计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600012

基于改进 CycleGAN 的人脸性别伪造图像生成模型

石　达　　芦天亮　　杜彦辉　　张建岭　　暴雨轩
中国人民公安大学信息网络安全学院　北京 100038
(1158083081@qq.com)

摘　要　深度伪造可以将人的声音、面部及身体动作拼接,从而合成虚假内容,用于转换性别、改变年龄等。基于生成对抗式图像翻译网络的人脸性别伪造图像存在容易改变无关图像域、人脸细节不够丰富等问题。针对这些问题,文中提出基于改进 CycleGAN 的人脸性别伪造图像生成模型。首先,优化生成器结构,利用注意力机制与自适应残差块提取更丰富的人脸特征;然后,借鉴相对损失的思想对损失函数进行改进,提高判别器的判别能力。最后,提出基于年龄约束的模型训练策略,减小了年龄变化对生成图像的影响。在 CelebA 和 IMDB-WIKI 数据集上进行实验,实验结果表明,与原始 CycleGAN 方法和 UGATIT 方法相比,所提方法能够生成更加真实的人脸性别伪造图像,伪造男性和伪造女性的平均内容准确率分别为 82.65%和78.83%,FID 平均得分分别为 32.14 和 34.50。

关键词:深度伪造;深度学习;生成对抗网络;图像翻译;图像生成;人脸性别伪造
中图法分类号　TP309;TP18

Generation Model of Gender-forged Face Image Based on Improved CycleGAN

SHI Da,LU Tian-liang,DU Yan-hui,ZHANG Jian-ling and BAO Yu-xuan
College of Information and Cyber Security,People's Public Security University of China,Beijing 100038,China

Abstract　Deepfake can be used to combine human voices,faces and body movements into fake content,switch gender and change age,etc. There are some problems of gender-forged face images based on generative adversarial image translation networks such as the irrelevant image domain changes easily and insufficient face details in generated images. To solve these problems,an generation model of gender-forged face image based on improved CycleGAN is proposed. Firstly,the generator is optimized by using the attention mechanism and adaptive residual blocks to extract richer facial features. Then,with the aim to improve the ability of the discriminator,the loss function is modified by the idea of relative loss. Finally,a model training strategy based on age constraints is proposed to reduce the impact of age changes on the generated images. Performing experiments on the CelebA and IMDB-WIKI datasets,the experimental results show that,compared with the original CycleGAN method and the UGATIT method,the proposed method can generate more real gender-forged face images. The average content accuracy of fake male images and fake female images is 82.65% and 78.83%,and the average FID score is 32.14 and 34.50,respectively.

Keywords　Deepfake,Deep learning,Generative adversarial network,Image translation,Image generation,Facial gender forgery

1　引言

深度伪造技术(Deepfake)指利用深度学习模型合成足以混淆视听、肉眼无法识别的虚假图像和视频的技术,Deepfake 一词源于 Deep Learning 和 Fake 的组合。利用深度伪造技术合成并传播虚假图像会严重危害个人隐私、侵犯肖像权和名誉权等,已成为威胁网络安全的重要因素之一。滥用人脸图像伪造技术容易扰乱社会秩序,甚至导致国家冲突。

深度伪造大多依靠共享权重的自动编码器和生成对抗网络(Generative Adversarial Networks,GAN)来实现[1]。人脸伪造图像合成根据伪造程度的不同可以分为全脸合成、局部换脸以及面部属性和表情的修改,如 deepfake[2],FaceSwap[3],Face2Face[4]以及 Faceswap-GAN[5]等。但当前大多数伪造数据集都是通过换脸得到,缺少性别伪造等人脸属性伪造数据集。性别伪造可以看作人脸属性迁移的一种,人脸属性迁移大多采用图像翻译的思想,由生成对抗网络来实现。

到稿日期:2021-06-01　返修日期:2021-09-06
基金项目:国家重点研发计划(2017YFB0802804);中国人民公安大学 2020 年基本科研业务费重大项目(2020JKF101)
This work was supported by the National Key R&D Program of China(2017YFB0802804) and 2020 Fundamental Research Funds for the Central Universities of PPSUC(2020JKF101).
通信作者:芦天亮(lutianliang@ppsuc.edu.cn)

Isola 等[6]提出了一种有监督的图像翻译网络 pix2pix，该网络需要一一配对的数据作为输入，但对于人脸属性迁移任务来说，这种配对的数据集难以得到。Zhu 等[7]提出了一种循环一致性图像翻译网络，该网络不需要一一配对的数据集作为输入，但生成的图像质量不佳。Kim 等[8]提出 UGATIT 模型，其添加的注意力模块通过基于辅助分类器获得的注意力图来区分源域和目标域，帮助模型完成迁移任务，但这样容易改变图像无关区域的纹理。针对生成图像效果不佳、无关图像域改变等问题，本文提出了一种基于非配对图像翻译网络的人脸性别伪造图像生成模型，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，通过引入注意力机制和自适应残差块，使得模型在图像转换过程中更有效地学习和提取脸部特征，提高生成图像的质量。

(2)对损失函数进行改进，改善了当生成器优化良好时判别器无法判别真伪的情况，提高了判别器的判别能力。

(3)在训练策略上，提出了一种基于年龄约束的模型训练策略，减小了年龄变化对生成图像的影响。

本文将改进的生成模型在公开数据集 CelebA 和 IMDB-WIKI 上进行实验，通过主观视觉评价，并结合内容准确率和结构相似度的客观评价指标，验证了所提方法的有效性。

2 相关工作

人脸伪造图像的合成大多依靠深度学习技术实现，深度学习技术的发展极大地提高了图像特征提取的能力，进而提升了图像合成的水平。利用生成对抗网络可以学习源人脸图像与目标人脸图像之间的转换关系，从而提升生成图像的质量。此外，生成对抗网络还应用于人脸老化等方面的研究[9]。

