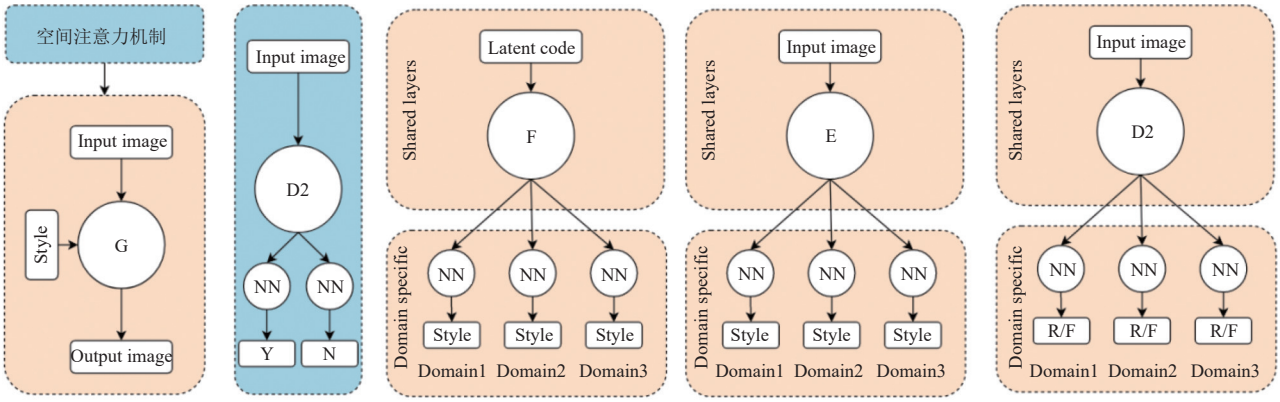
将循环一致性损失改进为对抗性循环一致性损失,

在 StarGAN-V2 的生成器后会引入一个额外的鉴别器, 定义为域反馈鉴别器. 其中域反馈鉴别器也是一个多任务的鉴别器, 包含 6 个预激活块, 这些块包含 *ReLU* 函数, 使用k个全连接层对生成器的每个域进行反馈. 在原始循环一致性损失中, 生成器需要确保当图像从原始域转换到目标域再回到原始域时, 生成的图像与原始图像尽可能相似. 这个损失的主要作用是保持生成图像的内容一致性, 避免模式崩塌和不合理的转换. 对抗性损失一致性函数的引入是为了提高模型的对抗性和鲁棒性. 具体来说, 通过引入额外的域反馈鉴别器, 可以增强生成图像的真实性, 提升循环一致性, 更好地输出反映特定域的图像.

2.1.5 图像反馈鉴别器 (image feedback discriminator)

图像反馈鉴别器的任务是负责判断输入的图像是否来自真实数据分布 (真实图像) 还是生成器生成的数据分布 (生成图像). 通过对抗训练与生成器相互作用, 帮助提升生成器生成逼真图像的能力, 其目标是通过与生成器对抗, 学习一个决策边界, 图像反馈鉴别器不断优化自身的判断能力, 以识别生成器生成的逼真图像, 这种竞争过程促使生成器生成更加逼真的图像, 同时使鉴别器能够更好地区分真实和生成的图像[21]. 鉴别器的反馈直接影响生成器的训练过程, 通过梯度反向传播和对抗损失, 生成器可以学习到更有效的图像生成策略[22]. 图像反馈鉴别器通常由卷积神经网络构成, 包括特征提取部分和判别部分.

改进后的 StarGAN-V2 模型组成与各部分组成如图 1, 其中蓝色部分为改进后新增的空间注意力模块与域反馈鉴别器.



(a) 生成器 (d) 域反馈鉴别器 (b) 映射网络 (c) 风格编码器 (e) 图像反馈鉴别器

图 1 改进的 StarGAN-V2 模型组成概述

该改进的 StarGAN-V2 模型使用对抗性目标训练整体框架, 给定图像$x \in X$和原始域$y \in Y$, 在训练过程中, 随机抽取一个潜在代码z和一个目标域$\tilde{y}$, 生成一个目标风格代码$\tilde{s} = F_{\tilde{y}}(z)$, 生成器以图像$x$和风格代码$\tilde{s}$作为输入学习并通过对抗性损失生成输出图像$G(x, \tilde{s})$.

$$L_{\text{adv}} = E_{x,y}[\log D_y(x)] + E_{x,\tilde{y},z}[\log(1 - D_{\tilde{y}}(G(x, \tilde{s}))] \quad (1)$$

其中, $D_y(\cdot)$表示鉴别器对应域y的输出, 映射网络学习提供可能在目标域y中的风格代码$\tilde{s}$, 生成器学习利用$\tilde{s}$生成与域$\tilde{y}$的真实图像无法区分的图像$G(x, \tilde{s})$.

