

融合小波快捷连接生成对抗网络的面部性别伪造

陈万泽, 黄丽清, 陈家祯, 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, 福建 福州 350117)

摘要: 面部属性编辑领域的主流方法由于数据以及模型架构的限制, 存在如下两方面的问题: 自编码器模型在编码以及解码阶段的瓶颈结构会丢失特征信息, 以及在解码过程中逐阶段对源域特征进行连续的样式注入, 使得生成图像中目标域信息占比过大, 导致生成图像丢失了源域的身份信息以及部分细节; 在人脸图像中, 图像人物的性别、种族或年龄等属性的差异会使得不同图像的频域信息组成差别较大, 在无监督的训练前提下, 当前主流网络框架并不能在样式注入阶段调整源域信息以及目标域信息之间的占比, 导致生成图像仍然存在伪影。针对上述问题, 基于生成对抗网络和图像到图像的翻译提出一种面部性别伪造模型, 即融合快捷小波连接生成对抗网络(WscGAN)。通过对自编码器结构添加快捷连接, 将不同编码阶段的输出通过小波变换进行特征级别的分解, 再引入通道注意力机制将其进行逐个处理, 以此动态改变在解码过程中不同频率的源域特征信息的比例, 最终实现面部图像在性别属性上的伪造。为了验证所提模型的有效性, WscGAN 分别在 CelebA-HQ 数据集和 FFHQ 数据集上进行了实验。实验结果表明, WscGAN 在 CelebA-HQ 以及 FFHQ 数据集上均领先现有先进模型, 其中弗雷歇起始距离分别提升 5.4% 和 19.8%, 特征度量相似度分别提升 1.8% 和 4.1%。此外, 定性的视觉对比结果充分体现了 WscGAN 可有效改善面部图像性别属性伪造的效果。

关键词: 图像生成; 生成对抗网络; 图像到图像的翻译; 面部属性编辑; 小波变换

中图分类号: TP37; TP391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023046

Gender forgery of faces by fusing wavelet shortcut connection generative adversarial network

CHEN Wanze, HUANG Liqing, CHEN Jiazhen, YE Feng, HUANG Tianqiang, LUO Haifeng

College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China

Abstract: The mainstream methods in the field of facial attribute manipulation had the following two defects due to

收稿日期: 2022-12-04; 修回日期: 2023-03-02

通信作者: 黄丽清, lqhuang@fjnu.edu.cn

基金项目: 国家自然科学基金(62072106); 福建省自然科学基金(2020J01168); 福建省教育厅(A类)(JAT210053)

Foundation Items: The National Natural Science Foundation of China (62072106), Fujian Province Natural Science Foundation (2020J01168), Fujian Provincial Department of Education (Rank A) (JAT210053)

引用格式: 陈万泽, 黄丽清, 陈家祯, 等. 融合小波快捷连接生成对抗网络的面部性别伪造[J]. 网络与信息安全学报, 2023, 9(3): 150-160.

Citation Format: CHEN W Z, HUANG L Q, CHEN J Z, et al. Gender forgery of faces by fusing wavelet shortcut connection generative adversarial network[J]. Chinese Journal of Network and Information Security, 2023, 9(3): 150-160.

data and model architecture limitations. First, the bottleneck structure of the autoencoder model results in the loss of feature information, and the traditional method of continuously injected styles to the source domain features during the decoding process makes the generated image too referential to the target domain while losing the identity information and fine-grained details. Second, differences in facial attributes composition between images, such as gender, ethnicity, or age can cause variations in frequency domain information. And the current unsupervised training methods do not automatically adjust the proportion of source and target domain information in the style injection stage, resulting in artifacts in generated images. A facial gender forgery model based on generative adversarial networks and image-to-image translation techniques, namely fused wavelet shortcut connection generative adversarial network (WscGAN), was proposed to address these issues. Shortcut connections were added to the autoencoder structure, and the outputs of different encoding stages were decomposed at the feature level by wavelet transform. Attention mechanism was employed to process them one by one, to dynamically change the proportion of source domain features at different frequencies in the decoding process. This model could complete forgery of facial images in terms of gender attributes. To verify the effectiveness of the model, it was conducted on the CelebA-HQ dataset and the FFHQ dataset. Compared with the existing optimal models, the method improves the FID and LPIPS indices by 5.4% and 11.2%, and by 1.8% and 6.7%, respectively. Furthermore, the effectiveness of the proposed method in improving the gender attribute conversion of facial images is fully demonstrated by the results based on qualitative visual comparisons.

Keywords: image generation, generative adversarial network, image-to-image translation, facial attribute manipulation, wavelet transform

0 引言

近年来,随着互联网技术的发展,短视频以及直播逐渐成为移动时代用户的常态化表达方式,同时自拍照片也成为人们分享生活的重要组成部分,这些用户需求刺激了对面部属性编辑技术的研究。面部属性编辑在社交平台上的广泛应用使短视频具有趣味性的同时满足了用户对自身形象进行改变的需求,不仅如此,面部属性编辑在人机交互和娱乐社交具有广泛应用,具有巨大的市场潜力。因此,研究人脸属性编辑具有重要的经济意义和社会价值。但是,深度伪造技术合成的虚假图像会严重损害公民肖像权以及危害个人隐私,已成为威胁网络安全的重要因素之一。目前,由 Goodfellow 等^[1]提出的生成对抗网络(GAN, generative adversarial network)得到了广泛的发展,并且由于条件生成对抗网络(cGAN, conditional GAN)^[2]能够合成满足输入标签的图像,研究人员利用 GAN 和 cGAN 这两种成熟技术实现了图像到图像的翻译(I2I, image-to-image translation)。众多基于生成对抗网络的模型在 I2I 中取得了优异的表现,其广泛应用于计算机视觉