2.1 生成对抗网络

生成对抗网络已被广泛应用于人脸图像的合成[10-12]，其原理源于零和博弈的思想，网络由两个“博弈者”组成，即生成网络 G 和判别网络 D，两者在不断“博弈”的过程中达到平衡点。生成对抗网络的目标函数如式(1)、式(2)所示：

$$V(D,G)=E_{x\sim P_{\text{data}}(x)}\log D(x)+E_{x\sim P_z}\log(1-D(G(z))) \tag{1}$$

$$\arg\min_G\max_D V(D,G) \tag{2}$$

对于来自简单随机分布的输入数据 z，通过生成网络 G 生成 $G(z)$，对于判别网络 D，判别输入数据 $G(z)$ 是否来自真实分布 $P_{\text{data}(z)}$，用 $D(G(z))$ 表示。在“博弈”过程中，生成网络 G 的目标是能够尽量生成接近真实分布的图像，以骗过判别网络 D，而判别网络 D 的目标是能够最大概率地判别出真实分布的数据与生成网络 G 生成的假数据。

2.2 人脸属性迁移

人脸属性指人脸中包含生物特征的内在属性，包括性别、年龄、种族等。人脸属性迁移可看作一类图像域到另一类图像域的转换，根据方法原理的不同，可以将其分为两类：基于特征编码的属性迁移和基于风格转换的属性迁移。

(1)基于特征编码的属性迁移

基于特征编码的属性迁移指利用各种方法改变图像的特征编码，进而实现属性迁移的任务。Upchurch 等[13]利用深度特征插值(Deep Feature Interpolation，DFI)的方法修改图像内容，该方法在预训练的特征空间内进行插值，以实现图像转换。Perarnau 等[14]提出可逆的条件生成对抗网络(IC-GAN)，该方法通过两个独立的编码器得到输入图像特征向量和目标图像特征向量，再由 CGAN 生成特定的图像。Wang 等[15]将源图像编码成内容向量加特征向量的形式，通过只改变特征向量的方式来完成图像重构。Liu 等[16]提出 UNIT 模型，该模型由两对生成对抗网络组成，通过共享生成器中编码器的潜在空间，来实现不同属性的编码和重建任务。Huang 等[17]提出多模态无监督图像翻译网络 MUNIT，它将图像表征为内容编码和风格编码，通过重新组合不同的内容编码和风格编码来完成多图像域的迁移。Park 等[18]提出一种新颖的风格注意力网络，通过学习内容特征和风格特征之间的语义相关性，能够在尽可能丰富风格样式的同时，保留图像的内容结构。Xiao 等[19]提出基于特征解耦的 ELEGANT 方法，通过交换具有不同属性图像的隐空间特征来实现迁移任务。Cho 等[20]提出 GDWCT(Group-wise Deep Whitening-and-Coloring Transformation)，通过学习的方式构建变换矩阵，利用正则化和分组计算的方式有效减少参数数量和提高计算效率。

(2)基于风格转换的属性迁移

基于风格转换的属性迁移指将人脸的属性迁移看作跨域的图像风格转换问题。Zhu 等[7]提出非配对图像翻译网络 CycleGAN，该网络由两个生成对抗网络组成，实现从源域到目标域再到源域的循环转换，并结合对抗损失和循环一致性损失将一类图像转换为另一类图像。Ma 等[21]提出双重一致性损失函数，使得生成图像与源图像在语义和风格上保持一致的情况下，学习内容图像与风格图像之间的关系。Choi 等[22]提出 StarGAN，只使用一个生成器和判别器来学习多个域之间的映射，通过改变目标图像域的标签信息来实现多个域之间的互相迁移。Sanakoyeu 等[23]提出风格感知损失函数，并通过一组编码器和解码器完成特定的艺术风格迁移。Wu 等[24]基于对偶生成对抗网络模型，对目标函数进行改进，附加两个新的损失函数并优化其参数，提升了图像翻译效果。Peng 等[25]通过在循环生成对抗网络中加入局部二值模式 LBP 算法的方法，来增强提取图像纹理特征的能力。Bao 等[26]提出 CVAE-GAN，将变分自动编码器与生成对抗网络相结合，通过改变输入到生成模型中的标签信息来生成特定类别的图像。

3 人脸性别伪造图像生成模型

3.1 模型整体结构

本文使用基于 CycleGAN 网络结构的模型，采用两对生成器和判别器，无须使用配对训练数据集，这也是本文选择以 CycleGAN 网络结构为基础的原因。基于此设计了人脸性别

转换的伪造图像生成模型，将生成的伪造图像用于扩充伪造图像数据集，同时验证了生成图像的有效性。

模型整体结构如图 1 所示，该模型主要由两个生成式对抗网络组成，图像域 X 经过生成器 G 得到伪造的图像域 Y_F，判别器 D_Y 判别伪造图像是否属于图像域 Y，伪造图像域 Y_F 经过生成器 F 得到重建的图像域 X_R。图像域 X 和重建的图像域 X_R 之间的循环一致损失保证图像翻译的有效性，指导图像域 X 到图像域 Y 的映射，减小无关图像域对图像翻译的影响。同样，图像域 Y 经过生成器 F 得到伪造的图像域 X_F，再经过生成器 G 得到重建的图像域 Y_R。