本文还使用了一个风格重建损失来使生成器在生成图像$G(x, \tilde{s})$时使用风格代码$\tilde{s}$. 训练了一个编码器E来进行多域的不同输出, 在实验中学习的编码器允许生成器变换输入图像, 反映参考图像的风格.

$$L_{\text{sty}} = E_{x,\tilde{y},z}[\|\tilde{s} - E_{\tilde{y}}(G(x, \tilde{s}))\|_1] \quad (2)$$

为了使生成器能够生成多样性的图像, 使用多样性敏感损失显式正则化, 使生成器探索图像空间并发现有意义的风格特征以生成多样化的图像. 其中, 目标风格码$\tilde{s}_1$和$\tilde{s}_2$由映射网络在两个随机潜码z_1和z_2条件下产生 ($\tilde{s}_i = F_{\tilde{y}}(z_i)$).

$$L_{\text{ds}} = E_{x,\tilde{y},z_1,z_2}[\|G(x, \tilde{s}_1) - G(x, \tilde{s}_2)\|_1] \quad (3)$$

为了保证源特性, 改进后的 StarGAN-V2 模型采用了对抗性循环一致性损失, 引入了一个域反馈鉴别器, 这种损失函数的设计灵感来自于 CycleGAN, 它通过强制生成器进行正向和逆向转换[23], 使模型在多域转换任务中更加稳健, 会使图像生成质量更好. 其中$\hat{s} = E_y(x)$为输入源图像的估计风格代码, y为x的原始域, 通过鼓励生成器使用风格代码$\hat{s}$重构输入图像, 生成器学习改变风格的同时保留输入图像的原始特征.

$$L_{\text{adv\_cyc}} = E_{x,y,\tilde{y},z}[\|x - G(G(x,\tilde{s}),\hat{s})\|_1] - \lambda E_{x,y,z} D(y) \quad (4)$$

改进后的 StarGAN-V2 的完整目标函数可以总结为式 (5), 其中, λ_{sty}、λ_{ds}、$\lambda_{\text{adv_cyc}}$是每 1 项的超参数.

$$\min_{G,F,E} \min_{D} L_{\text{adv}} + \lambda_{\text{sty}} L_{\text{sty}} - \lambda_{\text{ds}} L_{\text{ds}} + \lambda_{\text{adv\_cyc}} L_{\text{adv\_cyc}} \quad (5)$$

2.2 面部表情转移流程

生成器生成的图像包括潜在引导合成图像和参考引导合成图像. 两种合成方式的风格转移过程分别如图 2、图 3 所示.

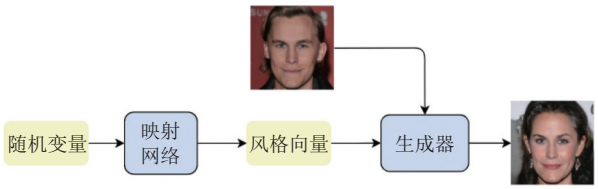


图 2 潜在引导合成

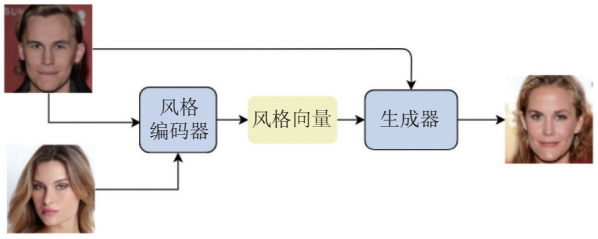


图 3 参考引导合成

潜在引导合成 (latent-guided synthesis) 是指利用潜在向量空间来控制生成图像的表情特征. 通过将目标表情的潜在向量输入风格编码器, 模型就可以根据这些潜在向量调整生成器的输出. 这种方法不仅允许直接控制生成图像的表情, 而且通过潜在空间的操作, 可以实现在不同风格和属性之间进行无缝转换, 从而实现更加灵活和个性化的图像生成.

另一种参考引导合成 (reference-guided synthesis) 是指使用一个参考图像, 即源图像, 来指导生成目标图像的过程, 从参考图像中提取风格特征, 并将这些特征应用到参考图像上. 首先使用风格编码器从源图像里提取风格特征, 提取到的风格特征包括发型、肤色等, 然后将提取的风格与参考图像融合生成新的目标图像. 这种方法通过改变参考图像, 可以灵活地控制生成图像的风格.

