Hans Journal of Data Mining 数据挖掘, 2023, 13(1), 23-35
Published Online January 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.131003

Hans汉斯

基于改进的MUNIT人脸图像性别转换模型

卢　维[1,2]，何　强[1,2*]

[1]北京建筑大学理学院，北京
[2]北京建筑大学，大数据建模理论与技术研究所，北京

收稿日期：2022年11月27日；录用日期：2022年12月27日；发布日期：2023年1月4日

摘　要

基于生成对抗网络的图像风格迁移算法已成为人脸图像性别转换的主流模型，但现有方法仍存在转化后的人脸图像模糊，背景图像扭曲，面部身份保留效果不好等缺点。针对上述问题，基于多模态无监督图像翻译网络(MUNIT)，本文提出了基于改进的人脸图像性别转换模型。首先对MUNIT模型生成器部分进行优化，在编码器部分加入动态实例归一化操作(DIN)，使编码器对人脸内容特征和风格特征的剥离更加精确；并在内容编码部分的残差块网络后加入混合注意力模块(CBAM)，使模型提取更丰富的人脸关键特征；此外，对CeleBA数据集的人脸图像根据属性进行筛选以及裁剪，减少了图像背景对于生成图像的影响，使模型更加专注于人脸特征的学习。根据实验对照情况，本文方法能够生成更加精细的人脸性别转换图像。

关键词

深度学习，生成对抗网络，风格迁移，无监督风格迁移，人脸性别转换

Gender Transformation Model of Face Image Based on Improved MUNIT

Wei Lu[1,2], Qiang He[1,2*]

[1]School of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Institute of Big Data Modelling and Technology, Beijing University of Civil Engineering and Architecture, Beijing

Received: Nov. 27th, 2022; accepted: Dec. 27th, 2022; published: Jan. 4th, 2023

Abstract

The image style transfer methods based on generative adversarial network have become the main-

*通讯作者。

文章引用: 卢维, 何强. 基于改进的 MUNIT 人脸图像性别转换模型[J]. 数据挖掘, 2023, 13(1): 23-35.
DOI: 10.12677/hjdm.2023.131003

stream model of face image gender transformation. However, for existing methods, the transformed face image is blurred, the background image is distorted, and the facial identity preservation effect is not good. Aiming at the above problems, this paper proposes an improved face image gender transformation model based on multi-modal unsupervised Image Translation Network (MUNIT). Firstly, the generator part of MUNIT model is optimized, and the dynamic instance normalization operation (DIN) is added to the encoder part to make the encoder more accurate in the stripping of face content features and style features. The mixed attention module (CBAM) is added after the residual block network in the content encoding part, so that the model can extract more abundant face key features. In addition, the face image of CeleBA dataset is screened and trimmed according to its attributes, which reduces the influence of image background on image generation and makes the model more focused on the learning of face features. According to the experimental situation, the proposed method can generate more refined facial gender conversion images.

Keywords

Deep Learning, Generating Adversarial Networks, Style Transfer, Unsupervised Style Migration, Facial Sex Conversion

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

人脸作为人最典型的外部身份特征之一，具有非常重要的研究价值。近年来，随着深度学习的不断发展，对于人脸的感知研究越来越多的受到关注。其中，人脸识别[1] [2]已经应用到社会生活中的方方面面，比如上班打卡，人脸支付等；人脸关键点检测[3] [4]，人脸 3D 重建[5]以及人脸美化[6] [7]都已经有了非常深入的研究。其中关于人脸的合成转换大多依靠生成对抗网络(GAN) [8]来实现。人脸图像性别转换可以看作人脸风格迁移的一种，需要在不改变原主身份的情况下生成异性的人脸图像，一般依靠生成对抗网络来实现。Karras 等[9]提出了一种无监督人脸属性迁移的网络 StyleGAN，它设计了新型生成器结构，将图像的高层语义信息进行解耦分离，可以在一定程度上对人脸进行合成。Zhu 等[10]提出了一种循环一致性图像转换网络 CycleGAN，该网络可以在无配对数据集情况下进行图像风格转换，但对于人脸性别转换任务来说，生成结果不够精细。Kim 等[11]提出 UGATIT 模型，将辅助分类器得到的特征图输入到注意力模块，以便于更好区分源域和目标域，使模型迁移效果更加优秀。但容易改变图像无关背景。Huang 等[12]在 2018 年提出多模态无监督图像转换网络(MUNIT)，它将图像的隐藏编码进一步细化为图像内容编码和图像风格编码，通过改变编码的方式来完成图像的风格交换，但对于特定人脸图像性别转换问题，其图像生成结果仍存在人脸图像模糊，背景图像扭曲，面部身份保留效果不好等缺点。针对上述问题，基于 MUNIT 模型，本文提出一种改进的人脸图像性别转换模型，并通过实验验证了其有效性。

本文的主要贡献如下：

1) 在网络结构上改进生成器，在编码器部分加入动态实例归一化操作(DIN) [13]，使编码器对人脸内容特征和风格特征的剥离更加精确。

2) 在内容编码部分的残差块网络后加入混合注意力模块(CBAM) [14]，使得模型能够更加有效地学习人脸图像中关于性别特征的部分，减少图像无用信息对于生成结果的影响。