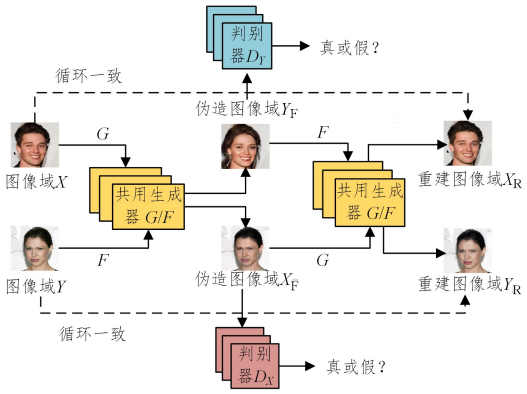


图 1　模型整体结构

Fig. 1　Structure of model

在模型训练阶段，首先，对于输入的图像域 X 和 Y，经过生成网络分别生成对应的伪造图像和重建图像；然后，计算生成网络的梯度并更新生成网络的权重参数；接着，计算判别网络的梯度并更新判别网络的权重参数；最后，按照保存模型的频率参数保存最新的模型。模型训练阶段的算法如算法 1 所示。

算法 1　模型训练阶段算法

输入：图像域 X，图像域 Y，模型训练 epoch N，总迭代次数 total_iters，保存模型频率 save_latest_freq

输出：模型文件

```
for each epoch in (1,N) do
    for each data indataset do
        生成图像域 X 和图像域 Y 对应的伪造图像和重建图像；
        将生成网络 G 和 F 的梯度设为 0；
        计算生成网络 G 的梯度；
        计算生成网络 F 的梯度；
        更新生成网络 G 和 F 的权重参数；
        将判别网络 Dx 和 Dy 的梯度设为 0；
        计算判别网络 Dx 的梯度；
        计算判别网络 Dy 的梯度；
        更新判别网络 Dx 和 Dy 的权重参数；
    end for
    if total_iters % save_latest_freq==0
        保存最新的模型；
    end if
end for
```

在模型测试阶段，首先加载已保存的最新模型；然后处理输入的图像，分别生成对应的伪造图像和重建图像；最后保存生成的伪造图像和重建图像。模型测试阶段的算法如算法 2 所示。

算法 2　模型测试阶段算法

输入：图像域 X，图像域 Y

　　最新的模型文件

输出：伪造图像域 Y_F，重建图像域 X_R，伪造图像域 X_F，重建图像域 Y_R

```
1. 加载最新的模型；
2. 设置测试参数；
3. for each data in dataset do
4.    获取输入图片信息；
5.    使用模型生成伪造图像；
6.    使用模型生成重建图像；
7.    保存生成的图像；
8. end for
```

3.2　生成器结构

为了提高生成图像的质量，使伪造图像更加真实、自然，本文引入了注意力机制和结合自适应层实例归一化的自适应残差块。生成器结构改进前后的对比如图 2 所示，生成器由下采样区域、自适应残差块和普通残差块以及注意力机制组成的中间区域和上采样区域 3 部分组成。

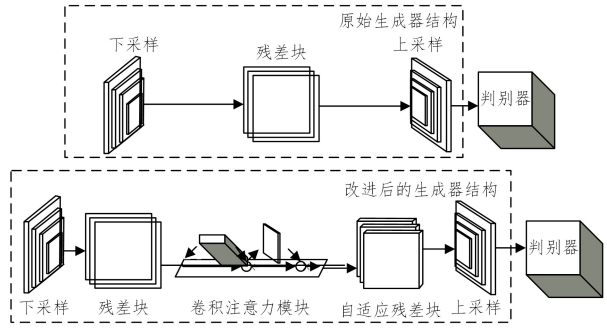


图 2　生成器结构改进前后的对比

Fig. 2　Comparison of generator structure before and after improvement

3.2.1　注意力机制

注意力机制采用仿生学的原理，模仿人类观察物品的方式，更加关注图像的局部特征，其最早应用于机器翻译领域，目前广泛应用于目标检测、图像分类等计算机视觉领域。注意力机制共分为两种：1）柔性注意力机制（soft attention），它是可微的，通过神经网络计算梯度并且经过前向传播和后向反馈来学习得到注意力的权重；2）硬性注意力机制（hard attention），它是不可微的，更加关注某一位置信息，通过增强学习来完成。本文采用卷积注意力机制[27]，如图 3 所示，它是一种结合通道维度和空间维度的注意力机制。

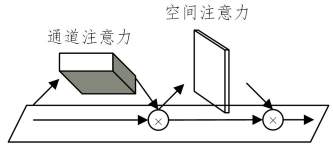


图 3　卷积注意力机制

Fig. 3　Convolutional block attention module

对于给定的特征图 $F\in R^{C\times H\times W}$，卷积注意力机制会沿着通道和空间两个独立的维度依次计算出一维的通道注意力图 $M_c\in R^{C\times 1\times 1}$ 和二维的空间注意力图 $M_s\in R^{1\times H\times W}$，整个过程如式(3)、式(4)所示。

$$F'=M_c(F)\otimes F \tag{3}$$

$$F''=M_s(F')\otimes F' \tag{4}$$

其中，$\otimes$ 表示对应元素逐个相乘。相比 SEnet[28] 只关注通道维度的注意力机制，卷积注意力机制可以取得更好的效果。通道注意力模块如图 4 所示。

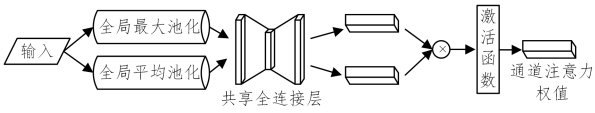


图 4 通道注意力模块

Fig. 4 Channel attention module

对于输入的特征图，首先使用全局最大池化和全局平均池化在空间维度上对其进行压缩，得到最大池化特征 $F^c_{\max}$ 和平均池化特征 F^c_{avg}，然后经过由多层感知机(Multi-Layer Perceptron，MLP)组成的共享网络和 sigmoid 激活函数得到通道注意力权值，计算过程如式(5)所示：

$$\begin{aligned}M_c(F)&=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))\\&=\sigma(W_1(W_0(F^c_{\text{avg}}))+W_1(W_0(F^c_{\max})))\end{aligned} \tag{5}$$