而空间注意力机制通过关注图像中特定的位置来增强对模型重要区域的感知. 本文的研究主要围绕人脸面部表情, 避免不了人脸细节方面的问题, 因此, 本文将空间注意力模块加入生成器的卷积层之后, 加入空间注意力机制之后的卷积层对输出的特征图进行加权, 更能突出图像中的重点区域, 帮助生成器更好地处理细节特征, 同时有效减少伪影的产生, 其中空间注意力机制模块如图 4.

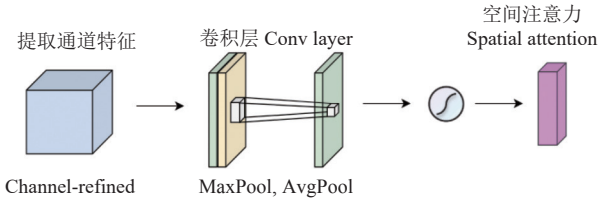


图 4 注意力模块

在改进后的 StarGAN-V2 模型中, 首先接收源域图像和目标域编码, 然后通过映射网络将随机噪声转换为风格向量, 并通过风格编码器将目标图像转换为风格向量. 生成器接收源图像和风格向量, 通过下采样和上采样块提取和重构特征, 生成目标域的图像. 生成的图像首先经过域反馈鉴别器, 域反馈鉴别器专门评估生成的图像是否准确反映特定的目标域, 从而提供反馈以优化生成器, 使其更好地生成具有目标域特征的图像. 接下来, 生成的图像再由改进的 StarGAN-V2 的图像反馈鉴别器进行评估, 图像反馈鉴别器负责区分生成图像与真实图像, 通过对抗训练进一步优化生成器的性能. 整个流程中, 使用改进的对抗性循环一致性损失函数, 确保生成的图像不仅逼真且在逆向转换回原始域时保持一致性. 通过这些改进, 模型能够更精确地生成多域图像转换结果, 并提升图像的质量和多样性, 模型流程图如图 5 所示.

2.3 实验评估

2.3.1 实验设置

本文使用 Python 语言和 PyTorch 框架进行实现并训练改进后的模型, 对原始数据进行了标准化处理, 生成器包含多个下采样和上采样块, 风格编码器和映射网络分别处理目标图像和随机噪声以生成风格向量. 引入一个域反馈鉴别器用于评估生成图像的特定域特征, 并通过对抗性循环一致性损失优化生成器, 同时使用图像反馈鉴别器进行图像真实性评估.

2.3.2 数据集设置

实验使用的数据集为 CelebA-HQ 数据集, 这是一个被广泛应用于人脸相关研究的高分辨率人脸图像数据集. CelebA-HQ 不仅提供了大量的人脸图像, 还包含

详细的人脸属性标注, 涵盖多种变化如性别、年龄、发型等, 每张图像都具有丰富的人脸表情和多样的细节特征. 在实验过程中, 我们则根据性别将 CelebA-HQ 数据集划分为两个主要域: 男性和女性. 所有图像均经过统一的 256×256 像素分辨率调整, 以确保在训练和测试过程中的一致性和有效性[24].

