

中图法分类号: TP391.41 文献标识码: A 文章编号: 1006-8961(2023)12-3784-13

论文引用格式: Liao Y H, Qian W H and Cao J D. 2023. MStarGAN: a face style transfer network with changeable style intensity. Journal of Image and Graphics, 28(12):3784-3796(廖远鸿, 钱文华, 曹进德. 2023. 风格强度可变的人脸风格迁移网络. 中国图象图形学报, 28(12):3784-3796)[DOI:10.11834/jig.221149]

风格强度可变的人脸风格迁移网络

廖远鸿¹, 钱文华^{1*}, 曹进德²

1. 云南大学信息学院, 昆明 650504; 2. 东南大学数学学院, 南京 214135

摘要: 目的 针对人脸风格迁移算法 StarGAN(star generative adversarial network)、MSGAN(mode seeking generative adversarial network)等存在细节风格学习不佳、迁移效果单一和生成图像失真等缺点, 提出一种能够降低失真并生成不同风格强度图像的人脸风格迁移算法 MStarGAN(multilayer StarGAN)。方法 首先, 通过特征金字塔网络(feature pyramid network, FPN)构建前置编码器, 生成蕴含图像细节特征的多层特征向量, 增强生成图像在风格传输时能学习到的风格图像的细节风格; 其次, 使用前置编码器对原图像及风格图像各生成一个风格向量并进行组合, 利用组合后的风格向量进行风格传输, 使生成图像具有不同的风格迁移强度; 最后, 采用权重解调算法作为生成器中的风格传输模块, 通过对卷积权重的操作代替在特征图上的归一化操作, 消除特征图中的特征伪影, 减少生成图像中的失真。结果 在 Celeba_HQ 数据集上进行实验, 与 MSGAN、StarGAN v2 等对比算法相比, 在参考引导合成实验中, MStarGAN 的 FID(Frechet inception distance score)指标分别降低了 18.9 和 3.1, LPIPS(learned perceptual image patch similarity)指标分别提升了 0.094 和 0.018。在潜在引导合成实验中, MStarGAN 的 FID 指标分别降低了 20.2 和 0.8, LPIPS 指标分别提升了 0.155 和 0.92, 并能够生成具有不同风格强度的结果图像。结论 提出的算法能够传输图像的细节风格, 生成具有不同强度的输出图像, 并减少生成图像的失真。

关键词: 人脸风格迁移网络; StarGAN; 风格强度; 特征金字塔网络(FPN); 权重解调

MStarGAN: a face style transfer network with changeable style intensity

Liao Yuanhong¹, Qian Wenhua^{1*}, Cao Jinde²

1. School of Information Science and Engineering, Yunnan University, Kunming 650504, China;

2. School of Mathematics, Southeast University, Nanjing 214135, China

Abstract: Objective The style transfer algorithm can transfer the style from the art image to the original natural image. The style image provides certain features, such as style texture and stroke, while the content image provides the contour structure. The goal of the style transfer algorithm is to synthesize a new stylized image with the texture stroke of the style image and the contour structure of the content image. The early face style transfer algorithm applies mathematical modeling to build a filter that counts the local features of the target image to understand its style. This algorithm then establishes a statistical model to describe the image style. However, the face style transfer algorithm only generates a single style, the resulting image style is not obvious, and needs to be modeled manually, thereby limiting its efficiency. With the rise of

收稿日期: 2022-12-22; 修回日期: 2023-04-03; 预印本日期: 2023-04-10

* 通信作者: 钱文华 whqian@ynu.edu.cn

基金项目: 国家自然科学基金项目(62162065); 云南省科技厅基金项目(202105AF150011, 202001BB050043); 云南省重大科技专项计划项目(202002AD080001); 云南省科技厅应用基础研究计划重点项目(2019FA044); 云南省中青年学术技术带头人后备人才项目(2019HB121)

Supported by: National Natural Science Foundation of China (62162065); Research Foundation of Yunnan Province (202105AF150011, 202001BB050043)

deep learning, the style transfer algorithm has started using the deep learning model as its core. Given that generative adversarial network (GAN) can generate images that satisfy certain distribution laws, we can generate a target image that is similar to the real image by training GAN. Therefore, GAN has been widely used in image style transfer algorithms. The main image style transfer algorithms are divided into two categories. The algorithms in the first category only improve GAN without using a pre-encoder, such as pix2pix and CycleGAN, while those in the second category use a pre-encoder. Due to the addition of encoders before the GAN structure, the resulting network structure becomes complex yet achieves highly realistic results, such as StyleGAN and StarGAN. To overcome the shortcomings of some face style transfer algorithms, such as StarGAN and MSGAN, which have poor detail style learning, insignificant style transfer effect, and generation of distorted images, we present a face style migration algorithm called multi-layer StarGAN (MStarGAN) with controllable style intensity. **Method** First, we construct the pre-encoder through the feature pyramid network (FPN) to generate multi-layer feature vectors containing image detail features. Compared with the original 1×64 feature vector, the pre-encoder constructed by FPN can output a 6×256 feature vector, which contains additional details of the original image. Therefore, the generated image can learn the detailed style of the style image during style transmission. Second, we use the pre-encoder to generate style vectors for the original and style images and then combine these vectors. We then use the combined style vector for style transmission. We can also adjust the number of layers of this vector so that the style of the generated image is biased to either the original or style image, hence resulting in different style transfer intensities for the generated image. Third, we introduce a new loss function to maintain balance in the style of the generated image and ensure that the image will not be too biased toward either the original or style image. Fourth, we apply the weight demodulation algorithm as our style transmission module in the generator. The traditional method AdaIN has been proven to distort the generated image. By replacing the normalization operation on the feature map with the operation of convolution weight, we eliminate the feature artifacts in the feature map and reduce the distortion in the generated image. **Result** We implement our model in Python and test it on the Celeba_HQ dataset with RTX2080Ti. Our model not only generates high-quality random face images but also makes the generated images learn the style of style images, including hair and skin color. Compared with the multimodal unsupervised image-to-image translation), diverse image-to-image translation, MSGAN, and StarGAN V2 algorithms, in the latent-guided synthesis experiment, the Frechét inception distance (FID) index of the proposed algorithm is reduced by 18.5, 39.2, 20.2, and 0.8, respectively, while its learned perceptual image patch similarity (LPIPS) index is increased by 0.181, 0.366, 0.155, and 0.092 respectively. In the reference-guided synthesis experiment, the FID index of the proposed algorithm is reduced by 86.4, 32.6, 18.9, and 3.1, respectively, while its LPIPS index is increased by 0.23, 0.095, 0.094, and 0.018, respectively. In sum, our algorithm can generate result images with different styles and intensities. **Conclusion** The proposed algorithm can transmit the detail style of the image, control the style intensity of the output image, and reduce the distortion of the generated image.

