文章编号：2095-2163(2022)12-0001-08 中图分类号：TP391 文献标志码：A

基于 Pix2Pix 的人脸素描图像生成方法研究

陶知众，王斌君，崔雨萌，闫尚义
（中国人民公安大学 信息网络安全学院，北京 100038）

摘 要： 鉴于 Pix2Pix 在图像风格转换等图像翻译任务中存在细节丢失、生成图像模糊等问题，无法满足当前人脸素描生成任务的目标要求，提出一种改进的 Pix2Pix 模型。通过引入基于自注意力机制的残差卷积模块，让 Pix2Pix 的生成器和鉴别器在训练过程中能够为人脸图像的不同区域和通道赋予不同的权重，从而提高生成的人脸素描图像的质量，并且对 Pix2Pix 生成器的损失函数进行改进，使其生成的人脸素描图像更具有手绘风格。同时，针对生成对抗网络训练困难的问题，对原 Pix2Pix 的训练方法进行了改进。通过与 Pix2Pix 和 CycleGAN 对比，使用改进的 Pix2Pix 模型在训练过程中损失函数收敛更快、收敛过程更稳定，且生成的人脸素描图像在细节保留、轮廓清晰度等方面优于原 Pix2Pix 等模型，验证了改进 Pix2Pix 模型在人脸素描生成任务中的有效性。

关键词： 人脸素描生成；图像风格转换；生成对抗网络；自注意力机制；Pix2Pix

Research on the generation method of face sketch images based on Pix2Pix

TAO Zhizhong, WANG Binjun, CUI Yumeng, YAN Shangyi
(School of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China)

【Abstract】 Considering that Pix2Pix has problems such as loss of details and blurred generated images in image translation tasks such as image style transfer, it cannot meet the current target requirements of face sketch generation tasks. An improved Pix2Pix model is proposed. By introducing the residual convolution module based on the self-attention mechanism, the generator and discriminator of Pix2Pix can assign different weights to different regions and channels of the face image during the training process, thereby improving the quality of the generated face sketch image, and the loss function of the Pix2Pix generator is improved to make the generated face sketch images more hand-drawn. At the same time, in view of the difficulty of training the generative adversarial network and the instability of the convergence process, the training method of the original Pix2Pix is improved. Through comparative experiments, it is found that the loss function of the improved Pix2Pix model converges faster and the convergence process is more stable during the training process, meanwhile the generated face sketch images are better than the original Pix2Pix and other models in terms of detail retention and outline clarity. The research verifies the effectiveness of the Pix2Pix model on face sketch generation tasks.

【Key words】 face sketch synthesis; image style transfer; generative adversarial networks; self-attention mechanism; Pix2Pix

0 引 言

图像风格转换是指将一幅图像从所在的原图像域转换到目标图像域，使其在保留图像原本内容的同时又能具有目标图像域风格的一种图像处理技术。图像风格转换在社交娱乐和艺术创作领域具有十分广阔的应用前景，因此受到学术界和企业领域的高度关注。早期的图像风格转换被看作是图像纹理生成问题，即通过设置一定的约束条件，使生成的图像既包含了原图像的语义内容，又具有目标图像域的纹理特征[1]。而自深度学习问世以来，很多基于深度学习的图像处理算法也已相继提出，利用深度学习来处理图像风格转换问题的各种研究也陆续展开。Gatys 等人[2]提出了一种基于卷积神经网络的图像风格转换方法，通过预训练的 VGG-19[3]模型提取输入图像的内容特征图和风格特征图，并使用在此基础上定义的内容损失函数和风格损失函数生成图像，该方法生成的图像效果优于许多传统的机器学习算法。Goodfellow 等人[4]提出的生成对抗网络(Generative Adversarial Networks, GAN)因其生成图像质量高、易于实现、兼容各种网络模型等优点而倍受关注，很多基于 GAN 的风格转换模型也取得