和图像处理领域,如图像合成^[3-4]、图像分割^[5]、风格迁移^[6]、图像修补^[7]、人体姿态估计^[8],以及面部属性操作^[9]。

I2I 旨在将图像从源域转移到目标域,保留内容表示的同时需满足目标域的属性要求。其中源域和目标域可以看作图像不同属性的集合,最终模型学习到的是不同视觉域之间的映射关系。CycleGAN^[10]利用 GAN 实现无监督未配对数据 I2I 任务,之后多种 GAN 的变体被应用于 I2I^[11-12]。I2I 的主要输入是图像以及标签,故大多数 I2I 领域的模型将传统 cGAN 的生成器替换为自编码器(AE, autoencoder)^[13]结构形成对抗式自编码器(AAE, adversarial autoencoder)^[14]。尽管上述方法取得了令人振奋的进展,但受数据以及模型架构的限制,其在性别属性操作方面仍存在缺陷。首先,生成器为自编码器的结构时,输入图像需要被编码器压缩为较小尺寸的特征,解码需要在特征之上通过样式迁移完成属性转换,整个过程中往往受到 AE 结构的影响丢失重要的细粒度细节。在解码过程中逐阶段对源域特征进行连续的样式注入,会导致生成图像丢失源域的身份信息,同时使生成图像的样式过于参考目标域,导致视

觉质量不理想。其次,源域图像与目标域图像在生成图像中保留的信息会根据不同的翻译方向产生变化,但面部图像中人物的性别、种族或年龄等属性的差异会导致图像间频域信息组成的差别较大,并且传统方法没有针对频域信息进行优化,最终导致生成图像存在伪影。根据生成图像与源域图像的主要变动范围,判断男女性别域转换的理想情况,即保持面部的身份信息不变,源域所独有的独特属性(如男性的胡须、女性的长发)消失,转而拥有目标域的特色属性。主要修改区域集中在拥有高频信息的头发与包含低频信息的背景。因此在不同的转换方向,如何使源域和目标域的不同频率信息更好地融合成为性别转换 I2I 中的难点。

针对上述问题,本文基于 GAN 的 I2I 技术提出一种全新的融合小波快捷连接生成对抗网络(WscGAN, wavelet shortcut connection GAN)模型,实现了面部图像在性别属性上的伪造。模型同时满足基于标签和基于参考的图像合成,并且能够输出不同风格的图像,均符合目标域的语义要求。首先,为了更好地保留源域图像的特征,本文对传统 I2I 中常用的生成器结构进行修正,在自编码器结构中添加快捷连接,将下采样过程中各个阶段的特征予以保留,使得各阶段的编码信息可以输入相对应的解码特征中,不仅减轻了生成器合成丰富细节的压力,而且防止生成图像丢失源域图像的身份信息。其次,为防止生成图像过度参考目标域信息,并能够实现在生成过程中模型可以更好地处理不同频域的信息,本文通过小波变换将编码特征分解为 4 个不同频带的特征,并添加通道注意力机制对不同频带的信息进行逐个处理,最终在解码阶段与目标域特征信息进行融合,使得生成结果在源域与目标域的信息之间做出了理想的权衡。

为了验证所提模型的有效性,本文在 CelebA-HQ 数据集^[15]进行面部性别属性的转换实验,并与现有先进模型进行比较,实验结果表明,本文模型除了较好地保留了源域面部身份信息外,生成结果也更好地符合目标域的属性要求。此外,为了进一步验证本文模型的鲁棒性,利用在年龄、种族和图像背景变化较大的 FFHQ 数据

集^[16]进行额外的验证,最终取得了较为出色的结果。综合实验结果表明,本文模型不仅在定量指标中优于现有的先进模型,而且定性视觉结果也取得了显著改善。

1 相关工作

1.1 生成对抗网络 GAN

继 Goodfellow 等^[1]提出 GAN 之后,随着 GAN 体系结构^[17]、损失函数^[18-19]、正则化^[20]等技术的不断改进,GAN 训练过程趋于稳定,且生成的图像也更加真实。例如,WGAN^[19]利用瓦瑟斯坦距离重新设计了 GAN 的训练,平衡生成器和判别器的训练过程,试图解决 GAN 模式崩溃的问题。同时 GAN 有许多变体,深度卷积生成对抗网络^[21]用卷积神经网络代替多层感知器,以增强图像生成的能力。条件生成对抗网络^[2]通过嵌入额外的条件来控制图像的生成。InfoGAN^[22]在 cGAN 中引入互信息最大化技术,确保嵌入条件能够尽可能地调节图像的生成。BIGGAN^[23]通过正则化等方法保证大型生成对抗网络训练过程的稳定性。StyleGAN^[16]通过无监督方式对图像的高层语义属性进行一定解耦分离,最终实现了可控的图像生成。

1.2 基于 GAN 的 I2I

I2I 的目标是将输入图像 x 从源域 y 转换到目标域 y_t ,即原图像的整体风格变换为与目标域相同的样式,并且保留大部分 x 的内容。上述过程中的源域和目标域是一种标签,代表着图像的某种属性。例如,在人脸性别属性伪造的过程中,试图将一张男性照片 x 根据另一幅女性参考图像 x_t 转化为女性,理想结果是保留男性的身份信息,但性别特征(如头发、肤色等)会变为参考图像的样式。因此,需要训练一个模型来实现这种映射关系,用公式表示为 $T_{y \rightarrow y_t}$,其中 T 表示转换过程,在数学上可以将这个翻译过程建模为 $x_t = T_{y \rightarrow y_t}(x)$, x_t 是符合目标域特征的生成图像。

pix2pix^[12]通过 cGAN 使用成对数据(如图 1 所示)实现 I2I 任务,其对生成图像以及对应域的真实图像之间使用像素级 L1 损失,还添加条件对抗损失确保生成图像的质量。Tang 等^[24]提出基于级联语义引导下的多通道注意力图像翻译,

