计算机科学
COMPUTER SCIENCE

结合小波变换高频信息的可控面部性别伪造

陈万泽, 陈家祯, 黄丽清, 叶锋, 黄添强, 罗海峰

引用本文

陈万泽, 陈家祯, 黄丽清, 叶锋, 黄添强, 罗海峰. 结合小波变换高频信息的可控面部性别伪造[J]. 计算机科学, 2023, 50(11A): 221000241-10.

CHEN Wanze, CHEN Jiazhen, HUANG Liqing, YE Feng, HUANG Tianqiang, LUO Haifeng. Controlled Facial Gender Forgery Combining Wavelet Transform High Frequency Information [J]. Computer Science, 2023, 50(11A): 221000241-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000241

结合小波变换高频信息的可控面部性别伪造

陈万泽　陈家祯　黄丽清　叶　锋　黄添强　罗海峰
福建师范大学计算机与网络空间安全学院　福州 350117
(chen_lex@163.com)

摘　要　基于生成对抗网络(Generative Adversarial Network,GAN)的图像到图像的翻译(Image-to-Image Translation,I2I)技术在各种领域中取得了一系列突破,并广泛应用于图像合成、图像着色、图像超分辨率,特别是在面部属性操作方面获得了深入研究。为了解决目前I2I领域由于模型架构以及数据不均衡所导致的不同翻译方向的生成图像性能表现差异的问题,提出了一种HFIGAN(High Frequency Injection GAN)模型,实现了结合高频信息的可控面部性别伪造。首先在结合高频信息的小波模块中,将编码特征通过离散小波变换进行特征级的分解,将所得到的高频信息在解码阶段对等注入,使得在上采样过程中的源域与目标域之间的信息可以达成平衡状态。其次,针对I2I任务中多域转换在不同方向的翻译难度不一致的问题,通过对损失函数进行重新设计,将难易样本的损失进行放缩,提高难样本对模型的反馈,使模型更专注于难样本的训练从而提升模型性能。最后,提出基于风格特征的多样性正则项,将风格向量在不同空间中的距离度量添加至传统的多样性损失中进行监督,使得模型能在保持生成图像多样性的同时提升图像的生成质量。分别在CelebA-HQ数据集和FFHQ数据集上进行实验并验证了所提方法的有效性。在主流的I2I模型中结合所提损失进行了损失函数通用性验证。实验结果表明,与以往先进方法相比,HFIGAN在面部性别伪造方面性能更加优异,所提出的损失函数具备一定的通用性。

关键词:图像生成;生成对抗网络;图像到图像的翻译;人脸属性编辑;聚焦损失
中图法分类号　TP391

Controlled Facial Gender Forgery Combining Wavelet Transform High Frequency Information

CHEN Wanze,CHEN Jiazhen,HUANG Liqing,YE Feng,HUANG Tianqiang and LUO Haifeng
College of Computer and Cyber Security,Fujian Normal University,Fuzhou 350117,China

Abstract　Image-to-image translation(I2I) technology based on generative adversarial networks has made a series of breakthroughs in various fields,and is widely used in image synthesis,image coloring,and image super-resolution,especially in face attribute manipulation. To solve the issue of disparity in the performance of generated images in different translation directions due to model architecture and data imbalance,an high-frequency injection GAN(HFIGAN) model is proposed to achieve controlled facial gender forgery for transmitting high-frequency information. Firstly,in the wavelet module for transmitting high-frequency information,the features in the coding stage are decomposed at the feature level by discrete wavelet transform,and the obtained high-frequency information is injected reciprocally in the decoding stage,so that the information composition between the source and target domains is always in a more desirable ratio. Second,images' dynamic consistency loss addresses the inconsistent translation difficulty in different directions for multi-domain conversion tasks in I2I. By redesigning the loss function,we scale the loss of difficult and easy samples,improve the feedback of difficult samples to the model,and make the model focus more on training difficult samples to improve performance. Finally,the diversity regular term based on style features is proposed to add the distance metric of style vectors in different spaces to the traditional diversity loss for supervision,which enables the model to maintain the diversity of generated images while improving the quality of image generation. Experiments on CelebA-HQ dataset and FFHQ dataset verify the effectiveness of the proposed method. The generalization of the loss function is verified in the mainstream I2I model combined with the proposed loss in this paper. Experimental results show that HFIGAN has better performance in facial gender falsification compared with previous advanced methods,and the proposed loss function has some generality.

Keywords　Image generation,Generative adversarial network,Image-to-Image translation,Facial attribute manipulation,Focal loss

基金项目:国家自然基金面上项目(62072106);福建省自然科学基金(2020J01168,2022J01190);福建省教育厅科学基金(JAT210053)
This work was supported by the National Natural Science Foundation of China(62072106),Natural Science Foundation of Fujian Province,China(2020J01168,2022J01190) and Scientific Research Fundation of the Education Department of Fujian Province,China(JAT210053).
通信作者:陈家祯(jiazhen_chen@fjnu.edu.cn)

1 引言

随着计算机视觉的发展及直播、短视频行业的兴起，面部图像编辑技术的应用逐渐深入日常生活，包括人脸美颜、虚拟美妆等，巨大的市场刺激了工业界对更高效的人脸编辑工具的研究。由 Goodfellow 等提出的生成对抗网络[1]（GAN）近年来取得了长足的进步，并涌现出诸多与 GAN 相关的变体。其中条件生成对抗网络（Conditional GAN，cGAN）[2]以标签作为额外输入实现了图像的可控生成，这一思想使得图像到图像的翻译（Image-to-Image Translation，I2I）成为可能。图像到图像的翻译[3]旨在学习不同视觉域之间的映射，即给定一个源域的图像，并将其渲染成具有目标域样式的图像。I2I 广泛应用于语义图像合成[4]、图像超分辨率[5]、面部属性编辑[6]等各种任务。CycleGAN[7]在训练中使用图像间的周期一致性损失进而实现了无监督训练的 I2I。同时，I2I 在面部图像中的应用也促成了深度伪造的发展[7]，即将输入图像依据属性标签进行包括局部属性以及全局属性的面部属性编辑，而原始图像的图像结构以及身份信息得以保留。

面部属性编辑的目标是向面部赋予特定属性，包括局部属性（如胡子、眼镜等）或全局属性（年龄、性别变化等）。然而，包括文献[7]在内的单域翻译模型不能学习多个域之间的相互转换，因此需要多个模型来匹配诸多的人脸特征。当前方法[8-10]为了解决多域转换问题，引入多个属性的风格特征作为附加输入。这些由风格特征指导合成的 GAN 模型展现了不错的性能，但是由于模型结构、数据不平衡以及损失函数设计的限制导致生成图像的质量有待提高。

首先，I2I 模型的输入为图像，故多以卷积自编码器（Convolutional Autoencoder，CAE）为基础结构。在映射过程中，由编码器将输入图像进行特征压缩，于较小的特征图中进行风格迁移，从而使输入图像的特征包含目标域的语义信息，最终通过解码获得翻译后的生成图像。但由于以往 I2I 方法在翻译过程使用自适应实例归一化[11]（Adaptive Instance Normalization，AdaIN）连续修改编码特征的统计量，导致生成图像中所蕴含的源域信息及目标域信息的比例失衡，即丢失了源域信息并过多地参考目标域风格，使得图像质量较差且丢失了图像的身份信息。其次，由于不同域之间的数据往往存在较大的差别，并且多域模型在表征数据的同时会受到数据分布不均所带来的影响，从而出现模式崩塌或风格偏好，导致某一条翻译路径性能表现较好而其他翻译路径下的生成图像质量较差。不同的翻译方向的难度也不尽相同，例如在男女性别伪造过程中，男性到女性方向的性能表现往往要比女性到男性的方向更好，传统方法并没有考虑到这一问题并提出解决办法。最后，以往在 I2I 领域解决 GAN 模式崩塌问题的方法是增加多样性损失[10,12]，通常损失考虑的只是最大化两张生成图像之间的距离，但由于 I2I 模型输出的图像是由生成器和风格特征共同决定的，当随机采样到的风格特征对相差不大时，此时输出的图像对也应较为相似，所以单纯最大化图像对之间的距离反而会对图像保真度产生负面影响。