基金项目：网络安全新业态视角下的关键技术风险分析及防控对策研究(20AZD114)。

作者简介：陶知众(1997-)，男，硕士研究生，主要研究方向：机器学习、数字图像处理；王斌君(1962-)，男，博士，教授，主要研究方向：网络安全执法技术计算机应用方面的研究；崔雨萌(1998-)，男，硕士研究生，主要研究方向：命名实体识别、文本分类；闫尚义(1998-)，男，硕士研究生，主要研究方向：自然语言处理、文本分类。

通讯作者：王斌君 Email：732184870@qq.com

收稿日期：2022-10-13

重大突破,其中包括 CycleGAN[5]、StarGAN[6] 及 Pix2Pix[7] 等。研究可知,CycleGAN 模型通过添加循环一致性损失函数,解决了在图像风格转换任务中缺少监督训练数据集的问题。StarGAN 模型则解决了多个图像领域间风格转换的问题,使其可以只经一次训练便可实现多个图像风格间的转换。Pix2Pix 模型则在 cGAN[8] 的基础上,将 U-Net[9] 作为生成器,PatchGAN 作为鉴别器,如此一来则可以生成质量较高的图像,并且因为其结构简单,易于训练等特点,目前在图像生成领域比较流行。

由于人脸图像细节较为丰富,而采用 Pix2Pix 模型很难捕捉到这些细节中所包含的信息,导致生成的人脸画像在五官、脸部轮廓等细节丰富部位会出现模糊、信息缺失等问题。文中针对该问题,提出一种改进 Pix2Pix 模型。在 Pix2Pix 基础上,研究的主要创新点包括:

(1)在原 Pix2Pix 模型的生成器和鉴别器中引入自注意力模块(Self-Attention Mechanism,SAM),使模型能够更好地学习到人脸的空间轮廓特点,从而解决生成图像在人脸五官等部位细节模糊或缺失等问题。

(2)在原 Pix2Pix 生成器的损失函数中引入了内容-风格损失函数,使生成器生成的素描图像在不丢失原图像细节内容的同时,在观感上更接近手绘素描图像。

(3)针对原 Pix2Pix 模型训练难度大、难以收敛等问题,提出了改进的训练方法,进而降低模型整体训练难度,加速模型收敛。

1 相关基础理论

1.1 Pix2Pix

GAN 是一种由生成器(Generator)和鉴别器(Discriminator)共同构成的深度学习模型。其中,生成器负责学习训练集输入数据的概率分布规律并生成具有相似概率分布的输出数据;鉴别器负责评估输入数据来自训练集或生成器的概率。训练过程中生成器和鉴别器一同训练,鉴别器的训练目标是能够正确区分输入数据是来自训练集或者生成器,而生成器的目标是尽量使鉴别器做出错误的判断。通过让 2 个模型进行对抗训练,使生成器生成数据的概率分布更接近真实数据,而鉴别器对生成数据和真实数据的鉴别能力也随之提高,并最终达到一种平衡状态。目前,GAN 越来越受到学术界重视,尤其是在计算机视觉领域,许多基于 GAN 的深度学习模型也逐渐进入学界视野,并已广泛应用在如图像风格转换[4-6]、超分辨率[10-11]、图像复原[12-13] 等图像处理任务上,继而不断向着其他领域扩展,具有广泛的应用前景[14-15]。

Pix2Pix 是由 Isola 等人[7] 提出的一种专门用于处理图像翻译问题的条件生成对抗网络模型。该模型包含了一个生成器和一个鉴别器,其中生成器可以根据输入图像生成其在目标图像域的对应图像,而鉴别器则是尝试分辨输入图像的真实性。Pix2Pix 模型结构如图 1 所示。