解决了跨视角图像翻译问题。Park 等^[25]证明了正则化层会减弱图像中原有的语义信息, 进而提出了空间自适应正则化, 进一步提升图像翻译的质量。针对未配对图像的 I2I, CycleGAN^[10]为未配对的图像转换引入了周期一致性损失, 假定源域中的图像映射到目标域之后, 还能通过同样的方法映射回源域。MUNIT^[26]假设图像表示可以分解为域不变的内容特征以及特定域的风格特征, 实现了多个域之间的互相转换。通过引入一个域编码器, DRIT++^[27]进一步将 DRIT 扩展到多域图像转换。StarGAN^[28]使用目标域标签的嵌入以及添加辅助判别器进行分类, 利用单一的 GAN 模型实现多域的 I2I。PHStarGAN-v2^[29]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。



图1 I2I 中配对的图像数据
Figure 1 Paired image data in I2I

1.3 小波变换在深度学习的应用

频域分析在传统图像的处理中得到了广泛的应用, 其关键思想是根据图像空间域内的变化速度, 将像素从欧几里得空间映射至频率空间。与傅里叶分析等其他频率分析方法相比, 小波变换可以同时捕获空间信息和频率信息, 因此被认为是图像处理手段中更高效的一种方法。随着深度学习的发展, 一些研究者尝试将频域分析引入深度学习框架中以提升性能。

Demirel 等^[30]提出了基于离散小波变换 (DWT, discrete wavelet transform) 对高频子带图像进行插值的分辨率增强技术。Chen 等^[31]认为下采样操作会导致信息损失, 借助小波分解思想, 将原图分解为高频信息与低频信息, 在解码阶段利用这两个分量恢复原始图像。

Huang 等^[32]提出了基于小波卷积神经网络的人脸超分辨率模型, 通过预测高分辨率图像

的小波系数, 可以生成不同分辨率下的面部图像。Fujieda 等^[33]在同一个模型中集成了卷积神经网络以及小波变换操作, 通过纹理信息解决分类问题。Williams 等^[34]对小波池化进行了研究, 认为最大池化或均值池化是通过舍弃部分信息从而实现正则化的效果, 故提出通过离散小波变换替代池化操作, 从而保留图像中的全局信息。

2 WscGAN 模型

WscGAN 模型专注于面部性别属性伪造的问题, 其使用端到端的训练, 并可根据输入的标签属性精确编辑生成高质量的人脸图像。

2.1 模型结构

本文的目的是得到一个多域转换的模型。具体来说, 给定一个真实的图像 $x \in X$ 和一个随机的目标域标签 $y_t \in Y$, 将 x 翻译成具有目标域风格的图像 x_t 。其中, x 所属的源域标签为 $y \in Y$, X 代表数据集中所有图片, Y 代表数据集中所有图像所属域的集合。网络架构分别由生成器 (G)、判别器 (P)、样式提取器 (E) 以及样式映射器 (M) 组成, 如图 2 所示。

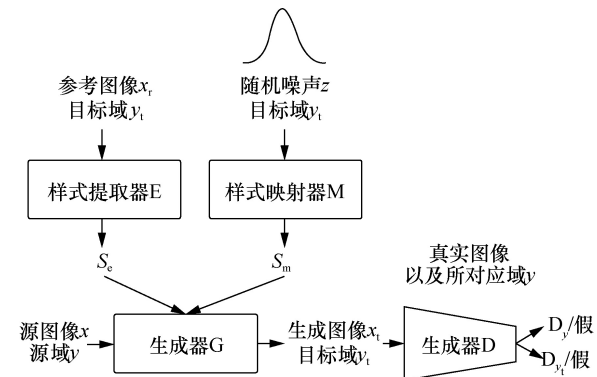


图2 网络架构
Figure 2 Network architecture

2.1.1 生成器

生成器的整体结构如图 3 所示。其输入包括 x 以及样式特征 s , 其中样式特征 $s \in \{s_m, s_e\}$ 来自样式映射器和样式提取器。G 的任务是将原始图像 x 根据包含目标域信息的样式特征翻译出图像, 不仅需要保留面部身份信息, 且图像整体需符合目标域语义要求, 生成过程用公式表示为