针对上述问题，为了改进生成图像丢失源域信息以及过多参考目标域的问题，本文提出了融合小波高频信息的注入生成对抗网络（High Frequecy Injection Generative Adversarial Network，HFIGAN）。具体来说，在解码过程中，解码特征与经过高频注入模块 HFI 处理过的源域特征高频信息进行合并，经过处理后的特征包含源域图像大部分的结构信息以及细粒度特征，使得上采样过程中的源域与目标域之间的信息可以达成平衡状态。针对由于数据分布不均衡而导致模型出现数据偏好以及不同翻译方向的性能不一致问题，本文提出图像动态一致性损失进行改进，参考聚焦损失（Focal Loss）[13]处理难易样本的思路，本文对 I2I 领域中常用的循环一致性损失进行改进。即将性能表现较差的翻译方向视作难翻译方向，在此方向中的各个样本所对应的循环一致性损失值相较易翻译方向普遍较大，故将难翻译方向的映射任务归类为难样本，反之则为易样本。通过损失函数将难易样本的损失进行放缩，从而使模型在训练过程中在保持易样本表征的同时更加关注难样本的训练。最后，针对传统方法中多样性损失的问题，本文将风格特征之间的在不同空间中的距离纳入多样性损失的考虑范围，即使用向量之间的 L1 距离以及余弦相似度进行额外的监督，提出风格特征指导的多样性正则项。从理论分析可知，L1 距离越小，余弦相似度越大，则风格特征越相似，与同一张输入图像对应的生成图像也越相似，故在正则项中增加风格特征在不同空间中的距离度量，可使生成器在保持原有生成图像多样性的同时提升图像质量。

本文的主要贡献如下：

1）提出了 HFI 模块，即在生成器的样式注入阶段对等地添加经过离散小波分解析出的高频特征，以此减轻生成器合成丰富细节的压力，并均衡解码阶段特征源域信息以及风格特征的占比，使得生成图片在保持身份信息的同时较好地参考了目标域信息，最终实现了结合高频信息的可控面部性别伪造。

2）提出了图像动态一致性损失，通过损失函数的设计将难易样本的损失值进行动态控制，使得模型更加注重难样本的训练，解决了样本分布不均衡所导致的翻译方向难度不同的问题。

3）提出了基于风格特征的多样性正则项，向多样性损失中额外添加风格特征在不同空间的距离进行监督，在保证多样性的前提下进一步提升生成图像质量。

4）在 CelebA-HQ[14]以及 FFHQ[15]数据集中证明了 HFIGAN 具有优良性能，同时在其他 I2I 模型中额外验证了所提图像动态一致性损失以及基于风格特征的多样性正则项具有通用性。大量的实验证明了本文所提方法相比以往方法在性别伪造方面的有效性，与当前先进方法相比，所提方法展现出了较为出色的图像质量。

2 相关工作

2.1 面部属性编辑

面部属性编辑旨在操作面部图像的单个或多个属性，生成具有目标属性的面部图像，同时保留图像其他原生细节，其中属性代表面部图像共有的特征，例如性别、发色、肤色、年龄等。近年来，面部属性编辑工作依靠生成对抗网络取得了良好的效果。AttGAN[16]通过对生成的图像施加属性分类约束保证编辑属性的有效性，并引入重构学习使得生成过程中不需要改变的属性得到保留，同时采用对抗学习保证了生成图片的质量。当目标人脸图像的属性比较复杂时，以往方法在属性控制的精细化中存在缺陷，基于此，STGAN[17]通过基于

差分属性标签下的 STU(Selective Transfer Units)模块完成选择性传输面部特征，最终实现了人脸高精度属性的编辑。StarGAN[18]为基于属性标签的合成模型，其使用单个模型实现多个域的转换。基于 GAN 强大的生成能力，面部属性编辑可以使用图像到图像翻译得以实现。L2M-GAN[19]将潜在空间中的特征向量分解为属性相关和属性无关，用风格转换器对分解后的属性相关向量和无关向量施加正交约束。MUNIT[8]同时使用内容特征以及风格特征来表征图像从而完成图像翻译的任务。PHStarGAN-v2[20]利用超复代数(Hypercomplex Algebra)来实现 I2I 生成模型，在确保图像质量以及翻译性能的前提下，减少了模型的参数量。Liu 等[21]针对图像在域内和域间插值的平滑性以及对风格空间的解耦提出了改进。

2.2 循环一致性在 I2I 中的应用

由于图像到图像翻译的数据多以无配对情况出现，故以循环一致性评估两个或多个样本之间的匹配程度是无监督 I2I 中常用的技术，其已被应用于许多计算机视觉任务，如图像对齐[22]、深度估计[23]、人脸属性编辑[6]等。如何利用循环一致关系学习到不同领域间的双向转换模型一直是 I2I 研究的热点问题，Hoffman 等[24]提出循环一致性对抗域自适应模型，其同时在像素级和特征级应用循环一致性，增强了模型对原始图像结构以及语义内容的保留。Zhu 等[25]使用循环一致性解决了在无成对数据时的 I2I 任务。

2.3 聚焦损失函数

Lin 等[13]发现图像会出现前景以及背景不平衡的现象，导致图像在目标检测的结果中不理想，他们进而提出聚焦损失来改善性能，聚焦损失将模型专注于难分类样本，增加难分类样本对模型的反馈从而提升性能。之后的一些工作在其他样本不平衡的状态下使用聚焦损失[26-28]。Ridnik 等[29]将聚焦损失应用于多标签分类，并提出不对称损失，该变体在多标签分类中比原始的聚焦损失具有更好的性能。Smith 等[30]提出了周期聚焦损失(Cyclical Focal Loss)，在网络的较早阶段更新时更多接收到易样本的反馈，难样本的训练在训练中间阶段之后进行，以此提高网络整体的泛化能力。

3 HFIGAN 方法

3.1 整体结构

假设 X 和 Y 分别对应图像和图像所属视觉域的集合，给定一个图像 $x_o \in X$ 以及对应的域标签 $y_o \in Y$ 和一个任意的目标域 $y_t \in Y$，则模型的任务是训练一个生成器 G(Generator)将 x_o 转换为拥有目标视觉域风格的图像 x_t。架构设置了映射网络 M(Mapping Network)和采样网络 S(Sampling Network)分别从参考图像 x_r 以及噪声 $\boldsymbol{z}$ 中获取风格特征，为 G 在执行过程中提供翻译依据。同时，HFIGAN 符合传统生成对抗网络的基本结构，添加鉴别器 D(Discriminator)对 G，M，S 进行对抗训练。整体结构如图 1 所示。