Key words: face style transfer network; StarGAN; style intensity; feature pyramid network (FPN); weight demodulation

0 引言

风格迁移技术可将艺术图像中的风格传输到原始的自然图像中,其中提供风格纹理和笔触等特征的图像称为风格图,而提供轮廓结构的图像称为内容图,风格迁移算法的目标是以风格图的纹理笔触与内容图的轮廓结构合成一幅新的风格化图像。图像风格迁移算法在航空航天,图像识别等领域中的应用广泛,如航拍测绘(Lin等,2017)、图像分类(Huang和Belongie,2017)等。人脸风格迁移算法作为风格迁移算法中一个分支,其内容图与风格图均

为人类面部肖像,目标为将风格图的面部风格特征与内容图的面部内容合成为一幅新的面部图像,如图1所示。如今人脸风格迁移算法已广泛应用于人脸识别(Nirkin等,2018)和照片美化(Li等,2018)等方面。

早期的人脸风格迁移算法(Shih等,2014)使用了数学建模构建滤波器的方法,对所要得到其风格的目标图像的局部特征进行统计,并建立统计模型(Portilla和Simoncelli,2000)描述其图像风格。这种算法只能针对一种单一风格进行生成,生成的结果图像风格不明显,且统计模型必须手动建模,因此算法效率较低,图2显示了此算法的风格迁移结果。



图1 人脸风格迁移算法示例

Fig. 1 Face style migration algorithm example
((a) original image; (b) style image; (c) result image)

近年来,基于深度神经网络的算法已应用于人脸风格迁移工作中,Gatys等人(2015a,b)首先提出基于深度神经网络的图像风格迁移算法,采用VGG-19 (Visual Geometry Group)网络(Simonyan和Zisserman, 2015)代替了传统风格迁移算法中的统计特征模型,算法效率得到了较大提升。但由于VGG-19网络所生成的图像分布较为均匀,此时的风格迁移算法与纹理合成算法较为类似,风格传输效果不明显。

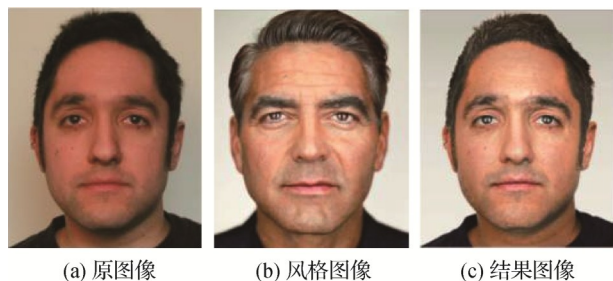


图2 早期人脸风格迁移算法示例(Shih等,2014)

Fig. 2 Early face style migration algorithm example
((a) original image; (b) style image; (c) result image)

为了解决上述问题,将生成式对抗网络(generative adversarial network, GAN)(Goodfellow等, 2020)应用于图像风格迁移算法中。由于GAN能够生成符合某一分布规律的图像,通过对GAN网络进行训练即能生成与真实图像较为相似的目标图像。Isola等人(2017)提出了pix2pix算法,使用条件生成对抗网络(Mirza和Osindero, 2014)学习从输入图像到输出图像的映射,以完成风格迁移任务,能够生成风格显著、质量较高的结果图像。Xie等人(2020)提出了一种基于相关对齐的总变分图像风格迁移模型。Zhu等人(2017)采用了双判别器的结构,训练集可以指向两个不成对的图像集合,解决了风格迁移网络前期划分图像空间较复杂的问题。

Choi等人(2018)提出的StarGAN(star generative adversarial network)采用前置编码器以实现风格迁移。此编码器可以实现图像到向量的转变,这个向量包含原本图像中的信息,再通过损失函数将向量规范至网络的潜空间中,以人为地修改网络中的潜变量。StarGAN能够在保留原图像特征的同时学习到目标图像的风格,由于采用了前置编码器,风格迁移的多样性及强度都有所提升,但其自身存在着以下问题:1)StarGAN生成的图像可能产生特征伪影,这种伪影有类似掩膜的性质,会在生成图像上呈现出多余的特征,造成生成图像失真;2)StarGAN对于目标图像细节部分的特征学习有所欠缺,其原因在于此网络使用的风格向量较小,无法很好地归纳目标图像中的信息;3)StarGAN对于每一对图像输入只有单一的风格迁移输出,不能较好地控制生成图像的风格强度。

为了解决以上问题,本文提出了MStarGAN(multi-layer StarGAN),采用权重解调(Karras等, 2020)方法解决StarGAN网络中出现的特征伪影问题。为了解决传统编码器结构臃肿、参数量大、效率较低等缺点,采用Lin等人(2017)提出的特征金字塔网络(feature pyramid network, FPN)模型作为前置编码器的基础结构,扩展了生成图像的风格多样性,并能得到不同风格强度的生成图像。本文的主要工作包括:1)提出一个能生成不同风格强度图像的人脸图像风格迁移网络,通过调整网络的参数,引入不同强度的目标图像特征,能够生成最多4幅风格迁移强度不同的结果图像;2)采用权重解调方法重新构造生成器,减小了风格传输中的失真,减少了生成图像中的特征伪影;3)采用风格强度损失,保持了内容图特征与风格图特征在风格迁移过程中的平衡,避免生成图像风格偏向原图或风格图。

1 相关工作

1.1 GAN

GAN网络可以生成满足特定分布规律的输出图像。Mirza和Osindero(2014)首先将标签信息编码为一个向量作为输入,约束隐变量与输出之间的关系,以控制GAN的输出。Denton等人(2015)将生成器由原本的多层卷积结构改为金字塔形结构,使用残差逼近,获得分辨率更高、质量更好的输出图像。

Shrivastava 等人(2017)首次将图像作为生成器的输入,较好地学习到输入和输出两个不同数据分布间的映射。之后,基于 GAN 的风格迁移算法(Elgammal 等, 2017; Sbair 等, 2018; Huang 等, 2017)得以提出。但由于 GAN 本身较难对隐变量进行修改,输出结果具有较大的局限性,仅通过 GAN 实现风格迁移算法无法很好地运用于实际。