3) 在训练策略上，对 CeleBA 数据集的人脸图像根据属性进行简单的筛选以及合适的裁剪，减少了图像背景对于生成图像的影响，使模型更加专注于人脸特征的学习。

本文将改进的人脸图像性别转换模型在数据集 CeleBA 上进行实验，通过主观视觉评价，以及基于内容准确率和结构相似度的客观评价指标，表明了所提方法的先进性。

2. 相关工作

人脸图像性别转换属于人脸风格迁移的特例，需要在保留原图像人脸的面部身份(内容特征)的条件下来转换人脸性别(风格特征)，且往往需要在无配对的性转数据集情况下完成迁移。相比于一般的风格迁移问题更具挑战性。现有方法仍存在迁移后图像模糊，背景图像扭曲，面部身份保留效果不理想等问题。多模态无监督图像翻译网络 MUNIT 模型被广泛研究应用于图像风格迁移，本文在 MUNIT 模型的基础上对人脸图像性别转换问题进行研究和改进，提出了一种基于改进 MUNIT 的人脸图像性别转换模型，并通过实验验证了其有效性。

2.1. MUNIT 模型

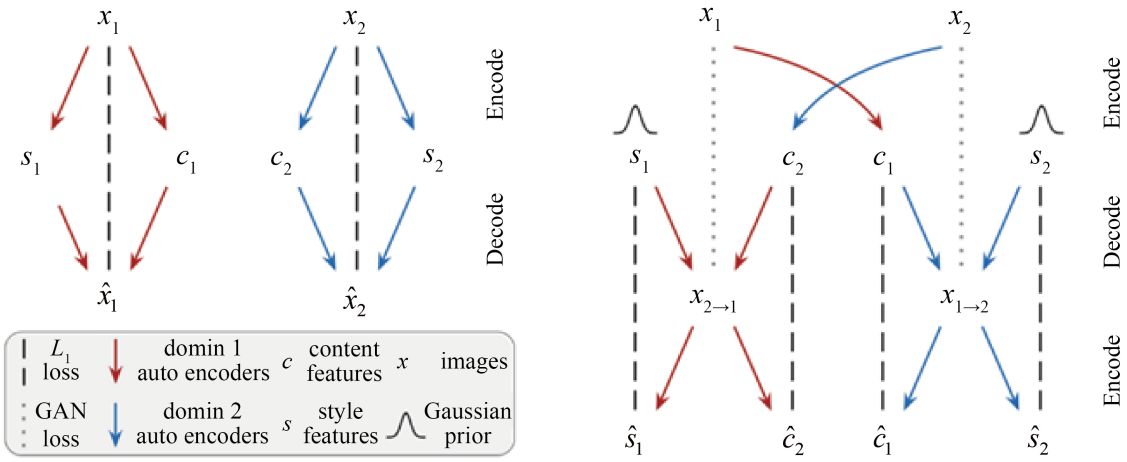
MUNIT 方法前身是 Liu 等人提出的非监督图像翻译模型(UNIT) [15]。在 UNIT 方法中，Liu 等人假设两个不同的图像风格域之间转化的实质是计算域的联合分布。在已知两幅图像边缘分布的情况下，网络通过参数的学习最终推断出风格迁移后联合分布的结果。作者利用编码器将原始图像域进行编码，计算时假设不同风格的图像共享隐藏编码，最终生成器(解码器)可以融合隐藏编码将目标图像变为不同风格的图像。相反地，编码器的作用就是将图像还原为隐藏编码。MUNIT 则在 UNIT 基础上提出了进一步的假设和实验。认为图像的隐藏编码可以进一步细化为相互独立的图像内容编码和图像风格编码。不同风格域的图像共享内容编码而独享自己的风格编码。图像的内容编码信息与风格信息不同，一般包含的都是高维信息。高维向量组成的特征矩阵作为风格编码可以描述图像中更多的结构信息和位置信息等。

MUNIT 作者认为风格编码 s 与内容编码 c 为相互独立的图像信息空间。在不同的域之间，内容编码空间是共享的。内容空间中包含一些图像内物体像素级属性，例如边缘信息、相对位置、朝向等信息，而风格编码则蕴含一些风格特征信息例如颜色、纹理等等。假设两个不同的域 $X1$ 和 $X2$ 的风格编码空间分别为 $s1$ 和 $s2$，图像共享的内容编码空间为 $c1$，图像的风格迁移过程如下式所示：

$$P(c1,s2)=G2\left(P(c1),P(s2)\right) \tag{1}$$

其中，$G2$ 为图像风格空间 $s2$ 的风格迁移生成器。编码器通过参数学习分别将风格编码空间$s2$ 和内容编码空间 $c1$ 从不同的图像域 $X2$ 和 $X1$ 中提取出来。作者假设前两者的分布相互独立，解码器就可以通过参数学习和损失函数的指导，学习到风格分布 $P(s2)$和内容分布 $P(c1)$的联合分布 $P(c1,s2)$。而学习到的联合分布就是将风格 $s2$ 融合到内容 $c1$ 的风格迁移图像结果。MUNIT 方法可以通过改变不同的风格编码进行多次风格迁移。

MUNIT 网络的主要结构如图 1 所示。网络结构与 CycleGAN [10]的循环对称结构类似。图像经过解码器E后生成对应的内容编码 c 和风格编码 s。将不同风格图像的风格编码 s 交换之后，利用生成器 G 还原成图像，完成一次单向的风格迁移过程。通过两个相同且对称的风格迁移过程，风格图像 $x1$ 和 $x2$ 分别变为 $x1\to2$ 和 $x2\to1$。如图 1 中(a)过程所示，图像需要经过重建损失，即将图像 x 通过编码器生成其对应风格内容编码后再次重新组合，确保生成器的图像生成能力准确，避免出现模式崩塌的情况。