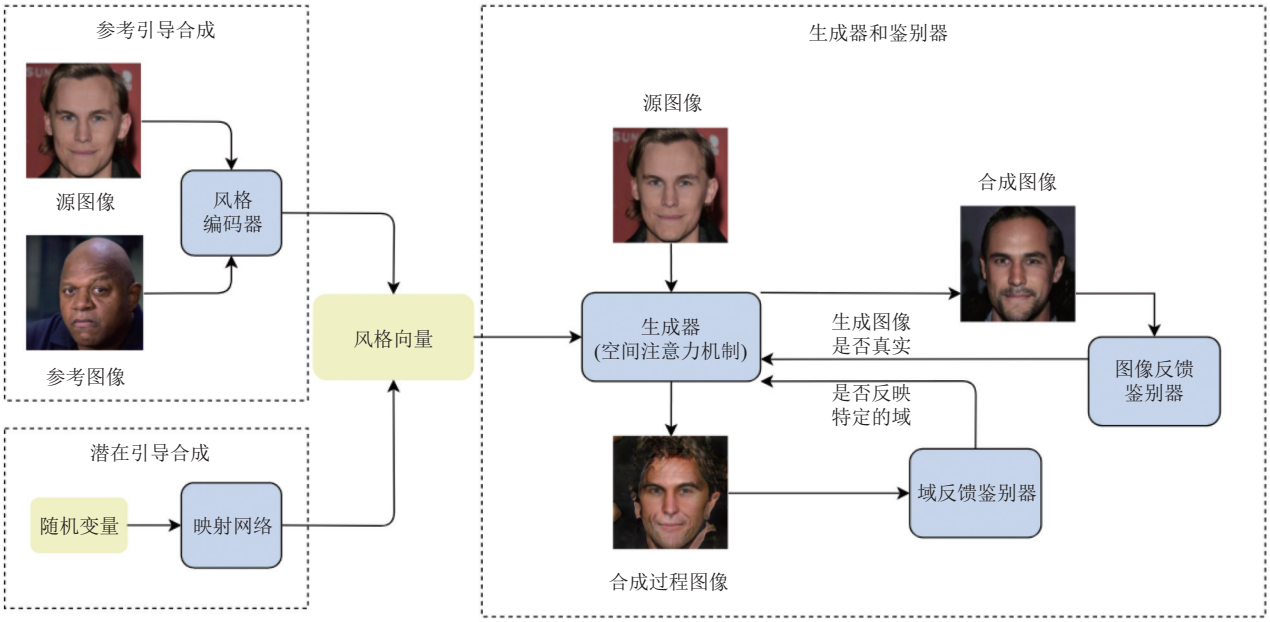


图 5 改进的 StarGAN-V2 模型流程图

在 CelebA-HQ 数据集的实验基础上, 本文还使用 FFHQ 数据集对模型进行测试以验证模型的鲁棒性和稳定性. FFHQ 数据集也是一个高质量人脸图像数据集, 主要用于训练和评估图像生成模型等任务, 相较于早期的 CelebA-HQ 数据集, FFHQ 在背景复杂度和多样性上等方面有一定提升, 涵盖了不同性别、种族、表情和发型, 包含了丰富的图像细节, 如光照和饰品等. 训练时经过筛选的图像经过统一的 256×256 像素分辨率调整来确保一致性.

2.3.3 评价指标与对比模型

本文的评价指标为 FID (Frechet inception distance) 值和 LPIPS (learned perceptual image patch similarity) 值[25].

FID 用来衡量生成图像和真实图像之间的分布差异, FID 值越低, 表示合成的图像与真实图像越接近, 所以在本文中 FID 值越小越好. LPIPS 用于衡量两个图像之间的感知相似度, LPIPS 值越低, 表示两种图像在感知上越相似, 本文生成图像与源图像越不相似越好, 即 LPIPS 值越高越好.

本文的对比模型采用经典的面部表情转移算法 MUNIT、DRIT、MSGAN 以及经典原始 StarGAN-V2, 并且加入了近年对 StarGAN-V2 改进的模型 StarGAN-V2Conv 进行对比, 所对比模型在上文均已经提到.

MUNIT 作为一种无监督图像到图像转换模型, 通过将图像表示分解为内容和风格两部分, 实现多模态生成. DRIT 与 MUNIT 类似, DRIT 通过内容-风格分离进行无监督图像转换, 并引入多样性损失以生成不同风格的图像. MSGAN 旨在控制生成图像的多样性, 通过在生成过程中保持模态多样性来提高模型的灵活性. StarGAN-V2 作为一种多域图像转换模型, 能够在有监督情况下进行多域、多模态的人脸图像生成. StarGAN-V2Conv 提出了一种 MDConv 注意方法, 嵌入具有多尺度融合功能的简化 PSConv 模块, 提高生成器特征提取能力. 将这些经典方法和改进的方法与我们的模型进行对比, 可以全面评估不同架构在多域面部表情转移中的表现, 尤其是在生成图像真实性、相似度方面的优势.

3 实验结果与分析

3.1 潜在引导合成结果分析

源图像被分为男性和女性两个域, 根据潜在向量生成的图像也包括男性和女性两种结果. 由于加入了

注意力模块和域反馈鉴别器来根据风格向量进行反馈, 最终合成的图像清晰度和自然度都有较大的提升, 潜在引导合成下的效果如图 6.

图 6 潜在引导合成效果图

在潜在引导合成过程中, 本文同时对比了与其他 5 个模型的图像视觉合成效果如图 7 所示, 从对比图中可以看出, 本文所提模型在图像的细节方面, 比如头发、面部纹路等方面均优于对比模型, 其中前 3 个模型虽然实现了表情转移, 但背景和各种纹理清晰度不足, StarGAN-V2 及 StarGAN-V2Conv 虽然实现了较好的转移效果, 但在头发、光影等细节方面有待提高, 本文方法在此基础上进行改进, 生成的图像更加清晰自然, 提高了真实性并降低了相似度.