1.2 编码器

目前主流的风格迁移算法分为两类:1)不使用前置编码器,仅对 GAN 进行改进的网络,如 pix2pix、CycleGAN(Zhu 等, 2017)以及 P²GAN(Zheng 和 Liu, 2020)等;2)使用前置编码器的网络,由于在 GAN 结构前增加了编码器,这种网络结构较为复杂,但能够得到更真实的结果,如 StyleGAN(style-based generative adversarial network)、StarGAN 等。StyleGAN 采用 8 层的多层感知机(multilayer perceptron, MLP)作为基础编码器,通过将随机变量输入进编码器中,能够生成符合生成器分布的风格向量,将此风格向量输入到生成器中,即能达到控制输出图像风格的目的;DualStyleGAN(dual style-based generative adversarial network)(Yang 等, 2022)在此网络的基础上将编码器扩展为两路 18 层 MLP,能够较为精细地控制图像细粒度风格;StarGAN 采用 ResNet(residual network)(He 等, 2016a, b)作为编码器,能够针对不同的图像风格域生成不同的风格向量。但这些网络采用的前置编码器效率较低,无法快速地将随机变量规范至网络的变量空间中。

1.3 风格学习模块

由于 GAN 本身不具有学习风格向量中特征的能力,需要额外的模块进行风格学习。目前主流的方法是在生成器中采用 AdaIN(adaptive instance normalization)(Huang 等, 2017)算法。AdaIN 通过输入原图像与风格图像的均值与方差,对输入图像进行归一化,使处理后的图像具有风格图像的部分特征。但 AdaIN 对输入图像的每个通道均进行了归一化,导致层与层之间相关联的信息受到破坏,输出图像原本信息量较高的部分进一步增强,导致结果图像失真。

本节提到的编码器、GAN 网络以及风格学习模块为目前人脸风格迁移网络的主要组成部分。网络存在以下问题:1)编码器结构较简单,提取的风格向量无法完整表达风格图像的特征,通过增加编码

器网络层数的方式能得到较好的风格向量,但参数量巨大,耗费时间久;2)采用 AdaIN 的方法会在特征图中产生特征伪影,使生成图像失真;3)无法控制生成图像的风格强度,传输多少风格到生成图像中是随机的。为了解决以上问题,本文提出了 MStarGAN。

2 本文方法

2.1 网络整体结构

MStarGAN 的整体结构如图 3 所示,整个网络包含 4 个子网络:Style Encoder、Mapping Network、生成器和判别器。其中,Style Encoder 提取输入图像中的特征信息,将其转化为风格向量 w ; Mapping Network 可以将随机变量转变为风格向量 w ;生成器和判别器即为 GAN 网络中的生成网络与判别网络。与 StarGAN v2 相比,MStarGAN 采用 FPN 网络作为编码器的主要结构,替代了原本的 ResNet 编码器,并需要原图像与风格图像作为编码器的输入,以提高生成的风格向量的精度及风格多样性。此外,本文还对 StarGAN v2 的生成器部分进行了改进,使生成图像不易失真。

采用编码器的风格迁移算法通常采用两种方式获取风格向量 w ,一种方式为将风格图像输入编码器,从中提取出符合风格图像分布的风格向量,使用此向量进行风格迁移的过程即为参考引导合成;另一种方式为将随机变量 z 输入编码器,由于 z 是随机生成的,导致生成的风格向量 w 表达的特征也为随机的,使用此向量进行图像风格迁移的过程即为潜在引导合成。

参考引导合成与潜在引导合成具有不同的功能。参考引导合成通过输入原图像及风格图像,能够生成具有风格图像风格的原图像;潜在引导合成不需要输入风格图像,通过输入原图像生成具有随机风格的图像。本文采用参考引导合成、潜在引导合成的风格迁移方法如图 4 所示。

设输入原图像为 x_o ,风格图像为 x_s ,风格向量为 w ,输出图像为 x_i ,随机变量为 z ,则两种不同的风格迁移方法如下:

1) 参考引导合成的风格迁移方法为

$$w = H_{SE}(x_o, x_s) \quad (1)$$

$$x_i = G(x_o, w) \quad (2)$$

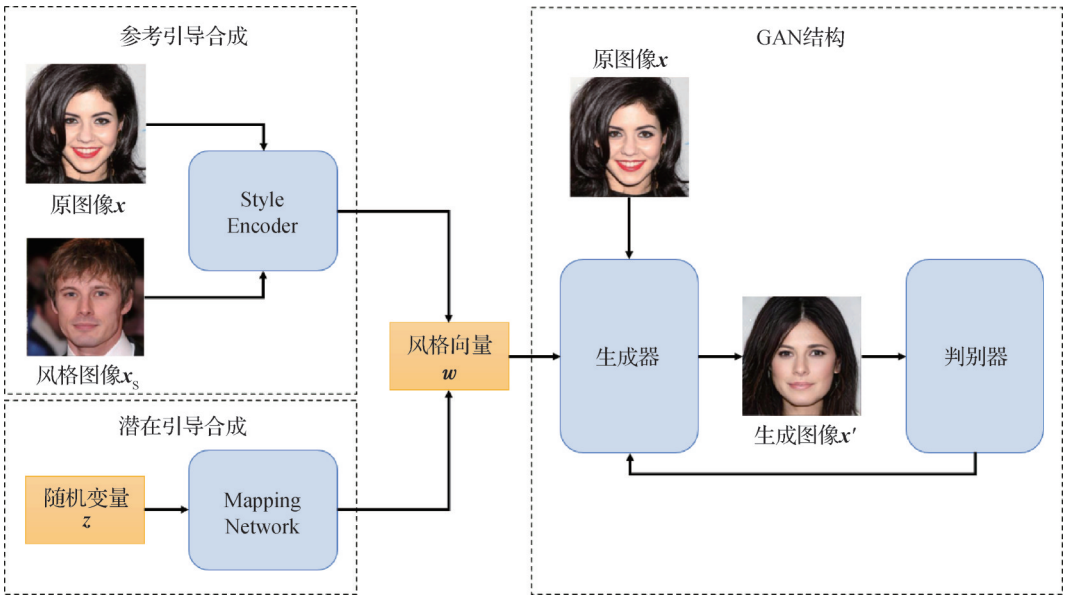


图3 网络整体结构

Fig. 3 Overall network structure



图4 两种风格迁移方法

Fig. 4 Two style transfer methods ((a) reference-guided synthesis; (b) latent-guided synthesis)

2) 潜在引导合成的风格迁移方法为

$$w = H_{MN}(z) \tag{3}$$

$$x_t = G(x_o, w) \tag{4}$$

式(1)一式(4)中, H_{SE} 为 Style Encoder 模块, H_{MN} 为 Mapping Network 模块, G 为生成器。

2.2 子网络结构

2.2.1 编码器

在 GAN 网络结构前添加的编码器模块 Style Encoder 用于提取风格输入图像 x_s 中的图像特征, 并将此特征编码为风格向量 w , 输入生成器模块中实现参考引导合成。本文的 Style Encoder 结构如图 5 所示。