(a) Within-domain reconstruction (b) Cross-domain translation

Figure 1. MUNIT style encoding and decoding process (from literature [12])
图 1. MUNIT 风格编解码过程(摘自文献[12])

2.2. 人脸属性迁移

人脸属性是指基于人脸各种生物特征所表现的隐藏信息，比较明显的属性例如性别，年龄以及种族。人脸属性迁移可看作图像的风格迁移。Isola 等[16]在条件生成对抗网络 CGAN [17]的基础上提出通用的非配对图像翻译网络，该网络将条件图像直接作为模型初始输入，生成的图像直接受到条件图像的影响，从而实现从一类图像转换到另一类图像的通用解决方案。Ma 等[18]提出一种双一致性损耗来训练带有对抗鉴别器的编码器解码器网络，鼓励输出在语义和风格上与主题相关的内容和风格图像对一致，可以使图像生成结果更加适应内容图像的主题。Choi 等[19]提出 StarGAN 模型，能够仅使用一个单一模型就实现多领域的图像转换，它允许在单个网络上同时训练带有不同领域的多个数据集从而实现多种不同风格的图像转换任务。Sanakoyeu 等[20]提出了一个风格感知损失函数，与一个 encoder-decoder 网络联合训练训练出特定艺术家风格。Wu 等[21]基于 DualGAN，在目标函数上附加两个新的损失函数，通过优化参数实验找到最优参数，提升了图像翻译效果。Peng 等[22]基于 CycleGAN 网络模型，通过加入局部二值模式 LBP 算法，增强了模型提取图像纹理特征的能力。Bao 等[23]提出了 CVAE-GAN 模型，通过将变分自动编码器(VAE)与生成对抗网络(GAN)相结合，改变生成模型中的输入标签信息来生成特定类别的图像。

关于图像风格迁移的模型方法多种多样，但对于具体的人脸图像性别迁移问题研究还有待研究。人脸图像性别转换问题有四大难点，一是难以获得配对的数据集，男女性别转换结果没有标准答案；二是难以界定或是捕捉人脸图像性别特征，需要找到决定人脸图像性别的关键特征；三是难以在性别(风格迁移)转换后保留原主的面部身份(内容特征)；四是难以消除图像无关背景域对于模型结果的影响，现有的图像风格迁移模型大多对输入图像的总体进行学习，导致人脸图像性别转换结果易受到无关背景域的影响。针对上述问题，石达等[24]提出来基于改进 CycleGAN 的人脸性别伪造图像生成模型，通过在循环生成对抗网络 CycleGAN 的生成器后加入混合注意力和自适应残差块，结合相对损失函数得到了不错的人脸图像性别转换效果，但仍无法解决无关背景域的影响。Liu 等[25]在多模态无监督图像翻译网络(MUNIT)的基础上引入新的人脸性别概率性掩膜，促进实现性别转移和身份保留的目标，同时通过人脸稀疏特征学习到关于人脸性别的决定性因素，最终获得了较好的性别转换效果，但对于人脸面部颜色，细节的部

分仍有改进的空间。

3. 基于改进 MUNIT 的人脸图像性别转换模型

本文方法基于 MUNIT 模型，将风格迁移之前的图像分离为内容特征和风格特征，随后对提取的风格特征进行迁移。在人脸图像性别转换这一具体问题上，内容特征代表了人的面部身份，风格特征代表了人的性别属性，包括眉、眼及唇形等各种面部细节。通过抽离性别特征，在固定内容特征的情况下，实现人脸图像性别的转换。本节将介绍改进的 MUNIT 网络模型以实现更好效果的性别转换图像生成结果。