其中，$W_0\in R^{C/r\times C}$，$W_1\in R^{C\times C/r}$，σ 表示 sigmoid 激活函数。与通道注意力不同，空间注意力主要关注位置信息，空间注意力模块如图 5 所示。

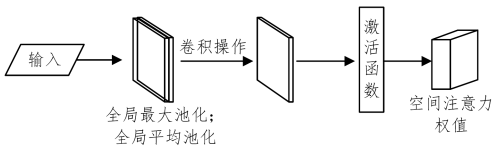


图 5 空间注意力模块

Fig. 5 Spatial attention module

首先在通道维度通过全局最大池化和全局平均池化得到两个二维特征图 $F^s_{\max}\in R^{1\times H\times W}$ 和 $F^s_{\text{avg}}\in R^{1\times H\times W}$，然后将两个特征图连接起来，使用卷积操作生成空间注意力权重图，再经过 sigmoid 激活函数得到空间注意力权值，计算过程如式(6)所示：

$$\begin{aligned}M_s(F)&=\sigma(f^{7\times 7}([AvgPool(F));MLP(MaxPool(F)]))\\&=\sigma(f^{7\times 7}([F^s_{\max};F^s_{\text{avg}}]))\end{aligned} \tag{6}$$

其中，$f^{7\times 7}$ 代表滤波器大小为 7×7 的卷积操作。

在原始模型中，通过卷积操作得到的特征图混合了跨通道和空间的特征信息，导致模型无法重点关注人脸关键区域和关键通道。

为此，本文在残差块后加入卷积注意力模块，强调模型关注通道和空间两个维度中有意义的特征。为了实现这一点，本文在卷积操作后依次应用通道和空间注意力模块，使模型既关注通道特征信息，又关注空间位置信息。通过强调通道信息和空间位置信息中有意义的特征或抑制无关的特征，使模型更加关注人脸五官，如眼、嘴等重要区域，以及该区域的关键通道，进而生成更加真实的人脸图像。

3.2.2 自适应残差块

自适应残差块指在卷积操作之间嵌入自适应实例归一化，自适应层实例归一化(adaptive layer-instance normalization)结合了实例归一化[29](instance normalization)和层归一化[30](layer normalization)，具体结构如图 6 所示。每个特征图包含图像的样式和纹理信息，实例归一化可以规范每个图像的样式，但前提是保证图像各通道之间无相关性，因为它仅对图像特征图本身做归一化。而层归一化则未假设通道之间存在相关性，它做了全局的归一化，却不能很好地保留图像样式结构。

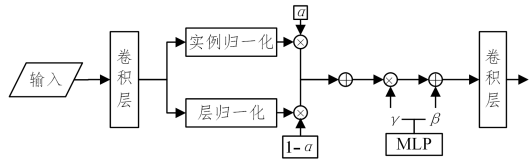


图 6 自适应残差块

Fig. 6 Adaptive residual block

自适应层实例归一化结合了实例归一化和层归一化的优点，并用它们共同指导后续残差块的工作，帮助注意力引导模型更灵活地控制样式和纹理的变化量。这种归一化方法既考虑了图像的样式结构，又考虑了各通道之间的相关性。实例归一化和层归一化的计算公式如式(9)、式(12)所示：

$$\mu_I(a)=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}a_{chw} \tag{7}$$

$$\sigma_I^2(a)=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}(a_{chw}-\mu_I(a))^2 \tag{8}$$

$$\hat{a}_I=\frac{a-\mu_I}{\sqrt{\sigma_I^2+\varepsilon}} \tag{9}$$

$$\mu_L(a)=\frac{1}{H}\sum_{h=1}^{H}a_h \tag{10}$$

$$\sigma_L^2(a)=\frac{1}{H\times W}\sum_{h=1}^{H}(a_h-\mu_L(a))^2 \tag{11}$$

$$\hat{a}_L=\frac{a-\mu_L}{\sqrt{\sigma_L^2+\varepsilon}} \tag{12}$$

$$y=\gamma[\alpha\hat{a}_I+(1-\alpha)\hat{a}_L]+\beta \tag{13}$$

$$\alpha\leftarrow \text{clip}_{[0,1]}(\alpha-\delta\triangle\alpha) \tag{14}$$

其中，$R^{H\times W\times C}$ 表示输入空间，$a\in R^{H\times W\times C}$，$h$，$w$ 为空间位置，c 为通道索引。式(13)中，γ 和 β 由多层感知机得到，α 是学习参数，δ 是学习速率，$\triangle\alpha$ 表示参数更新向量。α 在参数更新中被限制在[0，1]之间，以表示样式对当前任务的重要程度，α 越大越重要。

3.3 判别器结构

在判别网络中，用 70×70 的 PatchGAN 网络结构[6]来判别感受野为 70×70 的局部 patch 是否为真。传统的生成对抗网络的判别器是将输入映射成一个实数，以此来表示生成图像为真的概率；而 PatchGAN 是将输入特征图映射为一个 30×30 大小的输出，相当于对应输入特征图的 900 个 70×70

的局部 patch 为真的概率。用判别器对整个 $N\times N$ 大小的图像进行卷积，得到 30×30 大小的输出，然后求平均值，即为判别器最后的输出值。相比传统的判别器，PatchGAN 的感受野对应于输入特征图中的一小块区域，更关注图像细节。