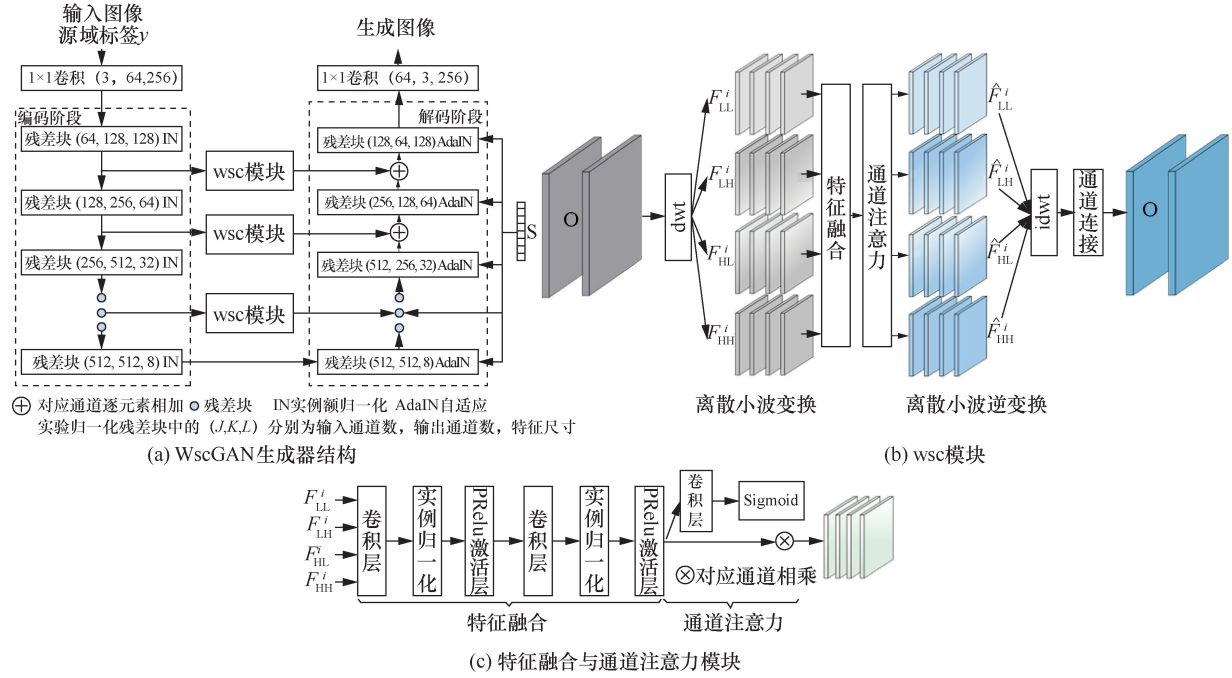


图3 生成器的整体结构
Figure 3 The overall structure of the generator

$x_t = G(x, s)$ 。网络整体结构以自编码器框架为主, 如图 3(a)所示。出于对提升图像翻译精度以及生成性能的考虑, 本文参考类 U-Net^[35]网络框架, 引入快捷连接进行模型设计。具体来说, 将编码过程中对应阶段的特征输入解码阶段, 以此降低自编码器在处理过程中对源域图像的内容损失, 防止生成图像丢失源域图像的身份信息。由于不同的转换方向需要修改的区域基本集中在头发和背景, 即高频信息与低频信息的处理, 因此, 本文模型在快捷连接处进行特征级别的小波分解, 得到 4 个不同的频带特征, 将其逐个处理之后再输入解码阶段。样式特征是生成器对源域图像进行风格化的依据, 在本框架中 s_m 或 s_e 都可以表达目标域的期望风格。将解码阶段的残差块中的归一化方法替换为自适应实例归一化^[36] (AdaIN, adaptive instance normalization) 进行样式迁移, 其中 AdaIN 影响每一层的特征统计量, 从而使得目标域的样式向量嵌入源域特征中。AdaIN 的具体表示如下。

$$\text{AdaIN}(F, s) = \sigma(s) \left(\frac{F - \mu(F)}{\sigma(F)} \right) + \mu(s) \quad (1)$$

其中, F 表示输入 wsc 模块的特征, $\sigma(\cdot)$ 和 $\mu(\cdot)$ 分

别表示标准差函数以及均值函数。将编码特征输入对等的解码阶段使得源域信息以及目标域的样式特征达到一种平衡态, 从而改进了传统模型中连续的样式注入。生成图像过于参考目标域的问题, 使得最终的生成图像保留了身份信息以及部分源域的细粒度特征。

2.1.2 判别器

判别器与传统模型类似, 均以真实图像与生成图像作为输入, 并鉴别其真伪。本文为了提高整体架构在 I2I 任务中的性能, 判别器具有多个输出的多任务网络, 在输入真假图像的同时需输入对应域的标签, 判别器的每个输出根据标签值进行二值分类, 以确定输入图像是属于对应域的真实图像还是生成图像。如图 2 所示, D_y 以及 D_{y_t} 表示判别器对图像预测为属于域 y 或 y_t 的概率。当 D_y 或 D_{y_t} 的值较大时代表判别器认为当前图片属于域 y 或 y_t , 反之则认为是生成的假图像。判别器的结构由与生成器编码阶段相同结构的下采样残差块级联组成, 在输出端添加卷积层保证输出的维度符合二分类要求。在判断输入图像的真实性的同时, 判别器还可以对相应的域进行判断, 使得在不增加额外参数的情况下保证输出的多样

性和高保真度。

2.1.3 样式提取器与映射器

样式提取器结构如图 4(a)所示, 其作用是根据设定的矢量维度提取样式, 并尽可能地与 G 相匹配, 使 G 能翻译出具有目标域信息的图像。更确切地说, E 根据输入的参考图像 x_r 和其所属的域, 提取生成器所需的样式特征 s_e , 用公式表示为 $s_e = E(x_r, y_l)$ 。样式提取器的结构与判别器类似, 同样使用生成器编码阶段的下采样残差块级联, 由于是多域转换模型, 故在结构末端添加多个平行的线性层确保样式提取器可以输出属于多个域的样式特征, 之后根据输入的目标域

标签 y_l 选择对应的样式特征 s_e 。

然而, 受数据集图像数量的限制, 由参考图得到的样式风格通常是有限的, 所以需要额外通过噪声映射来增加 s 的变化。与 E 不同, 映射器的输入为噪声向量 $z \sim N(0,1)$ y_l 。由于映射器的输入为向量, 故使用线性层以及激活函数堆叠将噪声向量升维至样式特征所需尺寸, 同时映射器是具有多个输出的网络, 与 E 类似, 根据标签 y_l 选择对应的样式特征 s_m , 结构如图 4(b)所示。输出的 s_m 具有与 s_e 相同的功能, 其中 $s_m = M(z, y_l)$ 。最终, 由 s_m 和 s_e 组成的样式特征 s 鼓励生成器在多个域之间翻译不同的图像。

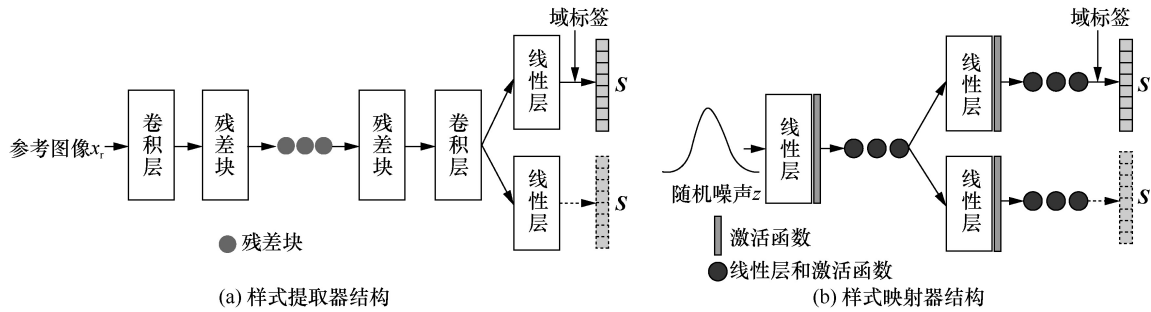


图4 样式提取器结构与样式映射器结构
Figure 4 Style feature extractor structure and style feature mapper structure