3.1. 模型整体结构

完整的改进 MUNIT 网络模型如图 2 所示，网络通过内容编码器和风格编码器分别提取并交换图像的内容特征和风格特征，最终完成人脸图像性别转换过程。

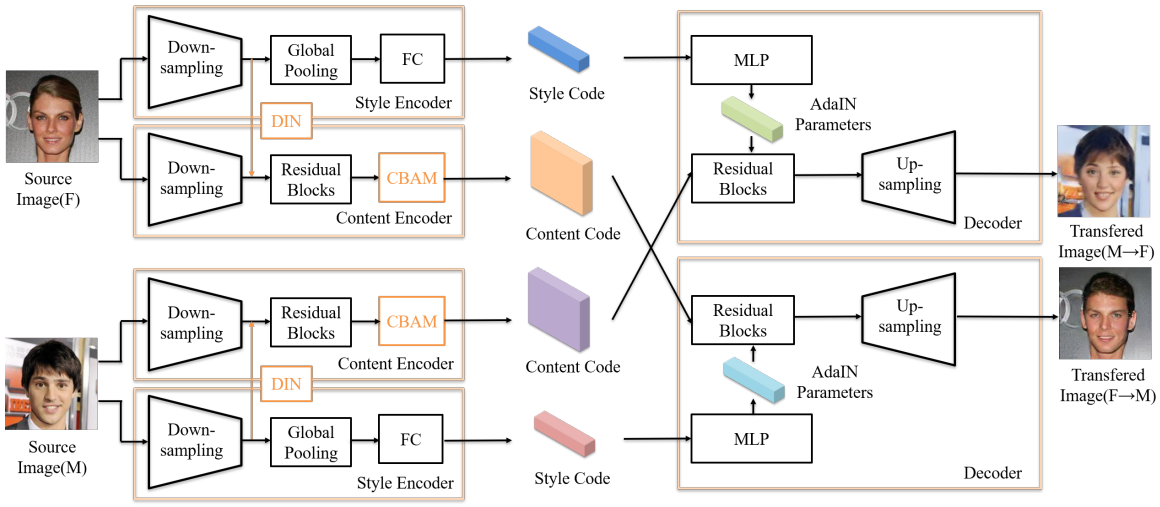


Figure 2. Improved MUNIT network structure diagram
图 2. 改进的 MUNIT 网络结构图

其中，风格编码器由下采样部分(Down-sampling)、全局池化层(Global Pooling)和全连接层(Fully Connected, FC)组成；内容编码器由下采样部分，动态实例归一化(DIN)残差模块(Residual Blocks)和混合注意力模块组成。而解码器则是由多层感知机(Multilayer Perceptron)、残差模块和上采样部分(Up-Sampling)构成。

3.2. 生成器结构

为了提高生成图像的质量，使性别转换后的图像更加真实自然，本文引入了动态实例归一化操作以及混合注意力机制。

3.2.1. 动态实例归一化(DIN)

动态实例归一化(DIN) [13]是近年来一个新的规范化模块，它可以实现更加灵活和有效的任意风格传输。DIN 的网络结构如图 3 所示，包括一个实例归一化以及一个动态卷积，它将图像的风格编码视为卷积参数进行学习，从而实现对内容图像进行风格迁移。本文将 DIN 引入到 MUNIT 模型对人脸面部风格与内容的解码当中，可以使人脸特征的剥离更加精确。

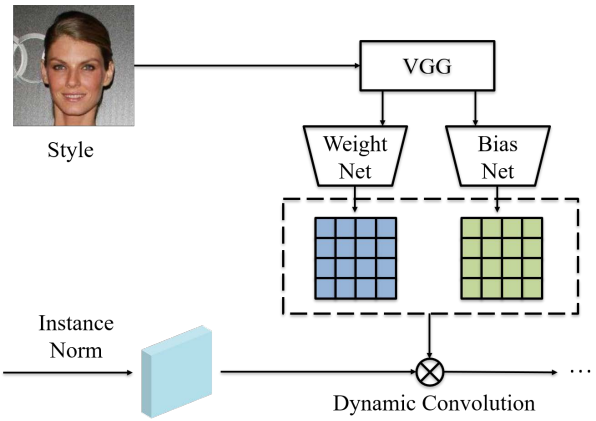


Figure 3. DIN network structure
图 3. DIN 网络结构

动态实例归一化包括实例归一化和动态卷积操作，其中 weight net 和 bias net 由简单的卷积层和自适应池化层构成。公式如下：

$$DIN\left(F_c^L\right)=IN\left(F_c^L\right)*W^L+b^L \tag{2}$$

其中 F_c 是内容输入的特征图，L 是特定层，W 是学习到的权重矩阵，b 是学习到的偏置向量。$IN(*)$ 是实例归一化操作，公式如下：

$$IN\left(x\right)=\frac{x-mean\left(x\right)}{std\left(x\right)} \tag{3}$$

根据该公式，可以看出利用 DIN 操作学习到的 W 其实就是风格的标准差，b 是风格的均值。

3.2.2. 注意力机制

注意力机制从本质上讲和人类的选择性视觉注意力机制类似，核心目标是从众多信息中选择出对当前任务目标更关键的信息，通过添加权重的方式，将重要程度高的特征进行强化学习，对重要程度较低的特征进行消减，从而提高深度学习网络模型的性能。

注意力模块大致可以分为三种：空间注意力(spatial domain)、通道注意力(channel domain)以及混合注意力(mixed domain)。空间注意力机制的经典是空间变幻网络(STN) [26]。STN 通过将原始图片中的空间信息变换到另一个空间中并保留关键信息，减少图像中无关信息对模型训练的干扰，从而提升了模型的性能。通道注意力机制的经典是 squeeze-and-excitation 网络(SENet) [27]。SENet 在通道维度增加注意力机制，它能够获取到特征图的每个通道的重要程度，然后用这个重要程度去给每个特征赋予一个权重值，从而让神经网络重点关注某些特征通道。将空间注意力机制和通道注意力机制进行结合，即混合域注意力机制(CBAM)。

图 4(a)为空间注意力机制结构图，通过拼接运算实现对特征图的平均池化和最大池化运算结果的融合，再经过全连接层和 Sigmoid 函数得到归一化的权重；图 4(b)为通道域注意力机制结构图，先将特征图进行全局平均池化和全局最大池化运算，两者相加后共同输入到多层感知器网络，经过全连接层和 Sigmoid 函数得到归一化的权重。

整体的混合注意力模块(CBAM) [11]流程如图 5 所示，输入特征图像经过通道注意力机制，将权重和输入特征图相乘后再送入空间注意力机制，将归一化权重和空间注意力机制的输入特征图相乘，得到最终的特征图。