FPN 最初应用于多尺度目标检测, 其目的是通过网络重建部分产生的多个高层特征去提取目标图像中不同尺度的信息。FPN 网络结构包括一个下采样层和一个上采样层, 对应层之间存在横向连接。FPN 通过简单的网络连接改变, 在基本不增加原有

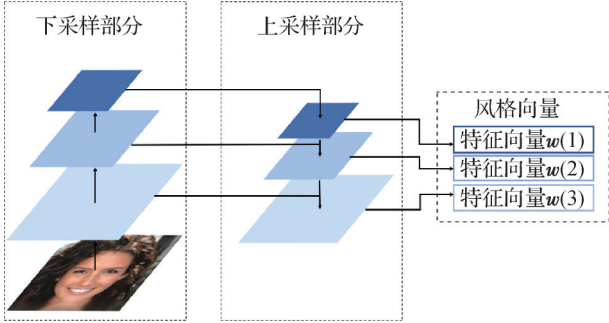


图5 Style Encoder网络结构

Fig. 5 Style Encoder network structure

模型计算量的情况下,大幅度提升了小物体检测的性能。通过高层特征进行上采样和低层特征进行自顶向下的连接,而且每一层都会输出结果,重建上采样层中高层的特征图更容易提取图像的细节部分特征,而底层更容易提取出图像的全局轮廓特征。

基于 FPN 能够生成多层信息不同的特征图的特点,本文采用 FPN 网络构造 Style Encoder,通过提取

FPN网络中重建部分的特征图,生成对应不同图像细粒度的图像风格向量,并根据所得到的图像风格向量进行风格迁移操作。

FPN网络的层数与所想得到的风格向量维度成正相关,FPN网络层数越多,风格向量的维度也就越高,表达的信息就越丰富,但生成器生成图像的性能会变低,由于层数增加,相邻层之间信息差异缩小,导致生成的不同风格强度的图像间差异缩小。因此,本文采用6层FPN网络构造Style Encoder,每层上采样层中能够输出 1×256 大小的向量,共能得到 6×256 大小的风格向量。后续参考引导合成实验表明,生成图像在包含了更多图像风格信息的同时,不同风格强度图像间仍存在较大差异。由于采用了FPN网络提取图像风格向量,因此这些向量也具有特征分层的特性,高层的向量包含图像的细节部分特征,而低层的向量包含图像的轮廓部分特征。由此,通过结合不同图像的风格向量并输入到生成器

中,生成具有丰富多样性的风格迁移图像。

为了对输出图像的风格强度进行控制,本文的编码器输入由原来的风格图像变为同时输入原图像及风格图像,因此,通过编码器可以得到两个风格向量,分别为原图像的风格向量 w_o 及风格图像的风格向量 w_s 。本文将 w_o 中的底层部分和 w_s 中的高层部分组合,得到具有原图像轮廓特征和目标图像细节特征的新的风格向量 w 。同时,还可以通过调整目标图像的层数分布与数量,以调整最终结果的风格迁移强度。整体过程如图6所示。

在GAN网络结构前添加的编码器模块Mapping Network主要功能为将随机变量 z 转化为风格向量 w ,输入生成器模块中实现潜在引导合成。通过这种方法生成的风格向量具有较好的随机性,能够满足随机风格迁移的需求。本文采用的网络结构为8层MLP网络,在潜在引导合成实验中能够得到较好的结果。

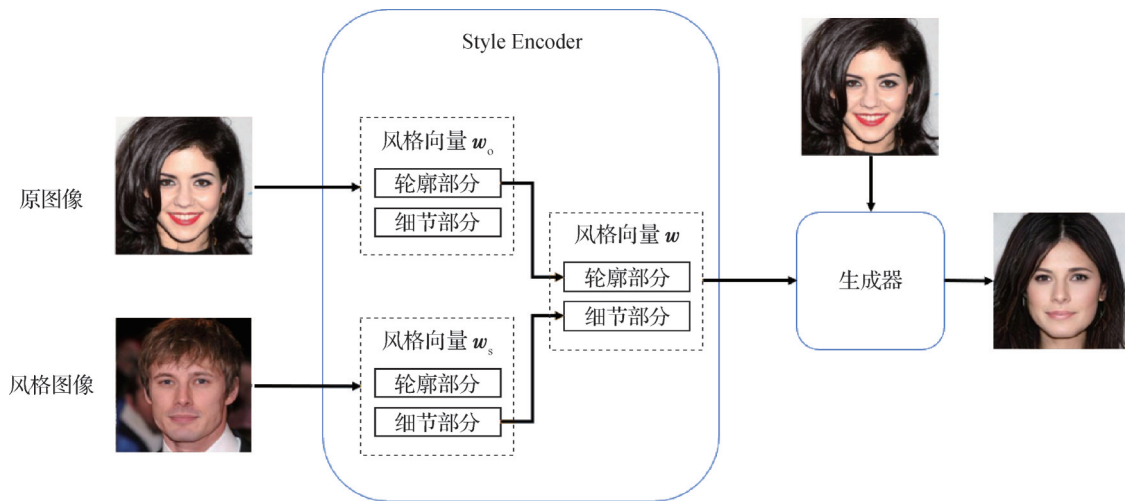


图6 风格迁移过程

Fig. 6 Style transfer process

2.2.2 GAN

生成器通过输入原图像 x 及风格向量 w 来得到具有 w 内含特征的图像风格迁移结果图像 x' 。本文生成器结构如图7所示。

生成器采用了经典的图像重建网络结构,由一个下采样区域和一个上采样区域两部分组成,其中下采样区域提取输入图像的图像特征,上采样区域进行图像重建,两部分之间采取了横向连接的方式,使输入图像特征能够传递到生成图像中,确保生成图像正确。为了防止横向连接输入的信息对生成图

像多样性带来影响,在横向连接中采用了掩膜(Wang等,2019)及高通滤波器处理,使通过此横向连接的图像信息只保留图像轮廓部分。

StarGAN v2在生成器中采用了AdaIN模块提取风格向量 w 中的特征信息,计算为

$$AdaIN(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y) \quad (5)$$

式中, $\mu(x)$ 和 $\sigma(x)$ 分别表示原图像的均值与方差, $\mu(y)$ 和 $\sigma(y)$ 分别表示风格图像的均值与方差,通过AdaIN方法使生成图像具有原图像与风格图像的特