2.1.4 融合小波变换的快捷连接

图像中人物的性别、种族以及年龄等属性的差异会使得图像之间频率信息的占比不同, 如女性人物图像高频信息的比重较大, 男性图像则反之。因此, 在性别属性转换时, 需要针对频域信息进一步处理。当对二维图像进行小波变换时(如图 5 所示), 小波沿着不同的频带将图像分割成 4 个相同大小的块, Ψ 为滤波操作。其中低通滤波器所处理的信号集中在图像光滑的表面, 主要与低频信号有关, 而高通滤波器则捕捉大部分的高频信号, 如垂直、水平和对角线的边缘。与原始图像相比, 哈尔小波变换后的图像具有更好的信息保持性, 故在本模型中选择哈尔小波变换作为基函数。哈尔小波变换包含两个池化操作, 分别为小波池化和小波逆池化。哈尔小波池化的内核如式(2)所示。

$$L = \frac{1}{\sqrt{2}}[1, 1]^T, \quad H = \frac{1}{\sqrt{2}}[-1, 1]^T \quad (2)$$

其中, L 指低通, H 指高通。用 L 以及 H 构成了

4 个内核, 分别为 LL^T, LH^T, HL^T, HH^T 。滤波后的图像对应图 5 最后一列的 a、b、c、d。通过对小波变换后的结果可视化可以发现, 低频分量保存了原图像的内容, 其他的则含有详细的纹理。

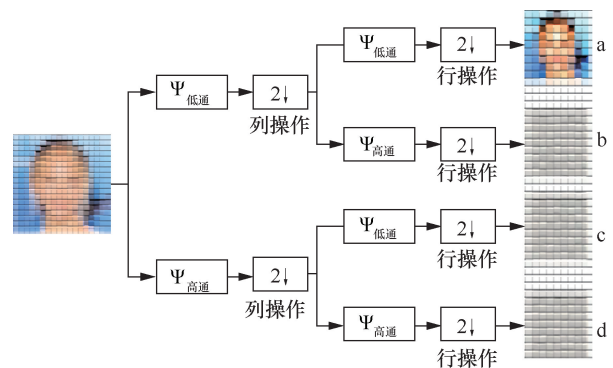


图5 小波分解结果可视化
Figure 5 Illustration of wavelet transformation

小波变换通常在图像层面上应用, 受到 WCT2^[37]的启发, 本文模型将小波变换应用于特征级别。wsc 模块如图 3(b)所示, 考虑要对解码阶段中不同尺寸的特征进行分解, 故本模块中小

波分解只递归使用一次。这些特征通过小波变换处理,获得不同层次的信息,这有助于根据模型翻译需求,保留原始图像的全局拓扑信息以及局部细粒度的纹理细节。将编码过程中每一阶段残差块的输出特征记为 F_{down}^i , 其中 $i \in \{0, 1, 2, \dots, 6\}$ 代表不同阶段的下采样,同理可得 F_{up}^i 。 F_{down}^i 经过小波池化得到 4 个不同的特征频带,分别为 F_{LL}^i 、 F_{LH}^i 、 F_{HL}^i 、 F_{HH}^i 。对不同频带的特征经过特征融合处理过后,再施加通道注意力,筛选出需要在解码阶段保留的特征信息。4 组特征进行上述处理后使用小波逆池化还原到处理前的尺寸,最终在通道维度上连接,传输至上采样阶段。整个过程用公式表示为

$$F_{\text{fre}_j}^i = \Phi_{\text{dwt}}(F_{\text{down}}^i) \quad (3)$$

$$\hat{F}_{\text{fre}_j}^i = \mathcal{Q}_{\text{att}}(\mathcal{I}_{\text{T}}(F_{\text{fre}_j}^i)) \quad (4)$$

$$F_{\text{wsc}}^i = \text{concat}(\phi_{\text{idwt}}(\hat{F}_{\text{fre}_j}^i)) \quad (5)$$

其中, $\Phi_{\text{dwt}}(\cdot)$ 、 $\phi_{\text{idwt}}(\cdot)$ 、 $\mathcal{I}_{\text{T}}(\cdot)$ 、 $\mathcal{Q}_{\text{att}}(\cdot)$ 、 $\text{concat}(\cdot)$ 分别代表离散小波变换、离散小波逆变换、特征融合、通道注意力以及通道连接。 $F_{\text{fre}_j}^i$ 以及 $\hat{F}_{\text{fre}_j}^i$ 代表处理前后的不同频带特征。 F_{wsc}^i 为 wsc 模块的输出,与对应阶段的 F_{up}^i 进行相加后输入下一阶段的解码过程。

2.2 损失函数

本节详细介绍损失函数,包括对抗损失、循环一致损失、样式特征重构损失和风格多样性损失。本文模型的整体损失函数可以总结为

$$\min_{G, M, E} \max_D (G, D, M, E) = L_{\text{adv}} + L_{\text{cyc}} + L_{\text{sr}} - L_{\text{sd}} \quad (6)$$

2.2.1 对抗损失

本文模型中对抗损失的目的是引导模型生成更加真实的图像以及使图像更符合目标域的风格。对抗损失选择了非饱和性的对抗损失^[1],其定义为

$$L_{\text{adv}} = \mathbb{E}_{x, y} [\log D_y(x)] + \mathbb{E}_{x, y_1, x_t, z} [\log(1 - D_{y_1}(x_t))] \quad (7)$$

其中, $x_t = G(x, s)$ 。 s 源于映射器 $s_m = M(z, y_1)$ 或提取器 $s_e = E(x_t, y_1)$ 。在训练过程中循环从 z 或 x_t 中采样得到样式特征。 $D_y(\cdot)$ 以及 $D_{y_1}(\cdot)$ 表示判别