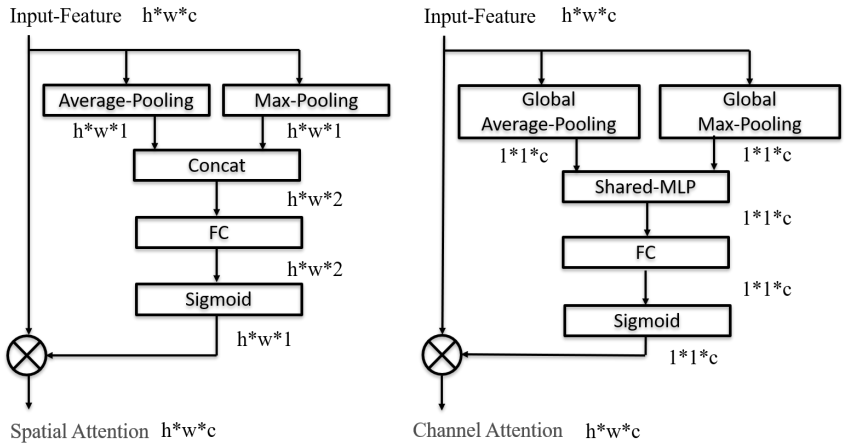


Figure 4. (a) Spatial attention module; (b) Channel attention module
图 4. (a) 空间域注意力机制；(b) 通道域注意力机制

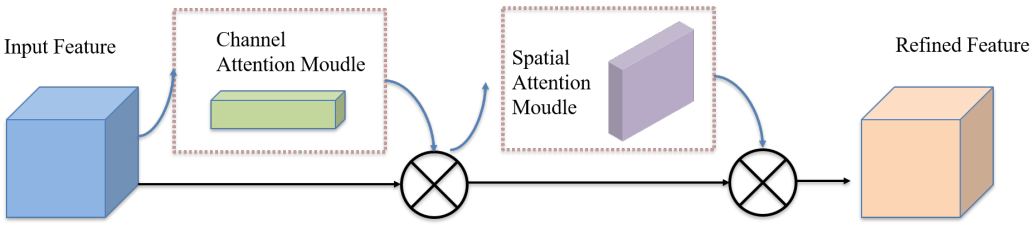


Figure 5. Convolutional block attention module
图 5. 混合注意力机制

3.3. 判别器结构

网络的生成器包括内容编码器，风格编码器以及解码器。鉴别器用来与生成器对抗以确保生成器的风格迁移效果提升。在网络训练过程中鉴别器判断生成图像和原图的风格差异，再通过损失函数的反馈指导生成器的训练。本文使用 PatchGAN 结构[28]，去掉了传统鉴别器中的全连接层，改用全卷积生成特征块，最后以输出特征块的均值来当作风格判定的概率。这种鉴别器结构的优点是突出了图像的全局特征，能更加准确判别出生成图像域和原始图像域的差别，从而有效地指导生成器的训练。

3.4. 损失函数

本文的损失函数与原始 MUNIT 模型相同，我们集合编码器，解码器，鉴别器的 loss，当作最后优化的目标，其为对抗性损失和双向重建损失项的加权和，如下：

$$\begin{aligned}&\min_{E_1,E_2,\mathrm{G}_1,\mathrm{G}_2}\min_{D_1,\mathrm{D}_2}L\left(E_1,E_2,\mathrm{G}_1,\mathrm{G}_2,D_1,\mathrm{D}_2\right)\\&=L_{GAN}^{x1}+L_{GAN}^{x2}+\lambda_x\left(L_{recon}^{x1}+L_{recon}^{x2}\right)+\lambda_c\left(L_{recon}^{c1}+L_{recon}^{c2}\right)+\lambda_s\left(L_{recon}^{s1}+L_{recon}^{s2}\right)\end{aligned}\tag{4}$$

这里的 $\lambda_x,\lambda_c,\lambda_s$ 是控制每项 loss 的权重参数。(4)式中前两项为对抗损失，使用 GANs 来匹配翻译后图像的分布和目标数据的分布。

$$L_{GAN}^{x2}=E_{c1\sim p(c1),s2\sim p(s2)}\left[\log\left(1-D_2\left(G_2\left(c1,s2\right)\right)\right)\right]+E_{x2\sim p(x2)}\left[\log\mathrm{D}_2\left(x2\right)\right]\tag{5}$$

这里 D_2 是鉴别生成的图像是否符合域 X_2 的分布，鉴别器 D_1 以及 L_{GAN}^{x1} 有类似定义。(4)式中第三项为

图像的重建损失，给定一个从数据分布中采样的图像，我们能够在编码和解码后重建它。

$$L_{recon}^{x1}=E_{x1\sim p(x1)}\left[\left\|G_1\left(E_1^c\left(x1\right),E_1^s\left(x1\right)\right)-x1\right\|_1\right] \tag{6}$$

(4)式中第四、五项为图像的内容风格损失，给出一个来自于 latent distribution 的 latent code(style 或者 content)，我们能够在编码和解码后重构它。

$$L_{recon}^{c1}=E_{c1\sim p(c1),s2\sim q(s2)}\left[\left\|E_2^c\left(G_2\left(c1,s2\right)\right)-c1\right\|_1\right] \tag{7}$$

$$L_{recon}^{s2}=E_{c1\sim p(c1),s2\sim q(s2)}\left[\left\|E_2^s\left(G_2\left(c1,s2\right)\right)-s2\right\|_1\right] \tag{8}$$

这里的 $q\left(s2\right)$ 表示先验分布 $N\left(0,1\right)$，$p\left(c1\right)$ 是由 $c1=E_1^c\left(x1\right)$ 和 $x1\sim p\left(x1\right)$ 给出。