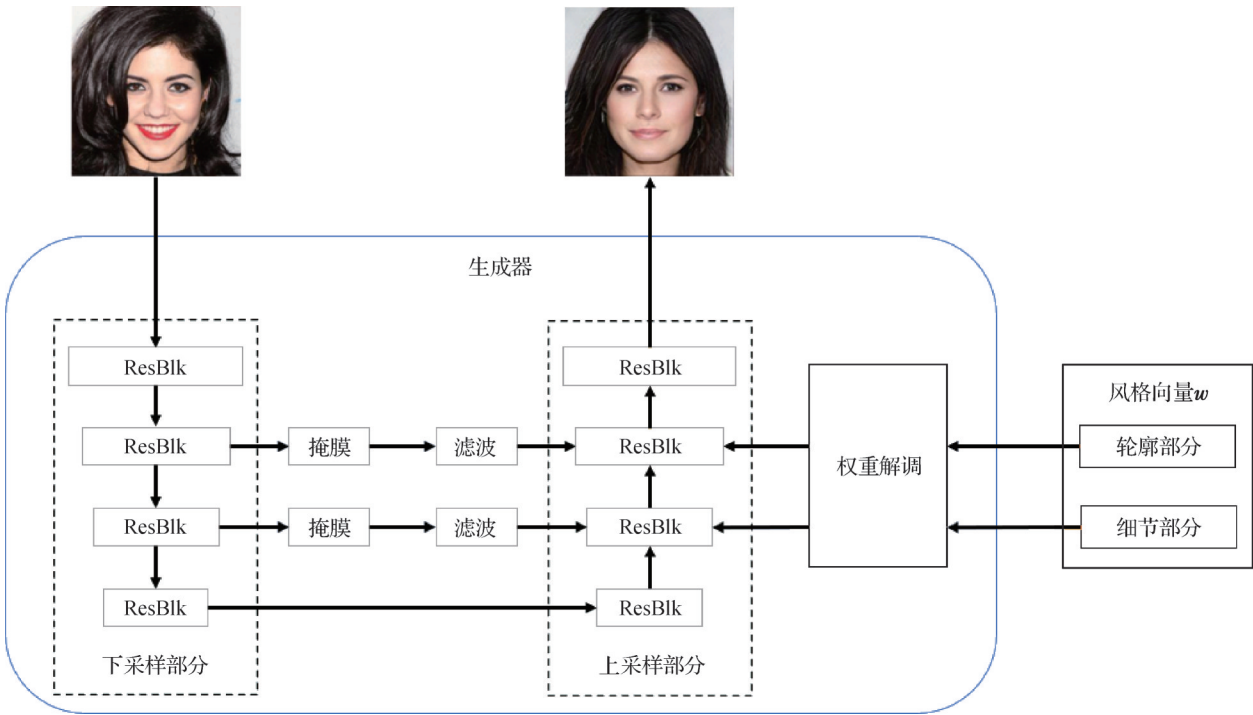


图7 生成器网络结构
Fig. 7 Generator structure

征。现有研究(Karras等,2020)表明,使用AdaIN模块是出现特征伪影问题的原因,如图8所示,左侧为生成图像,右侧为放大后的特征伪影区域。因此,基于权重解调不破坏特征图关联性的优点,本文采用权重解调算法代替AdaIN归一化算法作为风格的传输模块,该方法在保留完全可控性的同时消除了伪影。

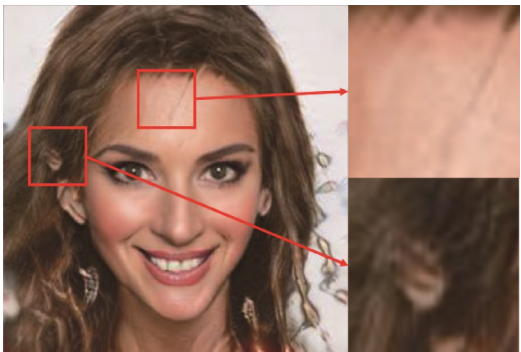


图8 特征伪影现象
Fig. 8 Water droplet artifacts

AdaIN结构可以拆分为两部分:标准化部分与调制部分。其中,标准化部分负责将输入标准化,调制部分负责将目标图像的信息通过参数的方式导入,即标准化外层的 $\sigma(y)$ 与 $\mu(y)$ 部分。这种方法分别对每个特征图的均值和方差执行归一化,因此

会破坏特征维度中相互关联的信息,导致生成器在生成图像时产生针对特征的局部峰值,从而产生特征伪影。权重解调算法通过去掉AdaIN中的归一化部分,并通过卷积的方式代替,避免了对关联特征信息的破坏,从而避免特征伪影问题。在调制之前插入一层卷积,缩放卷积权重计算为

$$w'_{ijk} = s_i \times w_{ijk} \tag{6}$$

式中, s_i 是第*i*个输入特征图的缩放比例。经过缩放和卷积后,输出激活的标准差为

$$\sigma_i = \sqrt{\sum_{i,k} w'^2_{ijk}} \tag{7}$$

获得标准差后,还需要对缩放后的权重进行解调,将输出特征图变为原来的单位标准差,解调部分计算为

$$w''_{ijk} = \frac{w'_{ijk}}{\sqrt{\sum_{i,k} w'^2_{ijk} + \epsilon}} \tag{8}$$

式中, ϵ 为极小量,避免分母为0。权重解调方法遵循归一化的思想,使输出的特征图有统一的分布均值和方差,在实现风格迁移的同时,避免了特征伪影的产生。

判别器Discriminator通过输入图像 x 以判断此图像的真伪,并输出一个0~1之间的值。输出越靠

近 1, 说明图像大概率为真, 反之则图像大概率为假。

本文的判别器结构为 6 层的 ResNet 组成的多任务判别器, 输出有多个枝干, 每个枝干对应着一个域, 枝干的输出即为输入图像在这个域中的概率。因此在训练生成器时, 判别器有两个输入: 生成器的输出图像 $\mathbf{x}$ 与此图像应属于的域 $\mathbf{y}$, 输出则为 $\mathbf{x}$ 属于 $\mathbf{y}$ 的概率。

2.3 损失函数

为了使生成器能够生成同时具有原图像与风格图像特征的图像, 本文主要使用了以下 5 类损失: 一般 GAN 损失、风格重建损失、风格多样性损失、图像重建损失和风格强度损失。

一般 GAN 损失用于规范生成器和判别器进行的博弈训练, 使判别器能够更精准地分辨真实图像与生成器生成的图像, 计算为

$$\mathcal{L}_{\text{adv}} = E_{\mathbf{x}, \mathbf{y}} [\log D_{\mathbf{y}}(\mathbf{x})] + E_{\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{z}} [\log(1 - D_{\tilde{\mathbf{y}}}(\mathbf{G}(\mathbf{x}, \tilde{\mathbf{s}})))] \quad (9)$$