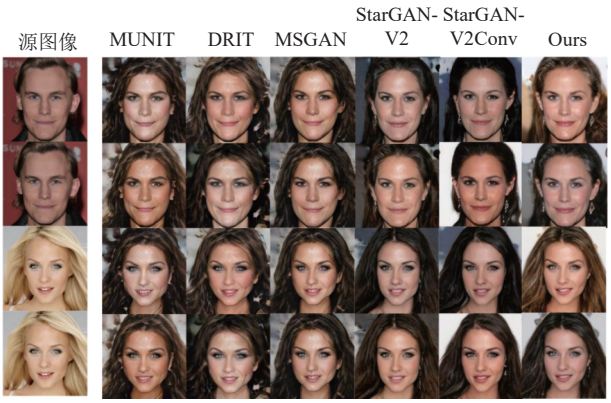


图 7 不同模型潜在引导合成对比

在潜在引导合成的实验中, 我们对比了 4 个经典算法与一个 StarGAN-V2Conv 模型在 FID 值和 LPIPS 值上的表现, 结果如表 1 所示. 结果表明, 改进后的 StarGAN-V2 模型在 FID 值上达到了 11.9, 明显低于其他模型, 较之原始 StarGAN-V2 模型, FID 值降低了 1.8, 相对于近年来对经典模型做出改进的模型 StarGAN-V2Conv 减少了 0.2, 验证了其生成图像质量更接近于真实图像. 同时, 该模型在 LPIPS 值上的表现也优于其他对比模型, 达到了 0.491, 较原始模型提高了 0.039, 相对于 StarGAN-V2Conv 提高了 0.04, 生成图像的感知相似度最高, 与源图像的视觉效果最不为接近.

我们在不同训练样本数量下, 将改进后的 StarGAN-V2 模型与 5 种方法模型的 FID 值与 LPIPS 值对比, 如图 8、图 9 所示, 展示了模型在 20 000–100 000 样本阶段的数据值变化.

表 1 潜在引导合成评价值分析

Method	FID	LPIPS
MUNIT	31.4	0.363
DRIT	52.1	0.178
MSGAN	33.1	0.389
StarGAN-V2	13.7	0.452
StarGAN-V2Conv	12.2	0.451
Ours	**11.9**	**0.491**
Real	10.8	—

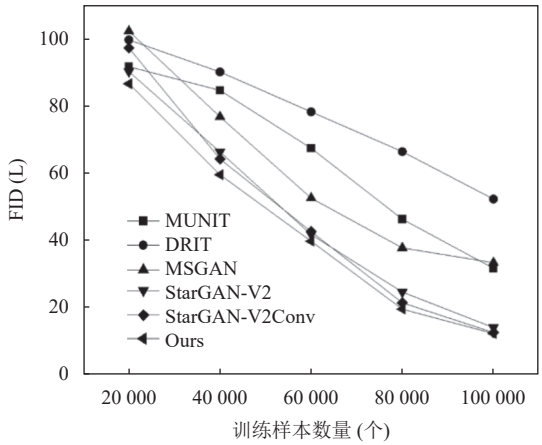


图 8 潜在引导合成 FID 值对比

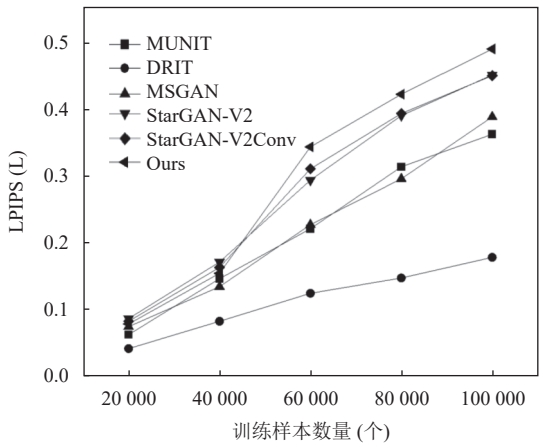


图 9 潜在引导合成 LPIPS 值对比

潜在引导合成中, 改进 StarGAN-V2 模型的 FID 值随着样本变化逐步下降, LPIPS 值则逐步升高, 当训练样本数量逐渐增加时, 改进后的 StarGAN-V2 模型在这两个指标上始终优于 MUNIT、DRIT、MSGAN

以及原始 StarGAN-V2 等模型.