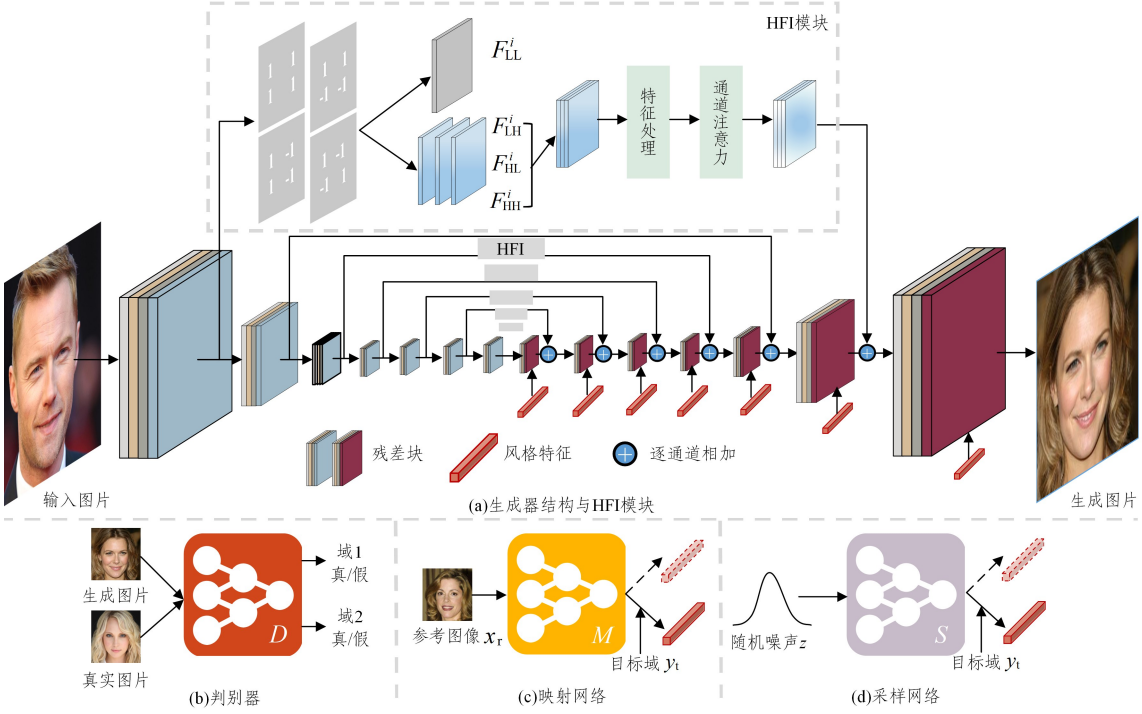

图 1 模型结构图

Fig. 1 Model structure

3.1.1 生成器

生成器结构如图 1(a)所示，给定特定域的风格特征 v，生成器 G 的任务是将输入的源域图像 x_o 翻译为拥有特定域语义 x_t，翻译的整体过程可抽象为 $x_t = G(x_o, v)$，其中 v 来源于映射网络 M 或采样网络 S。生成器模型的整体符合自编码器结构，为了完成映射过程，首先在编码阶段将图像压缩至较小的特征中，再输入至解码过程中对每一个阶段特征使用自适应实例归一化(AdaIN)进行样式注入，如下所示：

$$AdaIN(F^i_{De}, v) = \sigma(v)\left(\frac{F^i_{De} - \mu(F^i_{De})}{\sigma(F^i_{De})}\right) + \mu(v) \tag{1}$$

其中，F^i_{De} 为第 i 阶段解码残差块的输出，$i \in \{1, \cdots, n\}$，n 为残差块总数；$\sigma(\cdot)$ 和 $\mu(\cdot)$ 分别指标准差函数以及均值函数。通过式(1)修改特征各个通道的统计值从而达到风格迁移的目的。为了保护源域的身份信息并使图像可以合理地参考目标域的语义信息，本文在解码过程中的每一个样式注入阶段对等地输入了经过 HFI 模块处理后的源域特征。在 HFI

模块中，使用离散小波变换对编码后的源域图像信息进行筛选，由于高频信息包含了图像大部分的结构信息，故本文中选取高频特征用于后续处理，如图 1(a)中灰色框所示。生成器的整个过程如下所示：

$$F_i = G_{\mathrm{En}}^{i}(x_o) \tag{2}$$

$$F_i^{\mathrm{HL}}, F_i^{\mathrm{LH}}, F_i^{\mathrm{HH}} = \Omega_{\mathrm{DWT}}(F_i) \tag{3}$$

$$F_i^{\mathrm{HFI}} = \Phi(F_i^{\mathrm{HL}}, F_i^{\mathrm{LH}}, F_i^{\mathrm{HH}}) \tag{4}$$

$$x_t = G_{\mathrm{De}}^{n}(\hat{F}_{n-1} + F_{n-1}^{\mathrm{HFI}}), \hat{F}_1 = G_{\mathrm{De}}^{1}(F_n) \tag{5}$$

其中，G_{En}^{i} 以及 G_{De}^{i} 分别代表第 i 块编码残差块以及解码残差块；F_i 与 $\hat{F}_i$ 为输入至编码和解码残差块中的特征；F_i^{HL}，F_i^{LH}，F_i^{HH} 为离散小波变换在特征级别分解出的高频特征信息；F_i^{HFI} 为 HFI 模块的输出；Ω_{DWT} 代表离散小波变换操作，Φ 指后续的特征处理以及通道注意力。编码残差块与解码残差块结构相同，前者使用实例归一化，后者替换为自适应实例归一化。特征处理由卷积层、实例归一化以及激活层所构成的小模块级联两组组成。

3.1.2 鉴别器

鉴别器如图 1(b)所示，本文中使用的 D 是多任务鉴别器，输入的真假图像经过残差块降维后使用平行线性层使鉴别器拥有多个输出，其中每一个输出都是二值分类，意在判断输入图像真假的同时还可以对其所属域进行分类，从而确定输入图像是其语义域内的真实图像还是 G 生成的假图像 x_t。

3.1.3 采样网络和映射网络

给定从随机噪声中采样的向量 $\boldsymbol{z}$ 和任意域 y_t，设计拥有多输出的采样网络 S 来产生风格特征 $v = S_{y_t}(\boldsymbol{z})$，其中 $S_{y_t}(\cdot)$ 表示根据标签 y_t 选择的输出分支，如图 1(d)所示。S 由一个多层感知器(Multilayer Perceptron，MLP)外加多个输出分支组成，它可以为所有可用的视觉域提供样式向量。MLP 的目的是学习域间的公共样式信息。分支结构包括 3 个全连接层，学习特定于域的样式信息，最终输出预先设定维度大小的风格特征向量。映射网络 M 的输入为任意参考图像 x_r 及对应域标签 y_t，其过程表示 $v = M_{y_t}(x_r)$，其中 $M_{y_t}(\cdot)$ 为域 y_t 条件下 M 的输出，如图 1(c)所示。映射网络与鉴别器的多任务设计结构类似，M 可以从不同的参考图像中生成各种风格向量。映射网络主要包括多个级联的下采样块以及全连接层，前者同样是学习域间的公共样式信息，线性层则输出特定域的风格向量。

3.2 损失函数

本文通过损失设计，使映射网络和采样网络进行风格特征的精准获取，而生成器负责合成多样的图像，鉴别器负责与采样网络、映射网络以及生成器构成的图像输出系统进行对抗训练进而提升整体性能。所使用到的损失有对抗损失、图像动态一致性损失、图像动态重构损失、基于风格特征的多样性正则项以及风格重构损失。本文模型的整体损失函数可以总结为：

$$F(G,D,M,S) = \min_{G,M,S}\max_{D} L_{\mathrm{adv}} + L_{\mathrm{dc}} + L_{\mathrm{dr}} + L_{\mathrm{vr}} - L_{\mathrm{vms}} \tag{6}$$

其中，L_{adv}，L_{dc}，L_{dr}，L_{vr} 以及 L_{vms} 分别对应式(7)、式(8)、式(9)、式(13)以及式(12)。为了验证本文所提损失函数具有鲁棒性，需将损失函数添加至其他方法中进行验证。为了实验的公平性，我们没有刻意地调整不同损失函数的权重配比，故在所有实验中损失函数的权重均设定为 1。