3.4　损失函数

非配对图像翻译网络 CycleGAN 由两对生成器和判别器组成，用来将一类图片转换为另一类图片。该模型在两个生成对抗网络中都使用了对抗损失和循环一致性损失。假设现有两个图像域 X 和 Y，生成器 G 学习从 X 到 Y 的映射，对于 G 生成的图像，判别器 D_Y 判别它是否为真实图像，其对抗损失函数如下所示：

$$\max\ \mathcal{L}_{\mathrm{GAN}}(D_Y)=\mathbb{E}_{y\sim P_{\mathrm{data}}(y)}[\log D_Y(y)]+E_{x\sim P_{\mathrm{data}}(x)}[\log(1-D_Y(G(x)))] \tag{15}$$

$$\min\ \mathcal{L}_{\mathrm{GAN}}(G)=\mathbb{E}_{y\sim P_{\mathrm{data}}(y)}[\log D_Y(y)]+E_{x\sim P_{\mathrm{data}}(x)}[\log(1-D_Y(G(x)))] \tag{16}$$

本文借鉴了文献[31]的思想，将输入数据一半为真一半为假的先验知识代入判别器以提高判别器的判别能力，将绝对真假变成相对真假。将 $D_Y(y)$ 改为 $(D_Y(y)-D_Y(G(x)))$，将 $D_Y(G(x))$ 改为 $D_Y(G(x))-D_Y(y)$ 并改进损失函数。同时，融合最小二乘算法[11]，将对数运算变为残差平方运算，对判别为真但远离真实样本的假样本进行优化，更严格地惩罚远离决策边界的假样本，从而提高生成图像的质量。

$$\min\ \mathcal{L}_{\mathrm{RLSGAN}}(D_Y)=\frac{1}{2}\mathbb{E}_{\substack{y\sim P_{\mathrm{data}}(y)\\ x\sim P_{\mathrm{data}}(x)}}[(D_Y(y)-D_Y(G(x))-1)^2+(D_Y(G(x))-D_Y(y))^2] \tag{17}$$

$$\min\ \mathcal{L}_{\mathrm{RLSGAN}}(G)=\frac{1}{2}\mathbb{E}_{\substack{y\sim P_{\mathrm{data}}(y)\\ x\sim P_{\mathrm{data}}(x)}}[(D_Y(y)-D_Y(G(x)))^2+(D_Y(G(x))-D_Y(y)-1)^2] \tag{18}$$

同理，另一个生成器 F 学习从 Y 到 X 的映射，对于 F 生成的图像，判别器 D_X 判别它是否为真实图像，然后将 D_X 对 F 生成图像的判定结果反馈给 F 来指导网络进行训练。除了对抗损失外，还额外引入循环一致性损失。在学习 X 到 Y、Y 到 X 的映射时，图像域 X 经过生成器 G 生成伪造的图像域 Y_{F}，再经过 F 生成重建的图像域 X_{R}，使 X 与 X_{R} 之间的差异尽量减小，并计算两者之间的损失。同理，也应尽量使 Y 与 Y_{R} 之间的差异减小，循环一致性损失函数如式(19)所示：

$$\mathcal{L}_{\mathrm{CYC}}(G,F)=E_{y\sim P_{\mathrm{data}}(y)}[\|G(F(y))-y\|_1]+E_{x\sim P_{\mathrm{data}}(x)}[\|F(G(x))-x\|_1] \tag{19}$$

经过以上分析，可将模型整体损失函数表示为：

$$\mathcal{L}_{\mathrm{GL}}(G,F,D_X,D_Y)=\mathcal{L}_{\mathrm{RLSGAN}}(G,D_Y,X,Y)+\mathcal{L}_{\mathrm{RLSGAN}}(F,D_X,Y,X)+\lambda\mathcal{L}_{\mathrm{CYC}}(G,F) \tag{20}$$

其中，λ 为循环一致性损失的权重，模型最终的目标为：

$$\arg\min_{G,F}\max_{D_X,D_Y}\mathcal{L}_{\mathrm{GL}}(G,F,D_X,D_Y) \tag{21}$$

4　实验与分析

本文以性别转换为切入点进行人脸伪造图像合成。从生理学来讲，人脸会随着年龄的变化发生不可逆转的改变，面部骨骼也随着持续衰老发生动态变化，受此启发，本文采用基于年龄约束的训练策略来训练生成模型。

4.1　数据集

本文在选择数据集的过程中着重考虑带有性别和年龄标签的属性，但收集具有广泛生物学信息的人脸数据集具有挑战性。综合考虑后，本文选用公开数据集 CelebA 以及 IMDB-WIKI 数据集。CelebA 数据集包含 10 177 个人物身份的 202 599 张人脸图像，而且每张照片都有特征标注信息，包含性别、是否年老以及人脸特征点坐标等 40 多项信息。IMDB-WIKI 数据集包含从 IMDB 与 Wikipedia 中分别爬取的 461 871 张、62 359 张人脸图像，共计 524 230 张人脸图像。本文生成模型的训练集来自 CelebA 数据集，而测试集来自 CelebA，IMDB 与 WIKI 数据集。根据实验目的，本文根据图像标注信息进行数据集预处理，将数据集分为男性年轻(male-young)、男性年老(male-old)、女性年轻(female-young)、女性年老(female-old)共 4 组实验数据，并使用 python 开源计算机视觉库 OpenCV2 中的 resize()函数将图片大小调整为 256×256，以便进行训练。CelebA 数据集中人脸图像按实验分组的数据分布如图 7 所示。

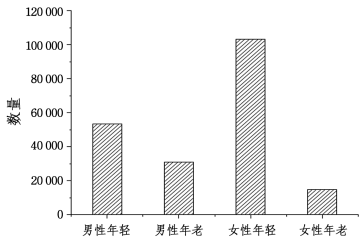


图 7　CelebA 数据集中人脸图像实验数据分布

Fig. 7　Distribution of facial images in CelebA

4.2　实验细节

本文实验环境如表 1 所列。

表 1　实验环境

Table 1　Environment of experiment

实验环境	版本
操作系统	Ubuntu 16.04 LTS
CPU	Intel(R) Core(TM) i9-10920X CPU @ 3.50GHz
GPU	NVIDIA RTX 3090
Pytorch	1.7.0
CUDA	11.0