器对图像预测为属于域 y 或 y_1 的概率。

2.2.2 循环一致损失

由于主流数据集中不包含 I2I 任务理想的成对数据,在这种条件下,为了保证 G 生成的图像具有源域图像的部分特征以及面部身份信息,有必要使用循环一致损失。本文采用了像素级的循环一致损失,具体如下。

$$L_{\text{cyc}} = \mathbb{E}_{x, y, s} [\|G(G(x, s), E(x, y)) - x\|_1] \quad (8)$$

其中, $\|\cdot\|_1$ 表示 1 范数, $G(x, s)$ 即源域图像 x 根据目标域相对应的样式特征 s 生成图像 x_t 。 $E(x, y)$ 是提取器 E 根据源域标签提取原图像的样式特征。 $\hat{x} = G(x_t, E(x, y))$ 表示将生成图像通过源域风格进行翻译的伪源域图像。通过最小化 L_{cyc} 可以在训练阶段使伪原始图像逐个像素地近似于输入图像,确保翻译后的图像保留了对应于原始图像 x 的大部分结构。通过对各个 x 进行循环一致损失监督,可以鼓励特征之间的一致性,并提高 G 、 E 和 M 的性能。

2.2.3 样式特征重构损失

样式特征是图像翻译过程中的关键。为了指导 G 可以生成更多样的图像,本文引入了噪声映射机制。由于映射器 M 的输入 $z \sim N(0, 1)$ 具有不确定性,故需要样式特征重构损失进行约束,使其包含的样式符合目标域风格,损失可以表示为

$$L_{\text{sr}} = \mathbb{E}_{x, z, y_1} [\|E(G(x, M(z, y_1)), y_1) - M(z, y_1)\|_1] \quad (9)$$

由式(9)可知,通过具有确定输入提取器 E 的监督,最小化 L_{sr} 消除了 z 对映射器 M 的负面影响,确保 z 映射到的特征是理想的目标域风格。同时最小化 L_{sr} 也可迫使 G 应用样式特征进行图像翻译。

2.2.4 风格多样性损失

为了保证生成器可以产生多样的图像,避免模式坍塌,本文借鉴 MSGAN^[38]、StarGAN-v2^[39] 等模型,使用风格多样性损失,具体如下。

$$L_{\text{sd}} = \mathbb{E}_{x, y_1, z} [\|G(x, s_{m1}) - G(x, s_{m2})\|_1] \quad (10)$$

其中, s_{m1} 和 s_{m2} 分别采样于不同的隐向量 $z \in N(0, 1)$ 。最大化 L_{sd} 值意味着根据不同的 z 所映射出的样式向量生成的图像差别越大,越具有样式多样性。因此使用风格多样性损失会使生成

器产生更多样化的图像。

3 实验结果与分析

3.1 实验设置

3.1.1 数据集

WscGAN 在 CelebA-HQ^[15] 以及 FFHQ^[16] 两个具有高分辨率人脸属性的数据集上进行训练和测试。其中 CelebA-HQ 包含 30 000 幅面部图像, 分辨率为 1 024×1 024。为了验证模型在性别属性操作任务中具有鲁棒性, 本文额外在 FFHQ 数据集进行完整训练以及验证, 其中 FFHQ 数据集包含 1 024×1 024 分辨率的 70 000 幅 PNG 格式高清人脸图像, 在年龄、种族和图像背景等方面风格多样且差异明显。

3.1.2 实验环境及参数设置

本文实验环境为 Intel Xeon Gold 5218 CPU 2.30 GHz 以及单个 NVIDIA Tesla V100S GPU 32 GB 的 PC 上进行训练。训练时采用 Adam 优化器, 其中 $\beta_1 = 0, \beta_2 = 0.99$ 。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。在每个训练步骤中, 判别器和生成器交替训练。生成器、判别器、样式提取器以及映射器的学习率均为 10^{-5} 。由于 GAN 的实验过程不稳定, 故本文进行了 3 次完整训练, 对每次的训练结果设置 10 个不同的随机种子进行测试, 得到的结果是所有测试结果的平均值。

3.1.3 评价指标

为了定量评估所提模型的性能, 本文采用两个指标: 弗雷歇起始距离^[40] (FID, Fréchet inception distance) 以及特征度量图像相似度^[41] (LPIPS, learned perceptual image patch similarity)。其中, FID 用于计算生成图像和真实图像之间的分布距离, 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 值更大。在本文设置中, 通过从标准正态分布采样 10 个不同噪声向量 z 映射为样式特征, 进而生成样本用于计算 LPIPS 值。为了实验的公平性, 本文使用 StarGAN-v2^[39] 官方代码中 FID 与 LPIPS 的计算方法。本次实验对比的模型有 MUNIT^[26]、DRIT++^[27]、

StarGAN-v2^[39]、PHStarGAN-v2^[29]。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。由于样式特征有两种获取渠道, 因此本文从由 s_m 参与生成的图像 (称之为潜在引导生成) 和 s_e 参与生成的图像 (称之为参考引导生成) 两方面来评估 WscGAN 的性能表现。

3.2 实验结果

3.2.1 潜在引导生成比较

为了评估 WscGAN 模型在基于潜在引导的翻译上的优秀性能, 与目前先进的 I2I 模型, 即 DRIT++、MUNIT、PHStarGAN-v2 ($n=3$) 以及 StarGAN-v2 在相同条件下进行比较。生成图中, DRIT++ 生成的图像并不符合目标域的语义要求, 且多样性较差, 面部身份信息丢失严重。MUNIT 虽然翻译出了符合目标域特征的图像, 但依旧保留了大量的源域信息, 以至于出现伪影, 且头发等部位的细节部分没有得到更好的处理。相比 MUNIT, StarGAN-v2 潜在引导生成的图像多样性较好, 其结果具有目标域的特征, 但在面部、头发等区域会出现不规则的形状, 背景较为凌乱。PHStarGAN-v2 为了节省模型参数量, 只能生成其他模型一半分辨率的图像, 故图像整体质量较差, 且背景及图像细节处理较为粗糙。WscGAN 展现了较为优秀的翻译性能, 可以很好地翻译出目标域的特征, 并且在细节表现方面更加细腻。添加快捷连接使得面部身份信息丢失较少, 并且有特征级小波分解以及注意力机制的加持, 生成的图像背景不会因为源域特征的干扰显得更加整洁清晰, 在观感上更加自然。为了突出 WscGAN 的性能, 选取面部有大范围遮挡的图像进行实验。与在 CelebA-HQ 数据集中类似, DRIT++ 生成图像依旧不符合目标域的语义要求, 且生成图像几乎不具备多样性。从 MUNIT 在男性域向女性域的转换方向中可以看出生成图像基本符合目标域的风格, 但存在大量伪影。StarGAN-v2 在各个转换方向的生成图像均符合目标域风格, 但生成图像背景杂乱。PHStarGAN-v2 受分辨率限制, 生成结果质量较差, 且由于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, 同时伪影较多。