式中, D 代表判别器, G 代表生成器, $\mathbf{x}$ 为输入图像, $\mathbf{y}$ 为图像所在的风格域, $\mathbf{z}$ 为随机变量, $\mathbf{s}$ 为生成的风格向量。此外, 本文还使用了 R1 正则化 (Mescheder 等, 2018), 通过梯度惩罚来稳定训练过程并防止模式崩溃, 计算为

$$R_1(\varphi) = \frac{\gamma}{2} E_{p(D(\mathbf{x}))} [\|\nabla D_{\varphi}(\mathbf{x})\|^2] \quad (10)$$

风格重建损失要求在进行风格迁移后, 得到的图像能够产生相同的风格向量, 具体为

$$\mathcal{L}_{\text{sty}} = E_{\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{z}} [\|\tilde{\mathbf{s}} - E_{\tilde{\mathbf{y}}}(\mathbf{G}(\mathbf{x}, \tilde{\mathbf{s}}))\|_1] \quad (11)$$

风格多样性损失要求生成的图像风格多样性尽可能高, 具体为

$$\mathcal{L}_{\text{ds}} = E_{\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{z}_1, \mathbf{z}_2} [\|\mathbf{G}(\mathbf{x}, \tilde{\mathbf{s}}_1) - \mathbf{G}(\mathbf{x}, \tilde{\mathbf{s}}_2)\|_1] \quad (12)$$

图像重建损失要求图像经过网络两轮转换后, 图像可以复原, 具体为

$$\mathcal{L}_{\text{cyc}} = E_{\mathbf{x}, \mathbf{y}, \tilde{\mathbf{y}}, \mathbf{z}} [\|\mathbf{x} - \mathbf{G}(\mathbf{G}(\mathbf{x}, \tilde{\mathbf{s}}), \hat{\mathbf{s}})\|_1] \quad (13)$$

风格强度损失为

$$\mathcal{L}_{\text{dsty}} = E_{\mathbf{x}, \mathbf{y}, \mathbf{z}} [\|\mathbf{G}(\mathbf{x}, \mathbf{s}) - \mathbf{y}_1\|_1] \quad (14)$$

$$\mathcal{L}_{\text{Dcoo}} = E_{\mathbf{x}, \mathbf{y}, \mathbf{z}} [\|\mathbf{G}(\mathbf{x}, \mathbf{s}) - \mathbf{x}_1\|_1] \quad (15)$$

这两个损失用于衡量生成图像与原图和目标图之间的差异。由于这两个损失互斥, 若将这两个损失的加权和作为一个整体损失, 能够在一定程度上

保证图像风格迁移的质量, 整体计算为

$$\mathcal{L}_{\text{Diff}} = \lambda_{D_{\text{sty}}} \mathcal{L}_{D_{\text{sty}}} + \lambda_{D_{\text{coo}}} \mathcal{L}_{D_{\text{coo}}} \quad (16)$$

最后, 总损失表示为

$$\min_{G, F, E} \max_D \mathcal{L}_{\text{adv}} + \lambda_{\text{sty}} \mathcal{L}_{\text{sty}} - \lambda_{\text{ds}} \mathcal{L}_{\text{ds}} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}} + \mathcal{L}_{\text{Diff}} \quad (17)$$

3 实验结果与分析

3.1 数据集构成

本文使用 CelebA-HQ 数据集对网络进行训练, 此数据集为高清人脸数据集, 包含 28 000 幅人脸图像。本文将 CelebA-HQ 数据集拆分为男性和女性两个域, 其中男性图像有 10 057 幅, 女性图像有 17 943 幅。除了这两个域标签外, 本文不使用任何其他的域标签。所有数据集中的图像大小均调整为 256×256 像素分辨率进行训练。

3.2 实验设置

本文实验使用 python 进行编程并在 NVIDIA RTX 2080Ti 上进行训练。对于 CelebA-HQ 数据集, 本文训练 100 000 轮, 每轮 2 个 batch。本文对比的基准模型为 MUNIT (multimodal unsupervised image-to-image translation) (Huang 等, 2018)、DRIT (diverse image-to-image translation) (Lee 等, 2018)、MSGAN (mode seeking generative adversarial network) (Mao 等, 2019)、pix2pix、CycleGAN、StyleGAN 以及 StarGAN v2 (Choi 等, 2020)。所有这些模型学习的均为两个域间的多模态映射, 实验条件与数据集均一致。

本文采用 FID (Fréchet inception distance score) (Heusel 等, 2017) 和 LPIPS (learned perceptual image patch similarity) (Zhang 等, 2018) 评估生成图像的视觉质量和多样性。通过计算数据集中每对图像域的 FID 和 LPIPS 及计算其平均值, 作为图像风格迁移的量化评价指标。对于所有生成模型, 取 200 幅生成图像并计算其平均值以进行量化评价指标对比。

3.3 潜在指导合成实验结果

潜在指导合成指不提供参考图像, 通过输入随机变量引导生成器进行图像生成。通过这种方法生成的图像风格由随机变量决定, 具有不确定性。本文通过 Mapping Network 结构生成随机风格向量, 实现潜在引导合成。输出结果如图 9 所示, 其中图 9(a) 为输入的原始图像, 图 9(b) 为男性域下输出的

随机风格迁移结果,图9(c)为女性域下输出的随机风格迁移结果。

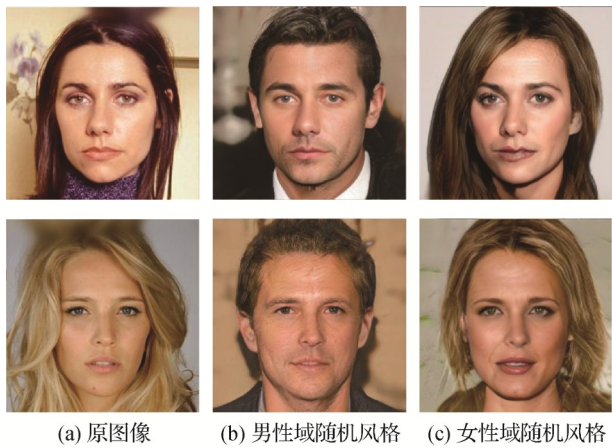


图9 潜在引导合成生成结果图

Fig. 9 Latent-guided synthesis results ((a) original images; (b) random male style; (c) random female style)

本文模型能够在潜在引导合成实验中生成风格差异较大且质量较高的风格迁移图像。量化指标如表1所示。表中同时对比了几个经典图像风格迁移算法。FID指标主要评估生成图像与原图像的相似性,此指标越小,说明与原图像越相似;LPIPS指标主要评估生成图像的多样性,此指标越大,说明图像多样性越丰富。由于本文在构建数据集时采用了性别别作为域标签,表中的FID及LPIPS指标为两组域标签指标的平均值。

表1 潜在引导合成实验量化指标

Table 1 Latent-guided synthesis quantitative indicator		
模型	FID	LPIPS
MUNIT	31.4	0.363
DRIT	52.1	0.178
MSGAN	33.1	0.389
StyleGAN	13.4	0.424
StarGAN v2	13.7	0.452
本文	12.9	0.544