3.2.1 对抗损失

作为 GAN 网络，对抗损失的作用是提升图片生成质量并保证其真实度。为了鼓励生成的图像尽可能接近数据集中包含的真实图像，本文使用非饱和性的对抗损失[1]，如下所示：

$$L_{\mathrm{adv}} = \log D_{y_o}(x_o) + \log(1 - D_{y_t}(G(x_o, v))) \tag{7}$$

如图 2(d)所示，在训练过程中，生成图像由参考图像 x_r 和潜在码 $\boldsymbol{z}$ 交替指导生成，故 v 包含映射网络和采样网络的两种输出。$D_{y_o}(\cdot)$ 以及 $D_{y_t}(\cdot)$ 表示鉴别器对于输入图像所属域的判断。最终生成器通过优化式(7)的对抗损失从而生成高质量 x_t。

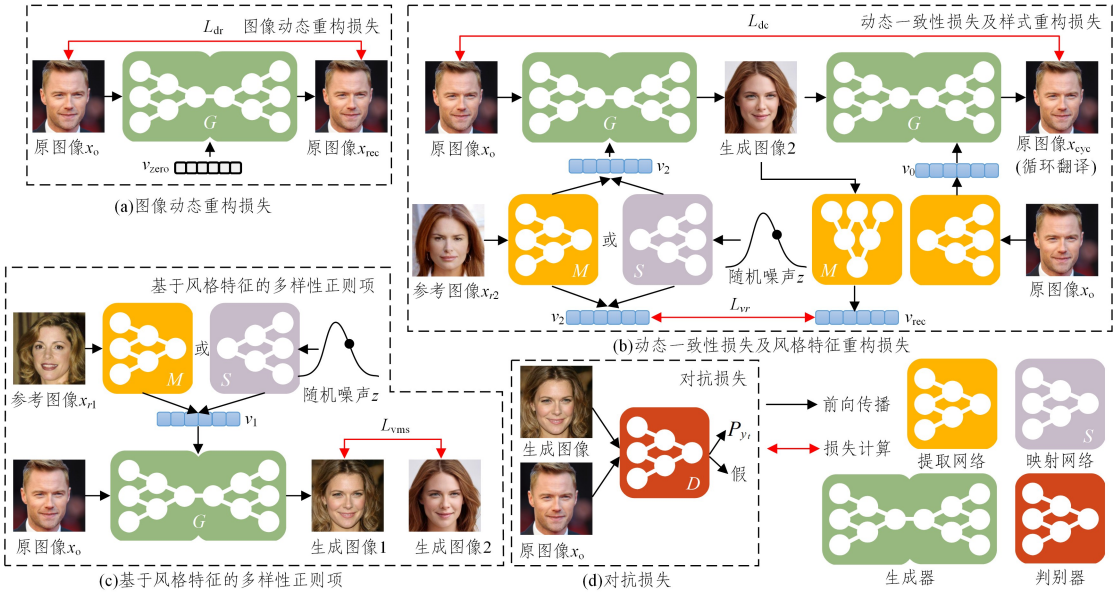

图 2 训练损失流程图

Fig. 2 Diagram of training loss

3.2.2 图像动态一致性损失

为了确保生成图像 x_t 能适当保留源域的内容特征(如姿态、表情等)，并使生成器在非成对数据的情况下重构输入图像 x，本文使用图像动态一致性损失(Dynamic Consistency

Loss)对生成器进行约束。在对传统方法的结果进行分析时发现，不同域数据的不平衡，例如在 CelebA-HQ 数据集中，女性域的图像数量相比男性域要多 71.3%，使得生成器容易出现数据偏向进而导致不同的翻译方向表现出不一致的性能表现，同时在客观指标上也显示男性域至女性域的翻译方向性能表现较好。本文参考聚焦损失[12]针对目标检测中样本不均衡的处理方法，将其思想应用于 I2I 图像间的 $L1$ 损失，提出动态循环一致性损失：

$$L_{\mathrm{dc}}=-\alpha\log\left(1-\left(\frac{\|x_o,x_{\mathrm{cyc}}\|_1}{2}\right)\right) \tag{8}$$

其中 $\|\cdot\|_1$ 表示 $L1$ 范数，α 为调制系数，x_{cyc} 为循环翻译后的图像。流程如图 2(b)所示，为实现循环一致性，图像必须要进行方向相反的两次翻译，故不可避免地会受到难路径的影响从而导致性能下降，因此本文将循环一致样本对之间 $L1$ 距离较大的认定为难样本，反之则为易样本。图像动态一致性损失的目的是消除生成器由于数据不均衡而出现的数据偏好问题。式(8)将传统方法使用的 $L1$ 损失进行放缩，如图 3(a)所示，黑色直线为传统的 $L1$ 损失，彩色曲线为经过式(8)放缩后的动态一致性损失，图 3(b)彩色曲线说明了动态一致性损失与 L1 损失之间的放缩倍率，高于黑色基准的部分表示损失被放大，α 的增加则代表认定为难样本的阈值变小。经过实验本文将 α 设定为 1.5。模型初始时损失大部分集中于 0.75 至 1.5 之间，经过动态一致性损失的处理，难样本的损失相比易样本放大了 2.23 倍，使得模型在训练过程中在保持易样本表征的同时更加关注难样本的训练。

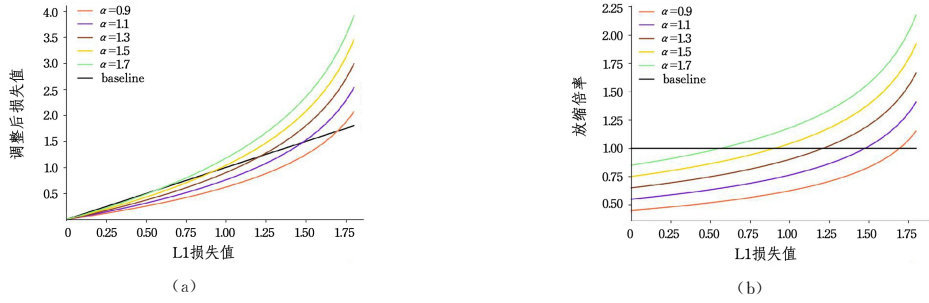

图 3　图像动态一致性损失曲线图(电子版为彩图)

Fig. 3　Curve of dynamic consistency loss

3.2.3　图像动态重构损失

为了生成器生成的图像可以更好地保留源域图像的部分特征以及面部身份信息，本文添加了图像动态重构损失(Dynamic Reconstruction Loss)，如图 2(a)所示。在翻译的过程中，将输入的风格向量替换为零向量，即解码阶段中的特征统计量不进行变化，从而重构输入图像。损失可表示为：

$$L_{\mathrm{dr}}=-\alpha\log\left(1-\left(\frac{\|G(x_o,v_{\mathrm{zero}}),x_o\|_1}{2}\right)\right) \tag{9}$$

其中，$\|\cdot\|_1$ 表示 1 范数，α 为调制系数且设定为 1.5，v_{zero} 代表维度与风格特征相同的零向量，也即无风格特征。依旧使用动态一致性损失进行一定比例的放缩，使得生成器可以更好地专注于难样本的训练。