WscGAN 在 FFHQ 数据集中同样展现了较优的性能,即使图像存在大范围遮挡,同样可以在保持身份信息的基础上翻译符合目标域语义要求的图像,且伪影较少,生成图像的多样性较好,背景相比其他模型更为整洁。

在客观指标方面,表 1 给出了基于潜在引导的生成方式下,本文模型在 CelebA-HQ 和 FFHQ 数据集中的 FID 和 LPIPS 指标。其中 FID 越小越好, LPIPS 则相反。同样与 DRIT++、MUNIT、StarGAN-v2 以及 PHStarGAN-v2 模型进行比较,可以看到在不同的数据集中,本文模型的 FID 指标均为最小,说明与先进模型相比,本文模型在生成图像的真实度方面是最佳的,并且 LPIPS 值较高,表明生成图像的多样性较好。最终结果表明,在融合了小波快捷连接的翻译过程中,不仅获得了质量较好的生成图像,并且能提升图像翻译过程中的多样性。

表 1 对不同模型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

模型	CelebA-HQ		FFHQ	
	FID	LPIPS	FID	LPIPS
MUNIT ^[26]	24.96	0.252	28.03	0.231
DRIT++ ^[27]	52.13	0.178	44.12	0.061
StarGAN-v2 ^[39]	13.71	0.452	19.21	0.455
PHStarGAN-v2 ^[29]	16.63	0.332	43.03	0.342
本文模型	13.26	0.462	14.95	0.480

3.2.2 参考引导生成比较

在基于参考引导生成图像的实验中,无论是在 CelebA-HQ 还是 FFHQ 数据集中,DRIT++、MUNIT、PHStarGAN-v2 等生成图明显丢失了源域图的面部细节,其中,在 FFHQ 数据集中 PHStarGAN-v2 由于过于参考目标域图,翻译图像的面部方向发生了改变。DRIT++以及 MUNIT 几乎没有根据参考图像进行翻译,且后者生成的图像涂抹严重。PHStarGAN-v2 翻译出的图像符合目标域的语义要求,但受限于分辨率,图像整体观感较差,且生成图像的头发不自然。本文模型所生成的图像保留了源域面部身份信息的同时很好地根据参考图像的样式进行了翻译。不仅如此,整体图像的光照、背景以及头发的自然程度

均更胜一筹。在源域图像有大范围遮挡的情况下,相比其他先进模型,WscGAN 参考了目标域的样式翻译出理想的图像。表 2 展示了不同模型在 CelebA-HQ 和 FFHQ 数据集中基于参考引导生成图像的定量比较,本文在各个指标下均取得了最佳的效果。

表 2 对不同模型基于参考引导生成的图像的定量比较
Table 2 Quantitative comparison of images generated by different methods based on reference guidance

模型	CelebA-HQ		FFHQ	
	FID	LPIPS	FID	LPIPS
MUNIT ^[26]	64.54	0.123	51.29	0.126
DRIT++ ^[27]	53.3	0.311	40.08	0.032
StarGAN-v2 ^[39]	23.84	0.388	28.94	0.409
PHStarGAN-v2 ^[29]	28.11	0.290	56.82	0.332
本文模型	22.06	0.393	23.90	0.420

3.2.3 消融实验

为了证明所提模型的有效性,本文在 CelebA-HQ 数据集中对生成器结构进行消融对比来说明各个模块对人脸属性编辑的影响。如表 3 所示,自编码器结构所体现的是 I2I 任务中生成器常用架构的性能。自编码器+快捷连接是对自编码器结构添加快捷连接但并不包含 wsc 模块。自编码器结构在编码阶段会丢失大量的源域特征信息,这导致生成图像的身份信息发生改变。在只添加快捷连接的实验中,每一个阶段的解码特征都可直接传输至对应的编码阶段,但由于没有 wsc 模块对快捷连接中的特征进行处理,翻译后的图像没有更好地符合目标域的语义要求。在表 3 中给出了各个消融情况下的 FID 与 LPIPS 值,客观的数据与生成结果都证明了融合小波分解的快捷连接表现出较好的性能。

表 3 消融实验结果
Table 3 Results of ablation experiments

不同方法	潜在引导生成		参考引导生成	
	FID	LPIPS	FID	LPIPS
自编码器	22.64	0.092	35.81	0.238
自编码器+快捷连接	19.98	0.125	19.63	0.124
自编码器+快捷连接 +wsc 模块	13.26	0.462	22.06	0.393

4 结束语

在面部性别属性伪造这一 I2I 任务中, 传统模型生成的图像往往存在丢失源域图像身份信息以及存在伪影等缺陷。针对上述问题, 本文提出了新的深度学习网络, 即融合小波快捷连接生成对抗网络。本文在自编码器中参考类 U-Net 网络向其添加快捷连接, 并嵌入了 wsc 模块, 其中 wsc 模块利用小波分解以及注意力机制对各级编码结果进行特征级的逐个处理, 然后输入相对应的解码阶段, 最终高质量实现了人脸性别属性转换任务。为了验证本文模型的有效性, 分别在 CelebA-HQ 和 FFHQ 数据集上进行了实验, 并利用 FID、LPIPS 指标进行客观量化评估。实验结果显示, 在两种数据集中 FID 指标分别提升 5.4% 和 19.8%, LPIPS 指标分别提升 1.8% 和 4.1%。此外, 在融合了小波快捷连接的生成器结构下, 本文模型相比其他先进模型具有更好的效果, 在提高图像翻译精度的同时, 提升了生成的多样性。