本文在预先分好组的数据集之间进行模型的训练，将输入和输出图像大小均设置为 256×256。在实验中，将批处理大小 batch size 设为 1，前 100 个 epoch 的学习率设置为 0.000 2，线性衰减为 0，将式(20)中的 λ 设置为 10。在训练过程中，采用 Adam[32]优化器进行梯度下降的优化。在生成器结构中，残差块的数量为 12，其中包括 6 个自适应残差块。

4.3　评价指标

对于生成对抗网络生成的人脸图像，评价其质量存在非常主观的因素，因为计算机很难像肉眼一样发现图片的细微变化。因此，本文采用主观视觉评价与客观指标评价相结合

的方法，客观评价指标采用内容准确率和结构相似度相结合的综合评价指标。为保证结果的客观性，在生成数据集中随机抽取 2000 张图像作为样本进行客观评价指标的计算。

(1)内容准确率。内容准确率能够很好地反映伪造图像与真实图像在内容上的差异。本文使用在 Adience 数据集上预训练的 Inception v3 网络作为分类模型。如表 2 所列，预训练的分类模型能够在 CelebA 数据集和 IMDB-WIKI 数据集的真实图像上达到较高的内容准确率。从理论上讲，如果生成的伪造图像足够真实，便能欺骗分类模型，使其作出错误的判断。我们将使分类模型作出错误判断的样本计入分类正确的样本数。以原始方法生成的伪造图像在分类模型上取得的内容准确率为基准，将使用本文方法生成的伪造图像送入分类模型进行分类，在得到其内容准确率后与基准值进行对比，准确率越高，代表伪造的图像越真实，效果越好。内容准确率的计算公式为：

$$\text{内容准确率}=\frac{\text{分类正确的样本数}}{\text{总样本数}} \tag{22}$$

表 2 Inception v3 分类模型在真实图像上的内容准确率

Table 2 Content accuracy of Inception v3 dassification model on real images

		真实女性	真实男性
内容准确率	CelebA	0.972	0.856
	Imdb	0.971	0.979
	Wiki	0.987	0.973

(2)结构相似度。本文选用 FID(Fréchet Inception Distance)指标用于衡量两组图像数据分布之间的相似度。FID 表征了真实图像与生成图像在特征空间上的距离，FID 分数越低，代表生成的图像质量越高。首先使用 Inception 网络提取全连接层之前的特征向量，然后使用高斯模型对特征空间进行建模，最后通过高斯模型的均值和协方差计算图像在特征空间上的距离，计算式如式(23)所示：

$$FID=\|\mu_1-\mu_2\|_2^2+T_r(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2-2(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)^{\frac{1}{2}}) \tag{23}$$

其中，μ_1 和 $\boldsymbol{\Sigma}_1$ 为真实数据集的均值和协方差矩阵，μ_2 和 $\boldsymbol{\Sigma}_2$ 为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

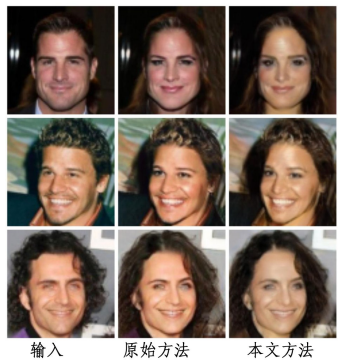
本文采用年龄约束的方法来训练生成模型，在年老组和年轻组两类数据集上分别进行原始方法和本文方法的对比实验。这样可以在进行横向对比的同时利用每种方法根据年龄变化再进行纵向对比。除此之外，本文还利用每种方法分别进行男性到女性的伪造实验和女性到男性的伪造实验，目的是更加全面地考察每种方法并进行对比。另外，本文采用控制变量的思想，所做实验均采用经过 100000 次迭代的生成模型，且同一种实验采用相同的测试数据，只保留生成方法和训练数据的不同。实验结果如图 8、图 9 所示。

(a)男性到女性

(b)女性到男性

图 8 年老组人脸性别伪造结果

Fig. 8 Facial gender forgery for old people face images



(a)男性到女性

(b)女性到男性

图 9 年轻组人脸性别伪造结果

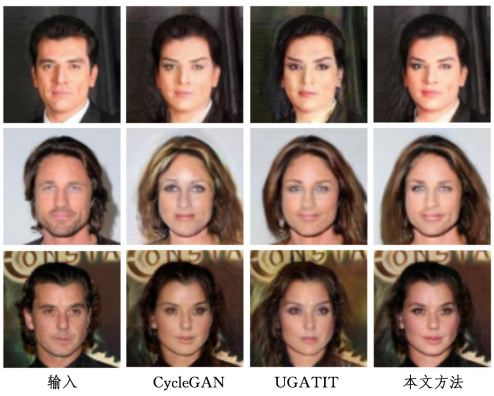
Fig. 9 Facial gender forgery for young people face images

年老组的伪造结果如图 8 所示，第二列为采用原始方法生成的人脸伪造图像，第三列为采用本文方法生成的人脸

伪造图像，图8(a)为男性到女性的伪造实验，图8(b)为女性到男性的伪造实验，下述实验结果展示均采用相同的分布。可以看出，本文方法在人脸部的细节变化上更加真实。例如，在男性到女性的伪造中的最后一组实验中，原始方法相比本文所提方法眼部变化更加夸张；在女性到男性的伪造实验中，采用本文方法生成的人脸图像嘴部有胡子的变化，这一变化在现实生活中是真实存在的。

相比年老组的伪造实验，年轻组的变化并不是特别明显，伪造结果如图 9 所示。在女性到男性的伪造实验中，虽然两种方法都表现一般，但在样式上仍发生了很多变化，如头发的变化。在男性到女性的伪造实验中，本文方法与原始方法相比，除了样式上的变化更加真实外，纹理也更加细腻，例如第二组实验的人脸眼部和嘴部的变化。