3.2.4　基于风格特征的多样性正则项

多样性损失在 I2I 任务中是为了迫使 G 探索更多的样式生成能力以及防止模式崩塌，传统方法[10,12]的多样性损失为：

$$L_{\mathrm{ms}}=\max_G\left(\frac{L_I(G(x,z_1),G(x,z_2))}{L_z(z_1,z_2)}\right) \tag{10}$$

$$L_{\mathrm{ds}}=E_{x,y_i,z}[\|G(x,v_1)-G(x,v_2)\|_1] \tag{11}$$

针对式(10)，如图 4(b)所示，z_1，z_2 随机采样于正态分布且作用类似于本文的风格特征，$L_I(\cdot)$ 为图像间的距离，$L_z(\cdot)$ 为采样向量的距离。当采样点相距较近时，此时分母过小从而使得 L_{ms} 的损失值急剧增大，影响模型训练的稳定性，导致生成效果变差。针对式(11)，即简单最大化 $G(x,v_1)$ 和 $G(x,v_2)$ 这两个由不同风格向量指导生成的图像之间的距离，以便生成器 G 尝试探索更多样的特征空间，如图 4(c)所示。但当风格特征 v 差异较大时，生成的图像应如同图 4(c)的 I_1 与 I_2，图像之间的 $L1$ 距离应较大。但当 I_2 与 I_3 由风格较为相似的两个向量 $\boldsymbol{v}$ 指导合成时，两者之间的图像差距应较小，此时最大化两者之间 $L1$ 距离可能会使 G 探索到不属于原有语义域的特征空间，从而影响图像之间的生成质量。本文认为在基于风格样式的 I2I 任务中不应该只是简单地最大化图像之间的距离，而是要额外考虑风格特征之间的差距。针对此问题，提出了基于风格特征的多样性正则项，如式(12)所示：

$$L_{\mathrm{vms}}=\frac{1}{N}\sum_i^N\frac{L_I^i(G(x_i,v_1^i),G(x_i,v_2^i))\cdot L_v^i(v_1^i,v_2^i)}{d_{\cos}^i(v_1^i,v_2^i)} \tag{12}$$

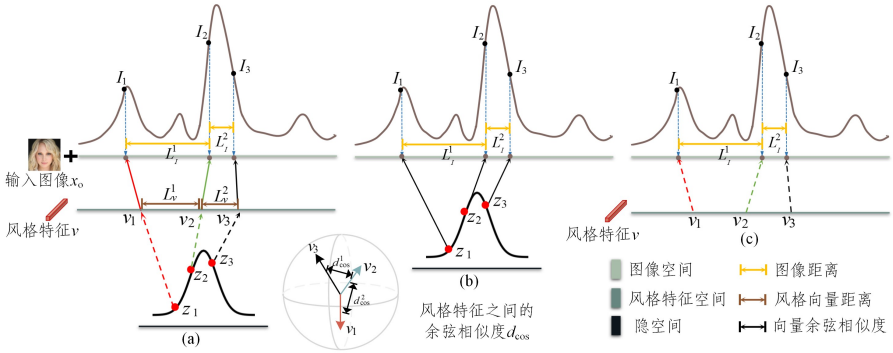

图 4　不同的多样性损失示意图

Fig. 4　Diagram of different diversity loss

损失计算流程如图 2(c)所示。式(12)中，$L_I^i(\cdot)$为图像间像素级 $L1$ 距离，$L_v^i(\cdot)$为风格特征间向量级 $L1$ 距离，$d_{\cos}^i(\cdot)$表示向量之间的余弦相似度，N 为批量大小，x_t^i 为某一个批量内的第 i 个生成图像，v_1^i 与 v_2^i 为两种不同的风格特征。如图 4(a)所示，因 x_t 是根据原图像以及风格特征合成而来，此时相关距离有 3 种，分别是图像距离 L_I、风格向量距离 L_v 以及样式向量余弦相似度 $d_{\cos}$。其中 L_v 和 $d_{\cos}$衡量了风格向量在不同空间中的度量。从理论上来说，L_v 越大，$d_{\cos}$越小，说明向量所蕴含的样式风格差别越大，反之则越小。因需要最大化 L_{vms}正则项，故将图像 $L1$ 距离设计为与风格特征 $L1$ 距离成正比，与 $d_{\cos}$成反比。且基于风格特征的多样性正则项不同于式(10)，其中由于潜在码 $\boldsymbol{z}$ 的维度较低，导致 $\boldsymbol{z}$ 之间的差值相比图像之间的距离有量级上的差距，使得式(10)的损失偏大从而影响模型训练的稳定性。综上，L_{vms}可以在保持多样性的同时，提升 I2I 模型的图像生成质量，并且相比传统方法没有添加额外的计算，可以轻易嵌入到不同的 I2I 框架中。

3.2.5 风格特征重构损失

风格特征是生成器在翻译过程中的唯一依据，为了确保生成器充分利用风格特征 v 的信息来生成图像 x_t，并消除采样网络 S 输入 $\boldsymbol{z}\sim N(0,1)$的不确定性，这里应用了风格特征重构损失，可表示为：

$$L_{\text{vr}}=\| M(G(x_o,S(\boldsymbol{z},y_t)),y_t)-S(\boldsymbol{z},y_t)\|_1 \tag{13}$$

其中，$M(G(x_o,S(\boldsymbol{z},y_t))$是重构后的风格特征。由式(13)可知，通过具有确定输入 M 的监督，最小化 L_{vr}消除了 $\boldsymbol{z}$ 对映射网络的负面影响，确保 $\boldsymbol{z}$ 映射到的特征是理想的目标域风格。同时最小化 L_{vr}也可迫使 G 应用风格特征进行图像翻译。

4 实验结果与分析

4.1 实验设置

4.1.1 实验环境及参数设置

本文实验在 NVIDIA Tesla V100S GPU 32 GB 环境中进行训练，并选择 Adam 优化器，其中 Adam 的两个相关参数设置为 $\beta_1=0$，$\beta_2=0.99$。将所有训练图像调整至 256×256 分辨率，输入图像的批量大小设置为 8，模型迭代 100 000 次。在每个训练步骤中，鉴别器和生成器交替训练。生成器、鉴别器、采样网络以及映射网络的学习率均为10^{-5}。

4.1.2 数据集

本文选择在 CelebA-HQ[14] 以及 FFHQ[15] 两个具有高分辨率人脸属性数据集上进行训练和测试。其中 CelebA-HQ 包含 3 万张面部图像，其中女性图像数量为 18943 张，男性图像为 11057 张，数据集分辨率为 1024×1024。为了验证模型在性别属性伪造任务中具有鲁棒性，本文额外在年龄、种族和图像背景等方面风格多样且差异明显的 FFHQ 数据集上进行完整训练以及验证，其中 FFHQ 数据集包含 1024×1024 分辨率的 7 万张 PNG 格式高清人脸图像。在 CelebA-HQ 数据集和 FFHQ 数据集的两个域中分别随机选择 1000 张以及 3000 张图像作为测试集。

4.1.3 评价指标

为了定量评估本方法的性能，本文采用两个指标：弗雷歇起始距离[31](Fréchet Inception Distance，FID)以及特征度量图像相似度[32](Learned Perceptual Image Patch Similarity，LPIPS)。其中，FID 用于测量两个数据分布之间的差距，即衡量生成图像的真实度，FID 值越低，生成图像的质量越高。LPIPS 用于计算来自同一属性图像对的相似性，若生成图像具有更好的多样性，则 LPIPS 值越大，反之亦然。为了实验的公平性，本文使用了 StarGAN-v2[10] 官方代码中 FID 与 LPIPS 的计算方法。本次实验对比的模型有 MUNIT[8]，DRIT++[9]，StarGAN-v2[10]，PHStarGAN-v2[20]。对于所有的对比模型，本文使用其官方的实现方案以及推荐的超参数进行训练，最终使用同样的方法进行评估。在损失通用性实验中，本文将所提损失嵌入至其他方法并保持原有设置进行完整训练，最后与原方法在同一数据集中进行对比。