3.2 参考引导合成结果分析

与潜在引导合成不同的是, 参考引导合成的策略利用了不同的参考图像, 可以将输入图像根据其性别分为 4 种转换情况: 男性转换为男性、男性转换为女性、女性转换为男性, 以及女性转换为女性. 这种方法利用参考图像的性别信息, 有助于生成更加符合目标性别的面部表情和特征, 参考引导合成效果如图 10、图 11 所示.

图 10 参考引导合成 (参考男性图像生成男性与女性)

图 11 参考引导合成 (参考女性图像生成男性与女性)

在参考引导合成过程中, 通过与 5 个模型对比展示了改进的 StarGAN-V2 模型在合成效果的优越性, 如图 12 所示. 可以看出, 改进后的模型在生成图像时克服了 MUNIT 和 DRIT 的失真问题, 相对于 MSGAN 提高了图像清晰度, 在细节方面优于 StarGAN-V2 原始模型及其变体, 主要体现在其细节比如头发、光影等方面, 在保持内容特征的同时, 在清晰度、自然性上有了一定的提升.

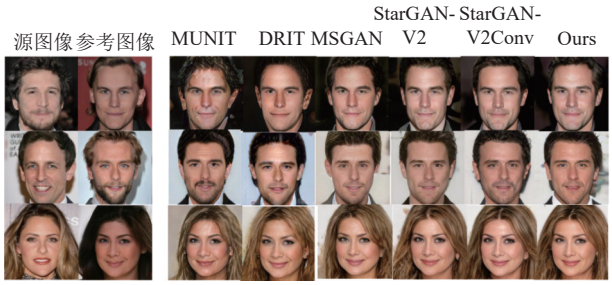


图 12 不同模型参考引导合成对比

参考引导合成中, 改进后的 StarGAN-V2 模型 FID 值达到 17.4, 比原始模型低了 6.4, 比 StarGAN-V2Conv 模型低了 2.8, LPIPS 值则达到了 0.426, 相对于比原始模型提高了 0.426, 比 StarGAN-V2Conv 模型高了 0.025, 均优于以上对比模型. 参考引导合成的 FID 值与 LPIPS 值与其他模型对比结果见表 2.

表 2 参考引导合成评价值分析

Method	FID	LPIPS
MUNIT	107.1	0.176
DRIT	53.3	0.311
MSGAN	39.6	0.312
StarGAN-V2	23.8	0.388
StarGAN-V2Conv	20.2	0.401
Ours	**17.4**	**0.426**
Real	10.8	—

参考引导合成在不同训练样本数量上的 FID 值与 LPIPS 值变化如图 13、图 14 所示.

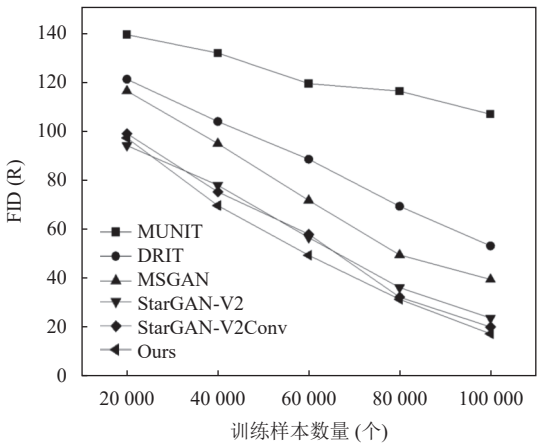


图 13 参考引导合成 FID 值分析

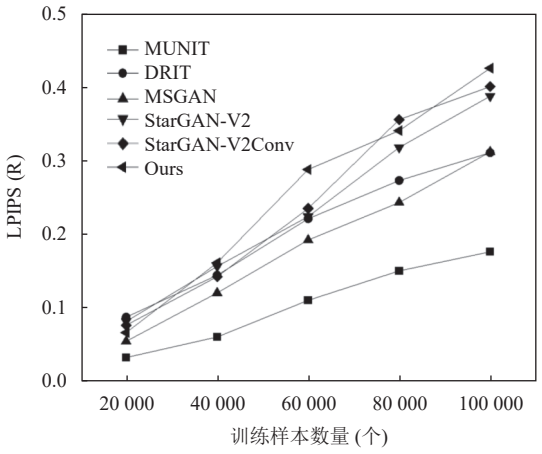


图 14 参考引导合成 LPIPS 值分析

在参考引导合成图像的方式中, 改进后的 StarGAN-V2 模型的 FID 值随着样本数量的变化逐步下降, LPIPS 值则逐步提升, 且随着训练样本数量的增加, 改进后的 StarGAN-V2 模型在这两个指标上的表现始终优于对比模型, FID 值稳定在 17.4, LPIPS 值为 0.426, 效果好于其他 5 个模型.