4. 实验与分析

4.1. 数据集

本文在综合考虑后，选用公开数据集 CelebFaces Attributes Dataset (CelebA)。CeleBA 数据集是一个大规模的人脸属性数据集，包括 10177 个身份，202599 张人脸图像，且每张照片都有特征标注信息，包含性别以及各种人脸特征等 40 多项信息。将 CeleBA 数据集的训练集输入模型进行训练，验证集和测试集输入模型进行测试。为减少无关背景因素对于图像生成结果的影响，我们对数据集的标注信息进行预处理，选取年轻人并对图像做合适的裁剪，将图片大小调整为 256 × 256。最后男性人脸训练集和测试集数量分别是 46,372 和 4564，女性人脸训练集和测试集的数量分别是 90,016 和 10,014。

4.2. 实验细节

本文实验使用的操作系统是 ubuntu18.04，CPU 是 15 核 AMD EPYC 7543 32-Core Processor，内存 80G，GPU 是 RTX 3090，显存 24G，Python 版本为 3.6.13，Pytorch 版本是 1.10.2，cuda 版本为 11.3。将预先处理好的数据集输入到模型进行训练。在实验中，模型训练次数为 1,000,000，batchsize 设为 1，学习率设置为 0.0001，将式(4)中的 λ_c,λ_s 设置为 10。在模型的训练过程中，使用 Adam [29]优化器对梯度下降进行优化。

4.3. 评价指标

图像风格迁移结果主观性非常大，概因计算机很难对转移前后的图像风格变化给出定性的评价结果。因此，本文将结合主观视觉评价与客观指标评价对模型结果进行解析。主观视觉评价将本文模型生成结果与同等条件下其他模型生成结果随机采样，依靠不同用户的评价选出人脸性别转换效果最优的模型。客观评价指标结合内容准确率和结构相似度进行综合评判。

4.3.1. 内容准确率

内容准确率即模型生成的伪造数据通过判别器的概率，也就代表了模型生成结果的有效性。本文使用 InceptionV3 网络[30]作为分类模型。将分类模型在 CeleBA 数据集中进行预训练得到基准的内容准确率，然后将本文模型生成的伪造图像输入到预训练后的分类模型中，如果伪造的图像足够真实可以通过分类模型，将其计入正确样本，最后将正确样本与输入样本数相除即可得到最后的内容准确率，准确率越高代表模型生成效果越好。

4.3.2. 结构相似度

本文基于 FID (Fréchet Inception Distance)指标来计算男女面部特征之间的相似度。FID 代表了真实人

脸图像与模型伪造的人脸图像的特征向量之间距离的一种度量。这种视觉特征是使用 Inception v3 图像分类模型提取特征并计算得到的。FID 在最佳情况下的得分为 0.0，表示两组图像相同。分数越低代表两组图像越相似，或者说二者的统计量越相似。FID 计算式如式(9)所示：

$$FID = \left\|\mu_1 - \mu_2\right\|_2^2 + T_r\left(\sum_1 + \sum_2 - 2\left(\sum_1 \sum_2\right)^{\frac{1}{2}}\right) \tag{9}$$

其中，μ_1 和 $\sum_1$ 为输入的人脸数据集的均值和协方差矩阵，μ_2 和 $\sum_2$ 为模型生成数据集的均值和协方差矩阵，T_r 表示矩阵对角线上元素的总和。

4.4. 效果评估

4.4.1. 主观视觉评价

本文将预处理过的 CeleBA 数据集输入到改进的 MUNIT 模型，原始 MUNIT 模型以及 CycleGAN 模型中进行训练和测试，横向对比每种方法的生成结果。本文所做实验均采用经过 1000000 次迭代的生成模型，且同一种实验采用相同的测试数据，只保留生成方法和训练数据的不同。实验结果如图 6、图 7 所示。

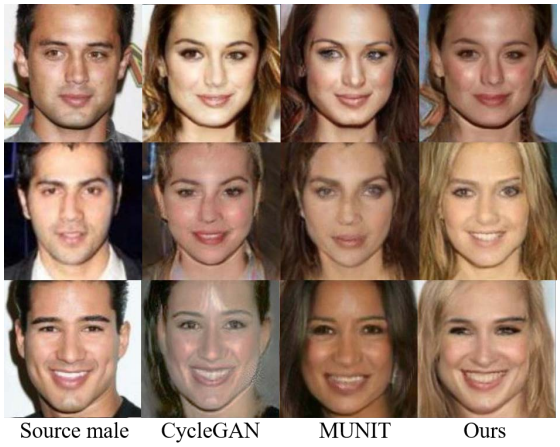


Figure 6. Male to female
图 6. 男性转为女性

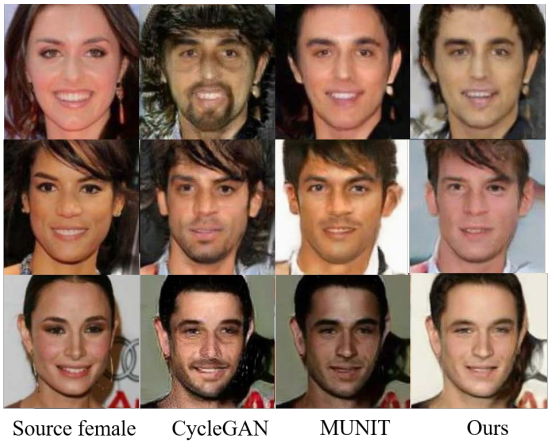


Figure 7. Feale to male
图 7. 女性转为男性

图 6，图 7 从左到右每列分别为原图像，CycleGAN 生成的性转图像，原 MUNIT 生成的性转图像以及本文模型生成的性转图像。从图 6 男性到女性的性别转换结果可以看出，本文方法在人脸部的细节变化上更加真实，背景保留效果较好。例如，在男性到女性的转换中的最后一组实验中，本文所提方法在脸颊褶皱的提取效果更好；在图 7 女性到男性的转换实验中，采用本文方法生成的人脸面部流畅度更高，图像质量更好，在无关背景域如头发的处理上更加优秀。整体上看，本文方法相较于 CycleGAN 以及 MUNIT 方法在人脸性别转换问题上表现的更好，但在性别转换过程中并未保持人脸肤色以及发型特征的一致性。