为了评估不同模型在生成过程中面部图像的身份信息的保持程度，本文设计了身份信息一致性实验，使用 VGGFace[33] 模型衡量原图像 x 与伪原图像 x' 两者特征之间的距离，其中 $x'=G(x,M_y(x))$，即原图像根据自身的风格特征生成的面部图像。VGGFace 是在具有 260 万张图像的大型的人脸数据集上进行训练，并且对外观变化(例如光照、年龄等)具有鲁棒性，使用 VGGFace 模型可以客观地衡量两幅图像的身份信息。

4.2 实验结果

4.2.1 基于潜在引导生成比较

潜在引导生成即生成器根据采样网络所获得的风格特征 $v=S_{y_t}(\boldsymbol{z})$进行图像合成。图 5 与图 6 为各方法分别在 CelebA-HQ 数据集以及 FFHQ 数据集中基于潜在引导生成的图像，其中(a)－(e)分别为 MUNIT[8]，DRIT++[9]，PHStarGAN-v2[20]，StarGAN-v2[10] 以及本文方法，图像左边为潜在引导生成的男性到女性的翻译方向，另一边则相反。

首先在图 5 中，图 5(b)所生成的图片基本集中于同一个样式，说明多样性较差，且图像并不符合对应目标域的语义要求，面部信息丢失严重，同时图像伴有较多伪影。图 5(a)所翻译的图像虽然基本符合目标域的语义要求，但多样性也较差，且伪影较多导致图像质量较差。相比图 5(a)与图 5(b)，图 5(c)和图 5(d)的图像多样性有明显提升，但由于文献[20]只能生成其他方法一半分辨率的图像，故图像(c)整体质量较差，且背景较为杂乱，头发以及面部细节处存在伪影。图5(d)中生成图片多样性较好，且生成图像符合目标域的语义要求，但在头发以及背景处出现不规则的形状，同时面部也没有捕获到源域图的细节。相比其他方法，HFIGAN 在图 5(e)中展现出了较好的翻译性能，细节表现上更加细腻，伪影较少且图像背景没有受到源域特征的干扰，显得更加整洁自然清晰，多样性以及图像质量在视觉效果方面处于较优水平。

在图 6 中选取了具有较大范围遮挡的面部图像。其中图 6(a)、图 6(b)对应的生成图像依旧存在遮挡区域的内容从而形成伪影，使得图像不符合目标域的语义要求，同时图 6(b)中图像多样性依旧较差。图 6(c)、图 6(d)相比图 6(a)、图 6(b)展现出了更好的图像多样性，但是由于遮挡部分的影响，图 6(d)方法生成的图像出现内容缺失，从而导致头部畸形，且头发不自然，背景有较大面积伪影，图 6(c)方法由于受到图像分辨率的影响，整体质量较差，背景杂乱，且因为过于参考目标域样式，部分图像丢失了源域的身份信息，出现了面部角度改变以及面部信息缺失。图 6(e)中图像展现出了较好的多样性以及图像质量，图像并没有受到左边源域图大面积遮挡的影响，依旧翻译出了符合目标域语义的图像，这说明 HFIGAN 对遮挡图像具有鲁棒性。

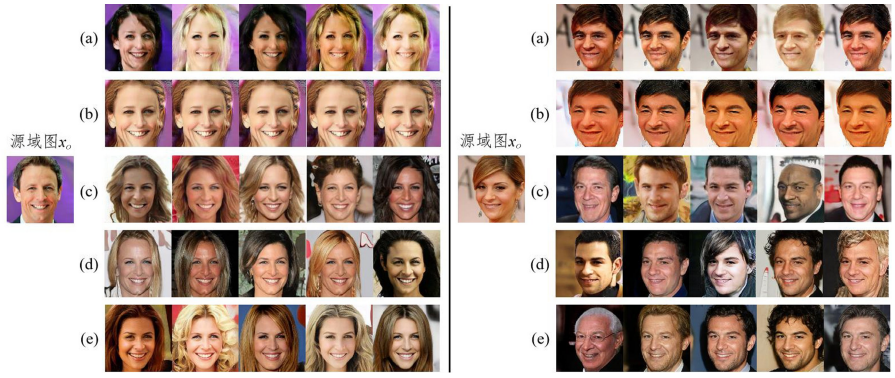

图 5　CelebA-HQ 数据集中基于潜在引导生成的图像

Fig. 5　Images generated based on latent guidance in CelebA-HQ dataset

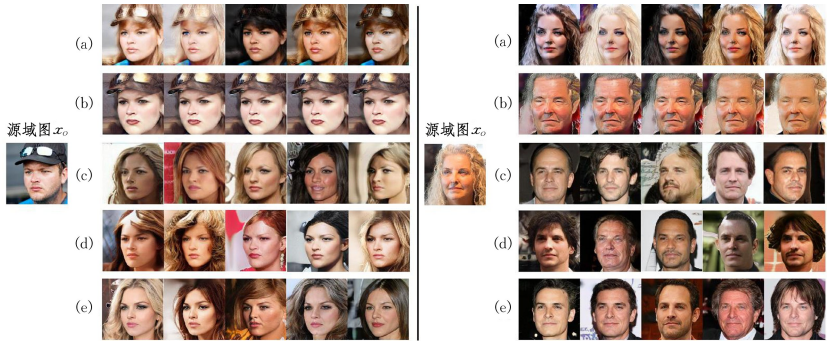

图 6　FFHQ 数据集中基于潜在引导生成的图像

Fig. 6　Images generated based on latent guidance in FFHQ dataset

表 1 列出了各个方法在 CelebA-HQ 数据集以及 FFHQ 数据集中基于潜在引导生成图像的 FID 与 LPIPS 客观指标数据。其中 FID 越小越好，LPIPS 则相反。FID 数据显示本文方法表现较佳，说明本文方法翻译后的图像的数据分布更接近于真实图像。另一方面，本文得到了最高的 LPIPS 分数，这证明了本文模型在给定单一的输入图像时，可以生成更多样化的生成图像。

表 1　对不同方法基于潜在引导生成的图像的定量比较

Table 1　Quantitative comparison of images generated by different methods based on latent guidance

方法	CelebA-HQ		FFHQ	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
MUNIT[8]	24.96	0.252	28.03	0.231
DRIT++[9]	52.13	0.178	44.12	0.061
StarGAN-v2[10]	13.71	0.452	18.76	0.455
PHStarGAN-v2 $n=3$[20]	16.63	0.332	43.03	0.342
本文方法	**13.31**	**0.476**	**17.01**	**0.479**
数据集图像	14.80	—	10.50	—

4.2.2　基于参考引导生成比较

参考引导生成即生成器根据采样网络所获得的风格特征 $v=M_{y_t}(x_r)$ 进行图像合成。

图 7 与图 8 分别给出了在 CelebA-HQ 数据集以及 FFHQ 数据集各方法于不同的翻译方向中所生成的图像（可放大查看细节）。(a)－(e)分别为 MUNIT[8]，DRIT++[9]，PHStarGAN-v2[20]，StarGAN-v2[10] 以及本文方法。(a)以及(b)方法中生成的图像并没有很好地结合目标域图 x_t 的语义信息，生成结果与源域图差别不大，且(a)方法中出现大面积伪影，(b)方法中丢失了大量的图像细节。(c)(d)(e)方法均较好地参考了目标域图 x_t 的信息，但(c)方法受生成分辨率影响，图像整体质量较差，细节损失较多，且容易受到遮挡影响。(d)方法对遮挡部位同样不具有鲁棒性，如图 7(a)第二行、图 7(b)第一行、图 8(a)第三行以及图 8(b)第二行，均没有很好地对遮挡部位进行处理，以至于出现大面积伪影和不规则形状，同时在(d)方法对应图像的彩色框所选定部分没有正确地保留原始图像的轮廓细节。相较以往先进方法，本文方法在面部图像存在不同角度或是大面积遮挡的翻译任务中均展现出了较优性能，有效地克服了外部干扰，生成的图像在光照、背景以及头发的自然程度都要更胜一筹，同时也还原出了源域图像的面部细节。