3.3 FFHQ 验证实验

在 CelebA-HQ 数据集上我们已经进行了训练和测试, 实验结果证明改进的模型的有效性, 一定程度上提高了面部表情转移生成图像的真实性, 有效降低了生成图像与源图像的相似度. 随后我们进一步利用 FFHQ 数据集对模型进行验证以评估模型的鲁棒性, 经过筛选之后选出适合表情转移的图像, FFHQ 数据集的高质量图像能够帮助我们更准确地评估生成图像的真实感和细节保留情况. 这一过程不仅增强了我们对模型在不同条件下表现的理解, 也为进一步的改进和优化提供了有价值的反馈. 潜在引导合成效果图如图 15, 参考引导合成效果图如图 16.

图 15 潜在引导合成

图 16 参考引导合成

可以看出所生成的面部表情图像，其真实性和清晰度都有很好的效果, 也很好地保留了图像的特征, 验证了改进的 StarGAN-V2 模型的稳定性和鲁棒性.

如表 3 所示, 在 FFHQ 数据集上, 潜在合成的 FID 和 LPIPS 为 12.2 和 0.473, 参考合成为 19.5 和 0.406, 说明模型在其他数据集上的有效性和鲁棒性.

表 3 模型在 FFHQ 数据集评价值分析

合成方式	FID	LPIPS
潜在引导合成	12.2	0.473
参考引导合成	19.5	0.406

3.4 消融实验

本文分别对加入空间注意力模块和加入新的域反馈鉴别器对模型的单独作用以及两种改进方法的共同作用进行对比, 生成的人脸面部表情结果如图 17 所示. 仅加入空间注意了模块之后, 在参考图像的特征方面表现更加突出, 生成的图像更加清晰, 但是在胡子头发等仍存在转换不明显的缺陷. 仅加入域反馈鉴别器之后, 增强细节的保留方面和多样性, 但是细节的清晰度不够高. 对比之下, 改进的 StarGAN-V2 生成的人脸表情图像中, 人脸的细节比如纹理、胡子等更加细节更加自然, 特征更加突出, 图像更真实.

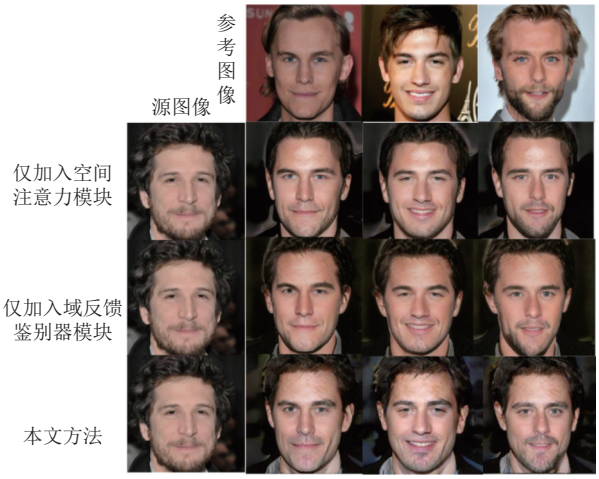


图 17 消融实验结果图

表 4 和表 5 所示为在潜在引导合成和参考引导合成的方式下 FID 值和 LPIPS 值的消融实验对比结果, 改进后的模型相对于仅加入空间注意力机制或者仅加入域反馈鉴别器的方法, FID 值更低, LPIPS 值更高, 即生成效果更好, 更符合真实图像.

表 4 潜在合成消融分析

方法	FID	LPIPS
仅加入空间注意力	12.4	0.451
仅加入新馈鉴别器	12.9	0.458
改进的模型	11.9	0.491

表 5 参考合成消融分析

方法	FID	LPIPS
仅加入空间注意力	22.4	0.404
仅加入新馈鉴别器	21.5	0.390
改进的模型	17.4	0.426

4 结论与展望

本文改进了 StarGAN-V2 模型, 专注于解决多域面部表情转移问题. 针对生成器, 本文引入了空间注意力模块, 增强生成器对图像细节区域的关注, 从而有效提高了生成图像的真实性. 为了能够在生成图像时更加能反映特定的域, 本文加入对抗性循环一致性损失, 引入了一个域反馈鉴别器, 能够在生成器生成的过程中不断提供域风格的反馈, 有效降低了源图像与生成图像的相似度. 并与近几年主流面部表情转移方法 MUNIT、DRIT、MAGAN、StarGAN-V2 模型进行对比, 同时加入近年对 StarGAN-V2 模型改进的 StarGAN-V2Conv 模型进行对比, 均取得了良好的结果.