我们随机选取 20 张人脸图像，男女各十张，输入到 CycleGAN，MUNIT 和本文方法生成的结果组合成问卷，交由 287 名用户进行评选，选取性别转换后效果最好的图像(模型)。所得结果如表 1 所示，显然，经过改进的 MUNIT 模型在人脸图像性别转换上表现的最好。

Table 1. User study comparison of the gender translation performance between CycleGAN, MUNIT and ours
表 1. 人脸性别转换在 CycleGAN，MUNIT 和本文方法所得结果的用户满意度调查

模型	CycleGAN	MUNIT	Ours
满意度占比	28.57%	30.87%	40.56%

4.4.2. 客观指标评价

1) 消融实验

本文在 MUNIT 的基础上逐步增加动态实例归一化和卷积注意力机制，下面将分别计算在不同改进策略下的内容准率和 FID 得分。

如表 2 所列，添加动态实例归一化操作后生成模型对内容和风格特征的剥离更加准确，在 CeleBA 数据集上，伪造女性和伪造男性的内容准确率分别提高了 0.048 和 0.044；继续添加混合注意力机制，使模型对人脸面部性别特征学习的权重增加，内容准确率再提高了 0.117 和 0.217。模型中添加动态实例归一化后，在 CeleBA 数据集上，伪造女性和伪造男性的 FID 得分分别降低了 10.91 和 3.67；继续增加混合注意力机制后，FID 再降低了 11.84 和 4.44。从表 2 和表 3 可以看出，本文在原始 MUNIT 模型上进行的改进是行之有效的。

Table 2. Content accuracy under different conditions on the CeleBA dataset
表 2. CeleBA 数据集上不同条件下的内容准确率

	MUNIT	MUNIT + DIN	MUNIT + DIN + CBAM
男转女	0.886	0.904	0.935
女转男	0.456	0.489	0.583

Table 3. FID scores under different conditions on the CeleBA dataset
表 3. CeleBA 数据集上不同条件下的 FID 得分

	MUNIT	MUNIT + DIN	MUNIT + DIN + CBAM
男转女	86.34	75.43	63.59
女转男	45.56	41.89	37.45

2) 与其他方法对比

本文方法与其他方法的内容准确率和 FID 得分的对比结果如表 4、表 5 所列。本文方法在男女性别转换的实验中内容准确率相较于其他方法都更加优秀，说明基于本文方法生成的人脸图像更加真实。基于本文模型的男转女 FID 得分低于原始的 MUNIT 模型，高于 CycleGAN 模型，说明本文方法在身份保留方面还有进步的空间，需要继续改进；在男转女的 FID 的得分结果在几种方法中最低，说明本文方法具有更好的模型性能，使模型的人脸生成结果更真实，效果更好。

Table 4. Content accuracy of each model on CeleBA dataset
表 4. CeleBA 数据集上各模型的内容准确率

	CycleGAN	MUNIT	Ours
男转女	0.876	0.886	0.935
女转男	0.379	0.456	0.583

Table 5. FID scores of each model on the CeleBA dataset
表 5. CeleBA 数据集上各模型的 FID 得分

	CycleGAN	MUNIT	Ours
男转女	37.45	86.34	63.59
女转男	43.47	45.56	37.45

5. 结束语

本文借鉴风格迁移的思想进行人脸图像性别转换。首先在多模态无监督风格迁移模型 MUNIT 的基础上，提出了一种融合动态实例归一化和混合注意力机制的人脸图像性别转换模型；然后，对数据集进行年轻面孔的筛选以及裁剪等预处理，减少年龄变化和无关背景对于生成图像质量的影响。通过最后的实验结果可得，本文所提方法对于人脸图像性别转换任务完成的更加优秀，生成的图像结果更加真实。虽然本文所提提高了人脸性别转换生成图像的质量，但仍存在性转前后人脸面部肤色以及发型发生显著差异的问题，这跟模型对于人脸数据整体进行训练迁移有关，这也将是我们后续需要研究的问题。

未来考虑将输入的人脸图像进行面部提取工作，对人脸部分进行针对性的模型训练，进一步消除无关背景对于迁移结果的影响；同时将针对人脸性别转移前后肤色变化的问题，设计新型损失函数，提高模型生成结果的前后一致性，生成更加优秀的人脸性别转换图像。

基金项目

北京建筑大学科学研究基金(KYJJ2017017, Y19-19, Y18-11)；住房和城乡建设部科学技术计划北京建筑大学北京未来城市设计高精尖创新中心开放课题(No. UDC2019033324, UDC201703332)；北京市教育委员会科学研究计划项目资助(KM202110016001, KM202210016002)。

参考文献

[1] Schroff, F., Kalenichenko, D. and Philbin, J. (2015) FaceNet: A Unified Embedding for Face Recognition and Clustering. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 815-823. https://doi.org/10.1109/CVPR.2015.7298682