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. 2014: 2672-2680.
- [2] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv Preprint arXiv: 1411.1784, 2014.
- [3] TANG H, XU D, YAN Y, et al. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020: 7867-7876.
- [4] CHEN B J, TAN W J, WANG Y T, et al. Distinguishing between natural and GAN-generated face images by combining global and local features[J]. Chinese Journal of Electronics, 2022, 31(1): 59-67.
- [5] GUO X, WANG Z, YANG Q, et al. Gan-based virtual-to-real image translation for urban scene semantic segmentation[J]. Neurocomputing, 2020, 394: 127-135.
- [6] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks[J]. arXiv Preprint arXiv: 1703.05192, 2017.
- [7] 王旭东, 卫红权, 高超, 等. 身份保持约束下的人脸图像补全[J]. 网络与信息安全学报, 2018, 4(8): 71-76.
WANG X D, WEI H Q, GAO C, et al. Identity preserving face completion with generative adversarial networks[J]. Chinese Journal of Network and Information Security, 2018, 4(8): 71-76.
- [8] TUNG H Y F, HARLEY A W, SETO W, et al. Adversarial inverse graphics networks: learning 2D-to-3D lifting and image-to-image translation from unpaired supervision[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 4364-4372.
- [9] LI X Y, ZHANG S C, HU J, et al. Image-to-image translation via hierarchical style disentanglement[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 8635-8644.
- [10] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 2242-2251.
- [11] TOMEI M, CORNIA M, BARALDI L, et al. Art2Real: unfolding the reality of artworks via semantically-aware image-to-image translation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 5842-5852.
- [12] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5967-5976.
- [13] KINGMA D P, WELING M. Auto-encoding variational bayes[J]. arXiv Preprint arXiv: 1312.6114, 2013.
- [14] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[J]. arXiv Preprint arXiv: 1511.05644, 2015.
- [15] KARRAS T, AILA, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[J]. arXiv preprint arXiv: 1710.10196, 2017.
- [16] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 4396-4405.
- [17] DENTON E L, CHINTALA S, FERGUS R. Deep generative image models using a laplacian pyramid of adversarial networks[J]. arXiv preprint arXiv: 1506.05751, 2015.
- [18] MAO X D, LI Q, XIE H R, et al. Least Squares generative adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 2813-2821.
- [19] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. arXiv Preprint arXiv: 1701.07875, 2017.
- [20] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[J]. arXiv Preprint arXiv: 1802.05957, 2018.
- [21] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint arXiv: 1511.06434, 2015.
- [22] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets[C]//Proceedings of Neural Information Processing Systems. 2016: 2172-2180.
- [23] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[J]. arXiv Preprint arXiv: 1809.11096, 2018.
- [24] TANG H, XU D, SEBE N, et al. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 2412-2421.
- [25] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition

- (CVPR). 2020: 2332-2341.
- [26] HUANG X, LIU M Y, BELONGIE S, et al. Multimodal unsupervised image-to-image translation[C]//Proceedings of European Conference on Computer Vision. 2018: 179-196.
- [27] LEE H Y, TSENG H Y, MAO Q, et al. DRIT++: diverse image-to-image translation via disentangled representations[J]. arXiv Preprint arXiv: 1905.01270, 2019.
- [28] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8789-8797.
- [29] GRASSUCCI E, SIGILLO L, UNCINI A, et al. Hyper complex image-to-image translation[J]. arXiv Preprint arXiv: 2205.02087, 2022.
- [30] DEMIREL H, ANBARJAFARI G. Discrete wavelet transform-based satellite image resolution enhancement[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(6): 1997-2004.
- [31] CHEN T S, LIN L, ZUO W M, et al. Learning a wavelet-like auto-encoder to accelerate deep neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).
- [32] HUANG H B, HE R, SUN Z N, et al. Wavelet-SRNet: a wavelet-based CNN for multi-scale face super resolution[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 1698-1706.
- [33] FUJIEDA S, TAKAYAMA K, HACHISUKA T. Wavelet convolutional neural networks for texture classification[J]. arXiv Preprint arXiv: 1707.07394, 2017.
- [34] WILLIAMS T, LI R. Wavelet pooling for convolutional neural networks[C]//Proceedings of International Conference on Learning Representations. 2018.
- [35] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015: 234-241.
- [36] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 1510-1519.
- [37] YOO J, UH Y, CHUN S, et al. Photorealistic style transfer via wavelet transforms[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2020: 9035-9044.
- [38] MAO Q, LEE H Y, TSENG H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 1429-1437.
- [39] CHOI Y, UH Y, YOO J, et al. StarGAN v2: diverse image synthesis for multiple domains[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 8185-8194.
- [40] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of Neural Information Processing Systems (NIPS). 2017: 6626-6637.
- [41] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 586-595.

[作者简介]



陈万泽（1995- ），男，甘肃兰州人，福建师范大学硕士生，主要研究方向为深度学习、生成对抗网络、图像生成。



黄丽清（1991- ），女，福建莆田人，福建师范大学讲师，主要研究方向为视频图像超分辨率、多媒体内容安全。



陈家桢（1971- ），女，福建福州人，福建师范大学副教授，主要研究方向为信息安全。



叶锋（1978- ），男，福建福州人，福建师范大学教授，主要研究方向为视频通信、视频图像处理。



黄添强（1971- ），男，福建仙游人，福建师范大学教授、博士生导师，主要研究方向为人工智能、大数据分析、数字媒体内容安全理论和技术。



罗海峰（1990- ），男，福建龙岩人，福建师范大学讲师，主要研究方向为二维/三维计算机视觉。