将本文方法与采用原始 CycleGAN 方法、UGATIT 方法生成的人脸性别伪造图像进行主观对比，实验结果如图 10 所示。



(a)男性到女性

(b)女性到男性

图 10　不同方法实验结果对比

Fig. 10　Comparison of multiple image translation methods

从实验结果可以看出，采用 CycleGAN 方法生成的人脸性别伪造图像效果最差，但清晰度较好；采用 UGATIT 方法生成的人脸性别伪造图像在多样性上表现更好，但有时会改变图像的整体色调，导致伪造痕迹明显；本文方法兼具图像的多样性和真实性，在主观视觉上的效果更好。将两部分实验结果进行对比可以发现，以上 3 种方法在男性到女性伪造实验中的效果优于女性到男性伪造实验中的效果。图10(a)包含更加丰富的纹理，相较于下半部分更加逼真。具体来看，在男性到女性的伪造实验中，最右侧图像相比中间图像五官更加细腻，妆容更加自然。虽然 3 种方法在女性到男性的伪造实验中表现一般，但最右侧图像仍优于中间图像。例如，在中间一组输入图像中，原始方法生成的伪造图像头发样式很不自然且眼部发生了扭曲变形。总体来说，在迭代次数相同的情况下，采用本文方法进行人脸性别伪造的效果在主观视觉上优于其他方法。

4.4.2　客观指标评价

(1)模型改进前后的对比

为了进一步体现所做改进工作带来的性能增益，本文在 CycleGAN 的基础上逐步增加卷积注意力机制和自适应层实例归一化，分别计算在不同改进策略下的内容准确率和 FID 得分。

如表 3 所列，添加卷积注意力机制后生成模型更加关注人脸部特征的学习，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 0.019 和 0.129；继续添加自适应层实例归一化，帮助注意力引导模型更灵活地控制样式和纹理的变化量，内容准确率再提高了 0.045 和 0.196。如表 4 所列，模型中添加卷积注意力机制后，在 CelebA 数据集上，伪造女性和伪造男性的 FID 得分分别降低了 3.55 和 7.12；继续增加自适应层实例归一化后，FID 再降低了 3.40 和 2.26。从表 3 和表 4 可以看出，生成模型在 IMDB 与 WIKI 数据集上的表现也十分出色，其内容准确率、FID 得分与 CelebA 数据集上的实验结果指标接近。这说明生成模型具备在多个数据集上的泛化能力。

表 3　不同条件下内容准确率的对比

Table 3　Content accuracy under different conditions

		CycleGAN	CycleGAN+CBAM	CycleGAN+CBAM+AdaLIN
男性到女性	CelebA	0.877	0.896	0.941
	Imdb	0.691	0.716	0.871
	Wiki	0.670	0.762	0.893
女性到男性	CelebA	0.387	0.516	0.712
	Imdb	0.344	0.486	0.686
	Wiki	0.364	0.478	0.703

表 4　不同条件下 FID 得分的对比

Table 4　FID score under different conditions

		CycleGAN	CycleGAN+CBAM	CycleGAN+CBAM+AdaLIN
男性到女性	CelebA	37.26	33.71	30.31
	Imdb	38.41	36.46	34.42
	Wiki	36.54	34.36	32.16
女性到男性	CelebA	43.35	36.23	33.97
	Imdb	43.16	38.24	35.27
	Wiki	40.16	36.26	36.16

使用类激活映射(Class Activation Mapping，CAM)可视化后，每种改进策略下的性能增益结果如图 11 所示，红色区域表示生成模型更加侧重的部分。CycleGAN 网络对于人脸面部的特征关注较少，其更多地关注头发等背景信息，导致

生成图像的人脸面部变化不明显。增加卷积注意力机制后，生成模型更加关注人脸面部特征的学习，但学习并不全面。在继续增加自适应层实例归一化后，人脸部特征的通道被赋予更高的权重，生成模型更加关注人脸面部特征的学习，从而使生成图像更加逼真。

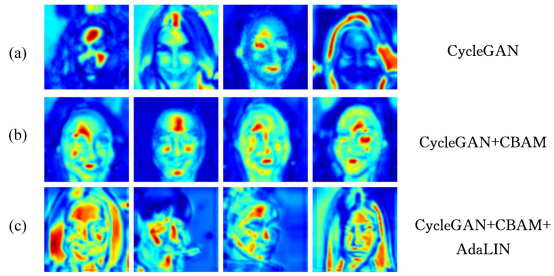


图 11 每种改进策略下的 CAM 热力图(电子版为彩色)

Fig. 11 CAM heatmaps generated by each improved strategy

综合以上对比实验来看，卷积注意力机制使生成模型更加关注人脸面部关键特征的提取，减少了无关区域的变化；自适应层实例归一化丰富了伪造图像的细节信息。实验证明了融合卷积注意力机制与自适应层实例归一化方法的有效性，所做改进工作提升了生成图像的真实性和多样性。

(2)与其他方法的对比

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 5、表 6 所列。本文方法在伪造女性和伪造男性的实验中内容准确率均高于其他方法，说明采用本文方法生成的图像更加真实，且生成的伪造图像细节更加丰富。采用本文方法生成的伪造女性和伪造男性图像的 FID 得分在所有方法中最低，说明生成的图像质量更高，具有更好的多样性。

表 5 不同方法的内容准确率对比

Table 5 Content accuracy under different methods

		CycleGAN	UGATIT	本文方法
男性到女性	CelebA	0.877	0.872	0.941
	Imdb	0.691	0.824	0.871
	Wiki	0.670	0.741	0.893
女性到男性	CelebA	0.387	0.472	0.712
	Imdb	0.344	0.432	0.686
	Wiki	0.364	0.453	0.703