注:加粗字体表示各列最优结果。

在潜在引导合成实验中,本文提出的模型能够达到更优的指标,对比原模型 StarGAN v2 可以获得更优的输出结果。StarGAN v2在潜在引导合成实验中采用了 AdaIN 归一化方法,这种方法在实现风格迁移后会出现特征伪影问题。而本文方法则使用

了权重解调方法,减小了结果图像的失真,因此在本实验中能够取得较好的结果。此外,本文模型在算法效率方面也优于 StarGAN v2,生成时间平均缩短了约 8%。特征伪影消除结果如图 10 所示,其中第 1 行为风格图像,第 2 行为本文得到的结果,第 3 行为 StarGAN v2 网络得到的结果。

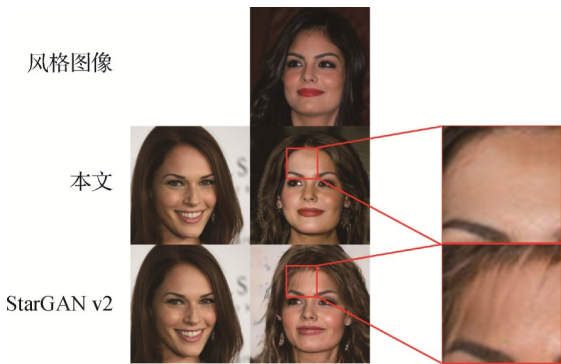


图10 特征伪影消除结果

Fig. 10 Remove water droplet artifacts results

3.4 参考指导合成实验结果

参考指导合成指提供目标图像,通过网络学习目标图像中的风格生成具有目标图像风格的原图像。本文通过 Style Encoder 结构实现参考引导合成,输出结果如图 11 所示,最左侧一列为原图像,第 1 行为风格图像,后两行为风格迁移结果。可以看到,本文得到的结果不仅保留了原图像的轮廓,同时学习到了风格图像的特征。



图11 参考引导合成生成结果图

Fig. 11 Reference-guided synthesis results

本文实验结果与 pix2pix 及 CycleGAN 算法结果的量化指标比较如表 2 所示。可以看出,本文提出的模型同样能达到最优指标。对比 StarGAN v2,本文模型提取了更丰富、层次更多的风格向量,能够

表 2 参考引导合成实验量化指标
Table 2 Reference-guided synthesis quantitative indicator

模型	FID	LPIPS
pix2pix	138.4	0.121
CycleGAN	73.2	0.284
MUNIT	107.1	0.176
DRIT	53.3	0.311
MSGAN	39.6	0.312
StarGAN v2	23.8	0.388
本文	20.7	0.406

注:加粗字体表示各列最优结果。

在细节层面上还原参考图像的特征。此外,StarGAN v2 使用的 Style Encoder 模块结构为 ResNet 网络,对于图像细节部分的信息提取较为困难。而本文模型采用了 FPN 网络较好地提取了图像的细节特征,并且对图像的特征进行分层处理,进一步提高风格迁移质量。图 12 显示了本文与其他算法结果的对比情况。

3.5 多风格强度实验结果

本文提出的模型可以通过调整 Style Encoder 中的参数 δ 调整输出图像的风格强度, δ 取值为 1~6 间的整数,数值越小越接近原图,越大越接近风格图,输出如图 13 所示。



图 12 参考引导合成对比图

Fig. 12 Reference-guided synthesis comparison results

((a) original images; (b) style images; (c) DRIT; (d) MSGAN; (e) StarGAN v2; (f) ours)

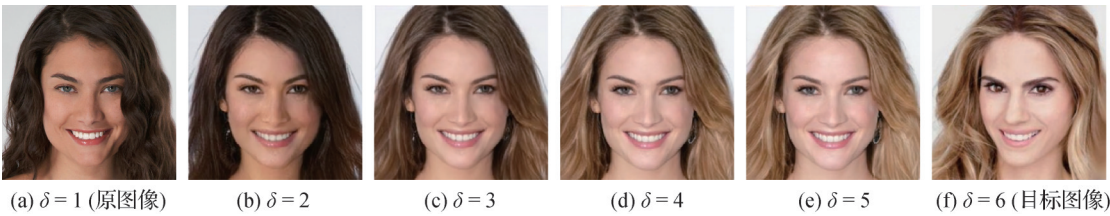


图 13 多风格强度实验结果图

Fig. 13 Multi-style intensity test results ((a) $\delta = 1$ (origin); (b) $\delta = 2$; (c) $\delta = 3$; (d) $\delta = 4$; (e) $\delta = 5$; (f) $\delta = 6$ (target))

由于StarGAN等人脸风格迁移网络提取了图像的整体风格特征,忽略了局部细节特征,因此不能较好地控制输出图像的风格。这些网络中生成的风格向量为一个耦合的整体,内部的风格特征难以单独解耦。而本文使用的Style Encoder模块所提取的风格向量具有多层结构,每一层所蕴含的风格特征各不相同。因此,通过输入不同的参数 δ ,Style Encoder可以将由原图像得到的风格向量 w_o 及由目标图像得到的风格向量 w_i 进行结合,生成不同的风格向量 w ,以实现风格强度的控制。

3.6 消融实验结果

本文在参考引导合成实验中,对比了FPN网络、权重解调方法和添加风格强度损失这3项改进对量化指标的提升,如表3所示。

表3 参考引导合成实验中的模型优化提升
Table 3 Reference-guided synthesis model optimization and promotion

模型	FID	LPIPS
StarGAN v2	13.7	0.452
+FPN网络	13.9	0.498
+权重解调	13.3	0.515
+风格强度损失	12.9	0.544

注:加粗字体表示各列最优结果。

由表3可知,在添加FPN网络之后,提升了整体模型的FID指标,这是因为FPN网络提取的风格向量大小为 6×256 ,大于原来的 1×64 ,包含了更多的风格特征信息,使生成器生成的图像更偏向于风格图像,而FID指标评估的是生成图像与原图像的相似度,因此FID指标得以提升。通过添加风格强度损失规范原图像与风格图像之间的风格平衡,本文模型即能得到更优秀的指标。

3.7 非正常面部实验结果

本文模型对非正常面部的风格迁移结果如图14所示。由于非正常面部图像与正常图像相比,存在不同程度的面部特征损失,且结构上具有较大差异,因此本文模型无法很好地对非正常面部图像进行风格迁移,亟待后续解决。

4 结论

本文针对人脸风格迁移的问题,提出了一种能



图14 非正常面部实验结果图
Fig. 14 Abnormal face test results
((a) hair covering; (b) eyeglass; (c) side face)