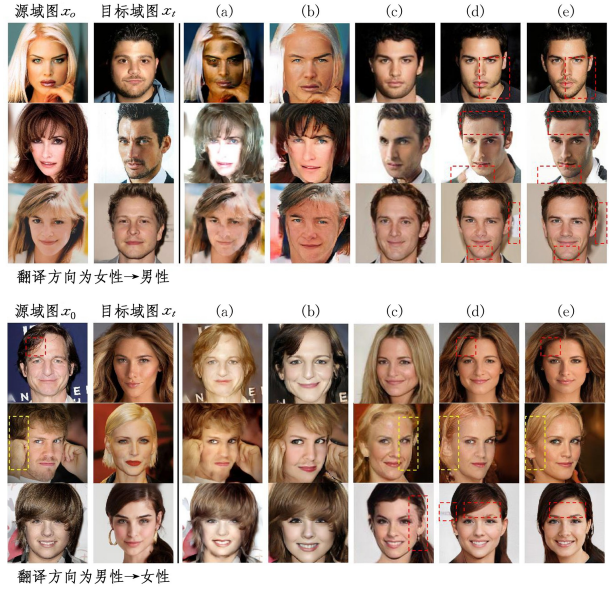

图 7　CelebA-HQ 数据集中基于参考引导生成的图像

Fig. 7　Image generated based on reference guidance in CelebA-HQ dataset

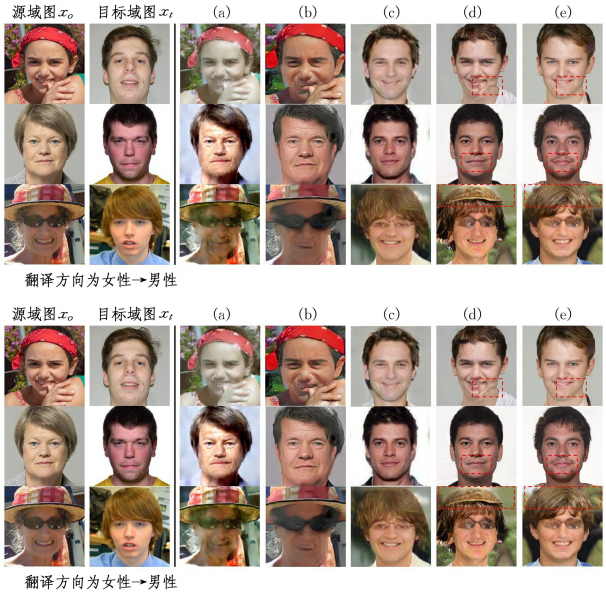

图 8 FFHQ 数据集中基于参考引导生成的图像

Fig. 8 Image generated based on reference guidance in FFHQ dataset

表 2 列出了基于参考的图像生成的定量比较结果。对于两个数据集，本文方法均获得了最低的 FID 分数；DRIT 获得了较高的 FID 以及最低的 LPIPS 分数，并且在图像结果中也说明其生成质量以及多样性较差。相比传统先进方法，本文方法在 LPIPS 没有落后的同时进一步提升 FID 分数，证明了本文所提方法可以在不影响图像生成多样性的前提下提升图像生成质量。

表 2 对不同方法基于参考引导生成的图像定量比较

Table 2 Quantitative comparison of images generated by different methods based on reference guidance

方法	CelebA-HQ		FFHQ	
	FID↓	LPIPS↑	FID↓	LPIPS↑
MUNIT[8]	64.54	0.123	51.29	0.126
DRIT++[9]	53.30	0.311	40.08	0.032
StarGAN-v2[10]	23.84	0.388	29.94	0.413
PHStarGAN-v2 $n=3$[20]	28.11	0.290	56.82	0.332
本文方法	**23.02**	**0.401**	**25.09**	**0.417**
数据集图像	14.80	—	10.50	—

4.2.3 消融实验

在消融实验中，本文在 CelebA-HQ 数据集中分别验证了 HFI 结构、图像动态一致性损失以及基于风格特征的多样性正则项的作用。表 3 列出了在不同配置下的 FID 和 LPIPS 分数。w/HFI 指向基准模型添加 HFI 模块，w/HFI+dr 指在前者的基础上只添加图像动态一致性损失，w/HFI+vms 即添加基于风格特征的多样性正则项。

表 3 模型消融实验

Table 3 Model ablation experiment

方法	潜在引导生成		参考引导生成	
	FID↓	LPIPS↑	FID↓	LPIPS↑
baseline	19.98	0.125	19.63	0.124
w/HFI	15.17	**0.519**	26.07	**0.455**
w/HFI+dr	13.20	0.469	**22.19**	0.394
w/HFI+vms	13.62	0.460	24.53	0.387
all	**13.31**	0.476	23.02	0.401

基准模型（baseline）是在 HFIGAN 模型的基础上删去 HFI 模块，使得编码特征可以直接作用于解码过程。相比基准模型，添加 HFI 模块之后性能大幅度提升。由于生成图像中源域和目标域信息占比较为合理，多样性表现较好，但在参考引导生成中 FID 分数退化严重。在上述基础中添加图像动态一致性损失之后，虽然 LPIPS 分数稍有下降，但无论是潜在引导生成或参考引导生成的 FID 分数都有较大提升。在 w/HFI+vms 的配置中，相比只添加 HFI 模块，FID 同样得到了提升，且 LPIPS 分数与 w/HFI+dr 相差不大，证明基于风格特征的多样性正则项可以在不影响图像多样性的同时带来图像质量的提升。

本文也对不同的多样性损失进行了消融实验，分别使用了文献[10]以及文献[12]方法所对应的多样性损失，结果如表 4 所列。需要特别指出的是，表 4 第一行中 w/[12]由于损失不稳定，在模型整体迭代至 20 000 次之后才开始计算多样性损失。文献[12]方法的多样性如式(10)所示，由于 z 的维度较低，导致不同 z 之间的距离相比图像之间的距离有量级上的差距，所以损失波动较大，在潜在引导生成以及参考引导生成中 FID 和 LPIPS 分数均最低。文献[10]方法对应式(11)，由于只是单纯地最大化不同风格特征所对应生成的图像间的距离，因此 LPIPS 表现较好，但图像质量较差，客观指标中则反映为 FID 分数较高。相比其他方法，多样性正则项添加了风格特征向量在不同空间中的距离度量，在 FID 与 LPIPS 得分中取得了较为均衡的成绩。

表 4 多样性损失比较

Table 4 Diversity loss comparison

方法	潜在引导生成		参考引导生成	
	FID↓	LPIPS↑	FID↓	LPIPS↑
w/[12]	37.89	0.391	26.78	0.353
w/L1 损失[10]	15.17	**0.519**	26.07	**0.455**
w/vms	**13.62**	0.460	**24.53**	0.387

4.2.4 损失函数通用性实验

为了验证图像动态一致性损失以及基于风格特征的多样性正则项在 I2I 任务中具有通用性，本文选取 I2I 先进方法并结合上述损失在 CelebA-HQ 数据集中进行验证。由于 MUNIT[8]网络直接将潜在码 z 作为风格特征，故针对文献[8]只验证图像动态一致性损失。