[2] An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., *et al*. (2022) Killing Two Birds with One Stone: Efficient and Robust Training of Face Recognition CNNs by Partial FC. 2022 *IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition (*CVPR*), New Orleans, 18-24 June 2022, 4032-4041. https://doi.org/10.1109/CVPR52688.2022.00401

[3] Zhang, K., Zhang, Z., Li, Z. and Qiao, Y. (2016) Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, **23**, 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[4] Sun, K., Wu, W., Liu, T., Yang, S., Wang, Q., Zhou, Q., Ye, Z. and Qian, C. (2019) FAB: A Robust Facial Landmark Detection Framework for Motion-Blurred Videos. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5461-5470. https://doi.org/10.1109/ICCV.2019.00556

[5] Feng, Y., Wu, F., Shao, X., Wang, Y. and Zhou, X. (2018) Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. 15*th European Conference*, Munich, 8-14 September 2018, 557-574. https://doi.org/10.1007/978-3-030-01264-9_33

[6] Nguyen, T., Tran, A. and Hoai, M. (2021) Lipstick Ain't Enough: Beyond Color Matching for In-the-Wild Makeup Transfer. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13300-13309. https://doi.org/10.1109/CVPR46437.2021.01310

[7] Sun, Z., Chen, Y. and Xiong, S. (2022) SSAT: A Symmetric Semantic-Aware Transformer Network for Makeup Transfer and Removal. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 2325-2334. https://doi.org/10.1609/aaai.v36i2.20131

[8] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C. and Bengio, Y. (2014) Generative Adversarial Nets. *Proceedings of the NIPS* 2014 *Workshop on High-Energy Physics and Machine Learning*, Montreal, 13 December 2014, 2672-2680.

[9] Karras, T., Laine, S. and Aila, T. (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4396-4405. https://doi.org/10.1109/CVPR.2019.00453

[10] Rai, H. and Shukla, N. (2018) Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.

[11] Kim, J., Kim, M., Kang, H. and Lee, K. (2020) U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation.

[12] Huang, X., Liu, M., Belongie, S.J. and Kautz, J. (2018) Multimodal Unsupervised Image-to-Image Translation. *ECCV* 2018: 15*th European Conference*, Munich, 8-14 September 2018, 179-196. https://doi.org/10.1007/978-3-030-01219-9_11

[13] Jing, Y., Liu, X., Ding, Y., Wang, X., Ding, E., Song, M. and Wen, S. (2020) Dynamic Instance Normalization for Arbitrary Style Transfer.

[14] Woo, S., Park, J., Lee, J. and Kweon, I. (2018) CBAM: Convolutional Block Attention Module. *ECCV* 2018: 15*th European Conference*, Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[15] Liu, M., Breuel, T.M. and Kautz, J. (2017) Unsupervised Image-to-Image Translation Networks.

[16] Isola, P., Zhu, J., Zhou, T. and Efros, A.A. (2017) Image-to-Image Translation with Conditional Adversarial Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5967-5976. https://doi.org/10.1109/CVPR.2017.632

[17] Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets.

[18] Ma, Z., Li, J., Wang, N. and Gao, X. (2020) Semantic-Related Image Style Transfer with Dual-Consistency Loss. *Neurocomputing*, **406**, 135-149. https://doi.org/10.1016/j.neucom.2020.04.027

[19] Choi, Y., Choi, M., Kim, M.S., Ha, J., Kim, S. and Choo, J. (2018) StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 8789-8797. https://doi.org/10.1109/CVPR.2018.00916

[20] Sanakoyeu, A., Kotovenko, D., Lang, S. andOmmer, B. (2018) A Style-Aware Content Loss for Real-Time HD Style Transfer. *ECCV* 2018: 15*th European Conference*, Munich, 8-14 September 2018, 715-731. https://doi.org/10.1007/978-3-030-01237-3_43

[21] Wu, H., Liu, Q., Wang, Y., Mathematics, S.O. and University, T. (2019) Face Image Translation Based on Generative Adversarial Networks. *Journal of Tianjin University* (*Science and Technology*), **52**, 306-314.

[22] Peng, Y.F., Wang, K.X., Mei, J.Y., *et al*. (2020) Image Style Migration Based on Cycle Generative Adversarial Net Works. *Computer Engineering & Science*, **42**, 699-706.

[23] Bao, J., Chen, D., Wen, F., Li, H. and Hua, G. (2017) CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2764-2773. https://doi.org/10.1109/ICCV.2017.299

[24] 石达，芦天亮，杜彦辉，张建岭，暴雨轩. 基于改进 CycleGAN 的人脸性别伪造图像生成模型[J]. 计算机科学, 2022, 49(2): 31-39.

[25] Liu, X., Wang, R., Peng, H., Yin, M., Chen, C. and Li, X. (2021) Sparse Feature Representation Learning for Deep Face Gender Transfer. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 4070-4080. https://doi.org/10.1109/ICCVW54120.2021.00454

[26] Jaderberg, M., Simonyan, K., Zisserman, A. and Kavukcuoglu, K. (2015) Spatial Transformer Networks. *NIPS* 2015, Montreal, 5 June 2015, 2017-2025.

[27] Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[28] Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J. and Catanzaro, B. (2018) High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 8798-8807. https://doi.org/10.1109/CVPR.2018.00917

[29] Kingma, D.P. and Ba, J. (2015) Adam: A Method for Stochastic Optimization.

[30] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308