够生成不同风格强度图像的算法MStarGAN,使用FPN构建前置编码器,采用特征分类的思想提取信息更丰富且表达特征分层的风格向量。通过对原图像及风格图像的风格向量进行不同层数的合成,可以对生成图像的风格迁移强度进行一定的控制。此外,采用权重解调算法作为生成器中风格的传输模块,能够生成细节更加丰富的结果图像。

实验结果表明,本文算法在人脸迁移算法领域中能够得到良好的量化指标。与目前主流的人脸风格迁移网络相比,本文提取到更多的局部细节特征、减小了生成图像失真并使生成图像风格可变。

尽管本文方法能够得到良好的结果,然而,对于原图像中的细节特征解耦仍存在不足,无法将单一特征独立解耦,如鼻子、眼睛等特征。此外,对于一些非正常面部图像,如戴墨镜、遮眼、侧脸等,由于非正常面部图像与正常面部图像存在较大结构性差异,本文的迁移效果不尽如人意。

在未来的工作中,考虑参考Re-GAN(史彩娟等,2021)等方法,对编码器及生成器进行优化,使网络能够解耦出更细节的特征,并考虑在判别器部分添加约束条件,对生成器生成的图像进一步优化,获得对非正常面部图像的高质量风格迁移结果。

参考文献(References)

Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J. 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8789-8797 [DOI: 10.1109/CVPR.2018.00916]

- Choi Y, Uh Y, Yoo J and Ha J W. 2020. StarGAN v2: diverse image synthesis for multiple domains//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8185-8194 [DOI: 10.1109/CVPR42600.2020.00821]
- Denton E, Chintala S, Szlam A and Fergus R. 2015. Deep generative image models using a Laplacian pyramid of adversarial networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 1486-1494
- Elgammal A, Liu B C, Elhoseiny M and Mazzone M. 2017. CAN: creative adversarial networks generating “art” by learning about styles and deviating from style norms [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/1706.07068.pdf>
- Gatys L A, Ecker A S and Bethge M. 2015a. Texture synthesis using convolutional neural networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 262-270
- Gatys L A, Ecker A S and Bethge M. 2015b. A neural algorithm of artistic style [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/1508.06576.pdf>
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139-144 [DOI: 10.1145/3422622]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016a. Identity mappings in deep residual networks//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 630-645 [DOI: 10.1007/978-3-319-46493-0_38]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016b. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/1706.08500.pdf>
- Huang R, Zhang S, Li T Y and He R. 2017. Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2458-2467 [DOI: 10.1109/ICCV.2017.267]
- Huang X and Belongie S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1510-1519 [DOI: 10.1109/ICCV.2017.167]
- Huang X, Liu M Y, Belongie S and Kautz J. 2018. Multimodal unsupervised image-to-image translation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 179-196 [DOI: 10.1007/978-3-030-01219-9_11]
- Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]
- Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of StyleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]
- Lee H Y, Tseng H Y, Huang J B, Singh M and Yang M H. 2018. Diverse image-to-image translation via disentangled representations//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 36-52 [DOI: 10.1007/978-3-030-01246-5_3]
- Li T T, Qian R H, Dong C, Liu S, Yan Q, Zhu W W and Lin L. 2018. BeautyGAN: instance-level facial makeup transfer with deep generative adversarial network//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea (South): ACM: 645-653 [DOI: 10.1145/3240508.3240618]
- Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]
- Mao Q, Lee H Y, Tseng H Y, Ma S W and Yang M H. 2019. Mode seeking generative adversarial networks for diverse image synthesis//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1429-1437 [DOI: 10.1109/CVPR.2019.00152]
- Mescheder L, Geiger A and Nowozin S. 2018. Which training methods for GANs do actually converge?//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 3481-3490
- Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/1411.1784.pdf>
- Nirkin Y, Masi I, Tuan A T, Hassner T and Medioni G. 2018. On face segmentation, face swapping, and face perception//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi'an, China: IEEE: 98-105 [DOI: 10.1109/FG.2018.00024]
- Portilla J and Simoncelli E P. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1): 49-70 [DOI: 10.1023/A:1026553619983]
- Sbai O, Elhoseiny M, Bordes A, LeCun Y and Couprie C. 2018. DesIGN: design inspiration from generative networks//Proceedings of 2018 European Conference on Computer Vision. Munich, Ger-

- many: Springer: 37-44 [DOI: 10.1007/978-3-030-11015-4_5]
- Shi C J, Tu D J and Liu J Y. 2021. Re-GAN: residual generative adversarial network algorithm. *Journal of Image and Graphics*, 26(3): 594-604 (史彩娟, 涂冬景, 刘靖伟. 2021. Re-GAN: 残差生成式对抗网络算法. *中国图象图形学报*, 26(3): 594-604) [DOI: 10.11834/jig.200069]
- Shih Y C, Paris S, Barnes C, Freeman W T and Durand F. 2014. Style transfer for headshot portraits. *ACM Transactions on Graphics*, 33(4): #148 [DOI: 10.1145/2601097.2601137]
- Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W D and Webb R. 2017. Learning from simulated and unsupervised images through adversarial training//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE: 2242-2251 [DOI: 10.1109/CVPR.2017.241]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/1409.1556.pdf>
- Wang X Y, Bo L F and Fuxin L. 2019. Adaptive wing loss for robust face alignment via heatmap regression//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul, Korea (South): IEEE: 6970-6980 [DOI: 10.1109/ICCV.2019.00707]
- Xie B, Wang N and Fan Y W. 2020. Correlation alignment total variation model and algorithm for style transfer. *Journal of Image and Graphics*, 25(2): 241-254 (谢斌, 汪宁, 范有伟. 2020. 相关对齐的总变分风格迁移新模型. *中国图象图形学报*, 25(2): 241-254) [DOI: 10.11834/jig.190199]
- Yang S, Jiang L M, Liu Z W and Loy C C. 2022. Pastiche master: exemplar-based high-resolution portrait style transfer [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/2203.13248.pdf>
- Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]
- Zheng Z T and Liu J Y. 2020. P²-GAN: efficient style transfer using single style image [EB/OL]. [2022-12-22]. <https://arxiv.org/pdf/2001.07466.pdf>
- Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

作者简介

廖远鸿,男,硕士研究生,主要研究方向为计算机视觉和风格迁移算法。E-mail: liaoyuanhongsd@gmail.com

钱文华,通信作者,男,教授,博士生导师,主要研究方向为图像处理 and 计算机视觉。E-mail: whqian@ynu.edu.cn

曹进德,男,教授,博士生导师,主要研究方向为复杂网络与复杂系统、神经动力学与优化、多智能体系统。

E-mail: jdcao@seu.edu.cn