表 6 不同方法的 FID 得分对比

Table 6 FID score under different methods

		CycleGAN	UGATIT	本文方法
男性到女性	CelebA	37.26	51.69	30.31
	Imdb	38.41	50.46	34.42
	Wiki	36.54	52.36	32.16
女性到男性	CelebA	43.35	52.90	33.97
	Imdb	43.16	51.24	35.27
	Wiki	40.16	48.26	36.16

通过在不同数据集上的实验结果可以看出，IMDB-WIKI 数据集上的内容准确率和 FID 得分均差于在 CelebA 数据集上的实验结果，但相差不大，这可能与 IMDB 和 WIKI 数据集人脸图像不对齐、图像拍摄形式多样化等原因有关。在未来的工作中，还将继续提高生成模型的泛化能力。

结束语 本文借鉴图像翻译的思想进行人脸性别伪造图像的生成。首先在非配对图像翻译网络 CycleGAN 的基础上，提出了一种融合卷积注意力机制和自适应残差块的人脸性别伪造图像生成模型；然后，采用相对损失的思想改进损失函数，提高了判别器的判别能力；最后，提出了基于年龄约束的模型训练方法，减小了年龄变化对生成图像的影响。实验结果表明，采用本文方法生成的人脸性别伪造图像在真实性和多样性上的表现更加出色。

虽然本文工作提高了伪造图像的质量，但依然无法完全避免无关图像域的改变，未来希望能够提高模型的稳定性和泛化能力并将其应用到其他人脸属性的伪造中，同时希望能够建立各种人脸属性伪造的数据集。

参 考 文 献

[1] BAO Y X, LU T L, DU Y H. Overview of Deepfake Video Detection Technology[J]. Computer Science, 2020, 47(9): 283-292.

[2] Deepfakes[OL]. https://github.com/deepfakes/faceswap.

[3] FaceSwap[OL]. https://github.com/MarekKowalski/FaceSwap/.

[4] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2387-2395.

[5] Faceswap-GAN[OL]. https://github.com/shaoanlu/faceswap-GAN.

[6] 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. 2017: 1125-1134.

[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 Vomputer vision. 2017: 2223-2232.

[8] KIM J, KIM M, KANG H, et al. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation[J]. arXiv: 1907.10830, 2019.

[9] WANG Z, TANG X, LUO W, et al. Face aging with identity-preserved conditional generative adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7939-7947.

[10] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// International Conference on Machine Learning. PMLR, 2017: 214-223.

[11] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2794-2802.

[12] GULRAJANI I, AHMES F, ARJOVSKY M, et al. Improved training of wasserstein gans[J]. arXiv: 1704.00028, 2017.

[13] UPCHURCH P, GARDNER J, PLEISS G, et al. Deep feature interpolation for image content changes[C]// Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. 2017：7064-7073.

[14] PERARNAU G，VAN DE WEIJER J，et al. Invertible conditional gans for image editing[J]. arXiv：1611. 06355，2016.

[15] WANG S M，LI S F. Multi-domain image conversion method based on feature vector transformation GAN[J]. Journal of Yunnan University（Natural Sciences Edition），2020，42（6）：1080-1090.

[16] LIU M Y，BREUEL T，KAUTZ J. Unsupervised image-to-image translation networks[J]. arXiv：1703. 00848，2017.

[17] 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）. 2018：172-189.

[18] PARK D Y，LEE K H. Arbitrary style transfer with style-attentional networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：5880-5888.

[19] XIAO T，HONG J，MA J. Elegant：Exchanging latent encodings with gan for transferring multiple face attributes[C]// Proceedings of the European Conference on Computer Vision（ECCV）. 2018：168-184.

[20] CHO W，CHOI S，PARK D K，et al. Image-to-image translation via group-wise deep whitening-and-coloring transformation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：10639-10647.

[21] MA Z，LI J，WANG N，et al. Semantic-related image style transfer with dual-consistency loss[J]. Neurocomputing，2020，406：135-149.

[22] CHOI Y，CHOI M，KIM M，et al. Stargan：Unified generative adversarial networks for multi-domain image-to-image translation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：8789-8797.

[23] SANAKOYEU A，KOTOVENKO D，LANG S，et al. A style-aware content loss for real-time hd style transfer[C]// Proceedings of the European Conference on Computer Vision（ECCV）. 2018：698-714.

[24] WU H M，LIU Q R，WANG Y H. Face image translation based on generative adversarial networks[J]. Journal of Tianjin University：Science and Technology，2019，52（3）：306-314.

[25] PENG Y F，WANG K X，MEI J Y，et al. Image style migration based on cycle generative adversarial networks[J]. Computer Engineering & Science，2020，42（4）：699-706.

[26] BAO J，CHEN D，WEN F，et al. CVAE-GAN：fine-grained image generation through asymmetric training[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2745-2754.

[27] WOO S，PARK J，LEE J Y，et al. Cbam：Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision（ECCV）. 2018：3-19.

[28] HU J，SHEN L，SUN G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7132-7141.

[29] ULYANOV D，VEDALDI A，LEMPITSKY V. Instance normalization：The missing ingredient for fast stylization[J]. arXiv：1607. 08022，2016.

[30] BA J L，KIROS J R，HINTON G E. Layer normalization[J]. arXiv：1607. 06450，2016.

[31] JOLICOEUR-MARTINEAU A. The relativistic discriminator：a key element missing from standard GAN[J]. arXiv：1807. 00734，2018.

[32] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv：1412. 6980，2014.

SHI Da，born in 1997，master. His main research interests include cyber security and artificial intelligence.

LU Tian-liang，born in 1985，Ph.D，associate professor，is a member of China Computer Federation. His main research interests include cyber security and artificial intelligence.

（责任编辑：喻藜）