表 5 列出了损失函数通用性实验的客观指标结果，加粗字体为最优结果，下划线字体代表次优结果。

表 5 损失函数通用性实验

Table 5 Experiment on universality of loss functions

方法	额外添加的损失	潜在引导生成		参考引导生成	
		FID↓	LPIPS↑	FID↓	LPIPS↑
MUNIT[8]	原方法	24.96	0.252	64.54	0.123
	w/dr	**22.38**	**0.255**	**59.27**	**0.136**
	w/vms	—	—	—	—
StarGAN-v2[10]	原方法	13.71	0.452	**23.84**	0.388
	w/dr	13.36	0.449	**20.95**	0.386
	w/vms	**13.34**	**0.469**	21.73	**0.394**
文献[21]	原方法	17.38	0.391	26.48	0.308
	w/dr	**15.31**	0.387	**23.29**	0.301
	w/vms	16.72	**0.431**	24.07	**0.351**

首先在文献[8]方法中，添加动态一致性损失之后 FID 分数相比原方法有较大提升。其次在文献[10]中，相比原方法，分别添加两个损失之后，LPIPS 略有提升，但 FID 指标

提升较大。最后在文献[21]方法中，添加的各个损失均有提升。说明动态一致性损失以及基于风格特征的多样性正则项可以在确保原有图像多样性的前提下稳定地提升图像生成质量，使得生成图像的分布更加靠近数据集图像。综上，表5结果说明本文所提两个损失在I2I任务中具有一定的通用性。

4.2.5 身份一致性实验

面部属性编辑任务需保持原图像与生成图像面部身份信息一致。本实验将原图像 x 以及 x 对应的风格特征输入至不同模型的生成器中，进而获得伪原图像 x'，其中 $x'=G(x, M_y(x))$。理想情况下 x 与 x' 是完全相同的，但受限于网络模型性能，x 与 x' 之间存在或多或少的偏差，所以本文添加已经预训练过的VGGFace[33]模型来衡量偏差大小。VGGFace需裁切出图像的面部区域之后再进行计算，可以较好地提取面部的身份信息。若网络在翻译过程中对于面部图像的身份信息保持越好，则 x 和 x' 在VGGFace模型中对应的输出就越相似。表6给出了不同方法对应的身份一致性距离，其中距离是VGGFace对应 x 和 x' 两者输出的平均绝对误差，距离越小说明身份信息保持得越好。其中由于DRIT++和MUNIT模型架构不同，因此在生成 x' 时将风格特征替换为零向量对应生成的风格特征。如表6所列，本文方法在身份一致性实验中表现最好，说明本方法在生成过程中很好地保留了身份信息。

表6 身份一致性距离

Table 6 Identity consistency distance

方法	MUNIT[8]	DRIT++[9]	StarGAN v2[10]	PHStarGAN v2 $n=3$[20]	本文算法
距离值	1.606	1.691	1.365	1.481	**1.263**

结束语 针对传统方法生成的图像源域和目标域信息占比不均衡而导致的图像质量较差，数据样本分布不均衡导致不同方向的翻译难度不同，以及传统多样性损失会影响I2I任务性能这3个问题，本文提出了对等注入离散小波分解的高频编码特征的HFI模块、图像动态一致性损失以及基于风格特征的多样性正则项，分别从模型结构以及损失函数来整体提升面部性别伪造任务的生成性能。通过与传统先进方法在不同的数据集中进行充分对比，并利用FID和LPIPS指标进行客观量化评估，最终实验结果证明，HFIGAN在HFI模块以及所提两个损失的加持下相比其他先进方法展现出了更好的面部性别伪造效果，并在损失函数的通用性实验中证明了所提损失函数在I2I任务中具有一定的通用性。本文模型是对性别这种面部全局属性进行修改，而如何提高局部属性的可控性（如肤色、年龄等）是今后工作的重点。

参 考 文 献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. Cambridge, MA, US: MIT Press, 2014: 2672-2680.

[2] MIRZA M, OSINDERO S. Conditional Generative Adversarial Nets [EB/OL]. (2014-11-06)[2022-08-16]. https://arxiv.org/abs/1411.1784.

[3] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 1125-1134.

[4] 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. Long Beach, CA, USA: IEEE, 2019: 2337-2346.

[5] LEDIG C, THEIS L, HUSZAR F, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4681-4690.

[6] LI X, ZHANG S, HU J, et al. Image-to-image Translation via Hierarchical Style Disentanglement [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2021: 8639-8648.

[7] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2223-2232.

[8] HUANG X, LIU M Y, BELONGIE S, et al. Multimodal unsupervised image-to-image translation[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 172-189.

[9] LEE H Y, TSENG H Y, MAO Q, et al. DRIT++: Diverse Image-to-Image Translation via Disentangled Representations [EB/OL]. (2019-05-02) [2022-08-16]. https://arxiv.org/abs/1905.01270.

[10] CHOI Y, UH Y, YOO J, et al. Stargan v2: Diverse image synthesis for multiple domains[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 8188-8197.

[11] HUANG X, BELONGIE S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[C]// IEEE. 2017.

[12] MAO Q, LEE H Y, TSENG H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 1429-1437.

[13] LIN T Y, GOYAL P, GIRSHICKR, et al. Focal Loss for Dense Object Detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[14] KARRAS T, AILA T, LAINE S, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation [EB/OL]. (2018-02-26) [2022-08-16]. https://arxiv.org/abs/1710.10196.

[15] KARRAS T, LAINE S, AILA T. A Style-Based Generator Architecture for Generative Adversarial Networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 4401-4410.

[16] HE Z, ZUO W, KAN M, et al. AttGAN: Facial Attribute Editing by Only Changing What You Want[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5464-5478.

[17] LIU M, DING Y, XIA M, et al. Stgan: A unified selective transfer network for arbitrary image attribute editing[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 3673-3682.

[18] CHOI Y, CHOI M, KIM M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Trans-

lation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, UT, USA: IEEE, 2018: 8789-8797.

[19] YANG G, FEI N, DING M, et al. L2M-GAN: Learning to Manipulate Latent Space Semantics for Facial Attribute Editing [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Virtual: IEEE, 2021: 2950-2959.

[20] GRASSUCCI E, SIGILLO L, UNCINI A, et al. Hypercomplex Image-to-Image Translation [EB/OL]. (2022-05-04) [2022-08-16]. https://arxiv.org/abs/2205.02087.

[21] LIU Y, SANGINETO E, NADAI M D, et al. Smoothing the Disentangled Latent Style Space for Unsupervised Image-to-Image Translation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2021.

[22] ZHOU T, KRÄHENBÜHL P, AUBRY M, et al. Learning Dense Correspondence via 3D-guided Cycle Consistency[C]// IEEE. 2016.

[23] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised Learning of Depth and Ego-Motion from Video[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[24] HOFFMAN J, TZENG E, PARK T, et al. Cycada: Cycle-consistent adversarial domain adaptation[C]// International Conference on Machine Learning. Pmlr, 2018: 1989-1998.

[25] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2223-2232.

[26] LI X, WANG W, WU L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012.

[27] SPIEGL B. Contrastive Unpaired Translation using Focal Loss for Patch Classification[J]. arXiv: 2109.12431, 2021.

[28] YUN P, TAI L, WANG Y, et al. Focal loss in 3d object detection[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 1263-1270.

[29] RIDNIK T, BEN-BARUCH E, ZAMIR N, et al. Asymmetric loss for multi-label classification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 82-91.

[30] SMITH L N. Cyclical Focal Loss[EB/OL]. (2014-02-16)[2022-08-16]. https://arxiv.org/abs/2202.08978.

[31] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium[C]// Neural Information Processing Systems(NIPS). Long Beach, CA, USA: MIT Press, 2017: 6626-6637.

[32] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 586-595.

[33] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep Face Recognition[C]// British Machine Vision Conference. Swansea, UK, 2015.

CHEN Wanze, born in 1995, postgraduate. His main research interests include facial image synthesis, facial attribute operation, etc.

CHEN Jiazhen, born in 1971, associate professor. Her main research interest is information security.