尽管取得了良好的结果, 本文方法仍面临一些挑战, 如对非正常面部图像 (如戴墨镜、遮眼、侧脸等) 的风格迁移效果有限, 这些区域会干扰模型提取特征, 尤其是在风格编码和重构阶段, 难以分辨表情细节, 而且此类训练样本往往占少数. 效果图如图 18 所示, 未来的研究可以进一步优化编码器和生成器结构, 并寻找更多的训练样本以及探索更有效的约束条件. 进一步提升生成图像质量和适应性, 特别是处理复杂面部图像时的表现.

图 18 面部遮挡合成困难

参考文献

1 卢情义. 基于领域适应的跨角度面部表情图像生成和识别 [硕士学位论文]. 南京: 南京邮电大学, 2021. [doi: 10.27251/d.cnki.gnjdc.2021.000998]

2 何子亮. 基于生成对抗网络的多域图像转换技术研究 [硕士学位论文]. 广州: 广东工业大学, 2020. [doi: 10.27029/d.cnki.ggdgu.2020.001306]

3 张飞飞. 基于生成对抗网络的数据驱动人脸表情识别研究 [博士学位论文]. 镇江: 江苏大学, 2019.

4 田鹏飞. 基于生成对抗网络的不可感知水印攻击算法研究 [硕士学位论文]. 济南: 齐鲁工业大学, 2024. [doi: 10.27278/d.cnki.gsdqc.2024.000231]

5 Creswell A, White T, Dumoulin V, *et al*. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 2018, 35(1): 53–65. [doi: 10.1109/MSP.2017.2765202]

6 Fan Y, Jiang XG, Lan SX, *et al*. Facial expression transfer based on conditional generative adversarial networks. IEEE Access, 2023, 11: 82276–82283. [doi: 10.1109/ACCESS.2023.3294697]

7 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. 1125–1134.

8 Kim T, Cha M, Kim H, *et al*. Learning to discover cross-domain relations with generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR.org, 2017. 1857–1865.

9 Liu MY, Breuel T, Kautz J. Unsupervised image-to-image translation networks. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 700–708.

10 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. 2223–2232.

11 Wu WQ, Yin YJ, Wang YY, *et al*. Facial expression recognition for different pose faces based on special landmark detection. Proceedings of the 24th International Conference on Pattern Recognition. Beijing: IEEE, 2018. 1524–1529.

12 Tomar V, Kumar N, Srivastava AR. Single sample face recognition using deep learning: A survey. Artificial Intelligence Review, 2023, 56(S1): 1063–1111. [doi: 10.1007/s10462-023-10551-y]

13 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. 172–189.

14 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. 35–51.

15 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.

16 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. 8188–8197.

17 刘鹤. 基于改进 StarGANv2 人脸属性风格化编辑方法研究 [硕士学位论文]. 南昌: 江西师范大学, 2023. [doi: 10.27178/d.cnki.gjxsu.2023.001615]

18 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

19 Viazovetskyi Y, Ivashkin V, Kashin E. StyleGAN2 distillation for feed-forward image manipulation. Proceedings of the 16th European Conference Computer Vision. Glasgow: Springer, 2020. 170–186.

20 尹淑婷, 丁一峰, 王纯纯. 基于 StyleGAN2 的属性提取人脸交换方法. 重庆工商大学学报 (自然科学版), 1–12. (2024-07-02) [2024-09-16]. http://kns.cnki.net/kcms/detail/50.1155.N.20240702.1141.002.html.

21 Yang C, Lim SN. Unconstrained facial expression transfer using style-based generator. arXiv:1912.06253, 2019.

22 Qiao FC, Yao NM, Jiao ZR, *et al*. Geometry-contrastive GAN for facial expression transfer. arXiv:1802.01822, 2018.

23 Welander P, Karlsson S, Eklund A. Generative adversarial networks for image-to-image translation on multi-contrast MR images —A comparison of CycleGAN and unit. arXiv:1806.07777, 2018.

24 赵威海. 基于 StarGAN 实现的人脸风格图像清晰化研究 [硕士学位论文]. 阜阳: 阜阳师范大学, 2023. [doi: 10.27846/d.cnki.gfysf.2023.000244]

25 Kapoor P, Bui TD. TinyStarGAN v2: Distilling StarGAN v2 for efficient diverse image synthesis for multiple domains. Proceedings of the 32nd British Machine Vision Conference. BMVA Press, 2021. 69.

(校对责编: 王欣欣)