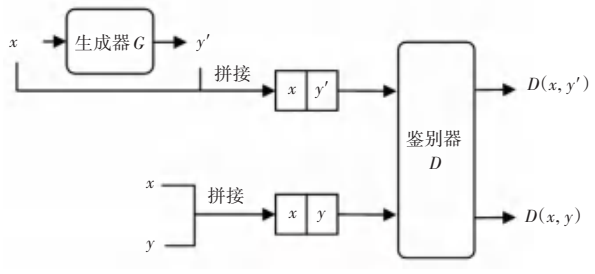


图 1 Pix2Pix 模型结构示意图

Fig. 1 Structure of Pix2Pix module

图 1 中,x,y 分别表示 2 个不同图像域 X,Y 中的图像。在训练生成器 G 时,将 x 输入到生成器中,生成具有 Y 图像域风格的图像 $y' = G(x)$。在训练鉴别器 D 时,则将 y 或 y' 和 x 一同输入到鉴别器 D 中,D 输出图像来自生成器 G 的概率。

Pix2Pix 模型的损失函数主要由条件对抗生成损失函数 l_{cGAN} 和 L_1 损失函数 l_{L1} 两部分组成,其中 l_{cGAN} 的表达式见如下:

$$l_{cGAN}(G,D) = E_{x,y\sim P_{data}(x,y)}[\log D(x,y)] + E_{x\sim P_{data}(x)}[\log(1 - D(x,G(x)))] \quad (1)$$

式(1)中,生成器以输入的真实图像作为条件,试图生成符合真实图像分布的对应虚假图像并欺骗鉴别器,因此生成器的训练目标是尽量减小;而鉴别器则在观察真实图像的基础上试图分辨输入的对应图像的真实性,因此鉴别器的训练目标是尽量增大。损失函数的表达式如式(2)所示:

$$l_{L_1}(G) = E_{x,y\sim P_{data}(x,y)}[\| y - G(x) \|_1] \quad (2)$$

损失函数用来确保生成器在生成虚假对应图像时,除了要考虑使虚假对应图像在概率分布上更接近真实对应图像外,还应使其在像素层面更接近于真实图像。因此,Pix2Pix 模型的最终损失函数具体如下:

$$G = \arg\min_G \max_D l_{cGAN}(G,D) + \gamma l_{L_1}(G) \quad (3)$$

其中,参数 γ 为 l_{L_1} 损失函数的权重,控制着条件对抗生成损失函数和 l_{L_1} 损失函数的相对重要性。

Pix2Pix 的生成器采用了 U-Net 框架。相较于

传统的编-解码器框架，Pix2Pix 生成器网络在第 i 卷积层和第 $n-i$ 卷积层之间增加了直连路径，其中 n 是生成器网络总层数，每一个直连路径会将第 i 层各信道信息拼接在第 $n-i$ 层各信道之后。通过增加直连路径，Pix2Pix 的生成网络可以使输入图像和输出图像共享低层信息，同时也确保了梯度信息能够在深层网络中有效传播，改善深层网络性能。同时，Pix2Pix 生成器网络还在某些层中使用了 *Dropout*，以取代 GAN 中作为输入的噪声。生成器的网络结构如图 2 所示。

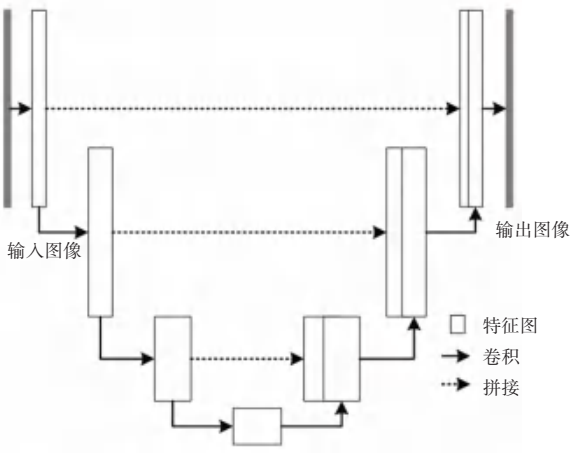


图 2　Pix2Pix 模型的生成器网络结构

Fig. 2　Generator network structure of Pix2Pix model

Pix2Pix 的鉴别器使用的是 PatchGAN 结构。不同于传统鉴别器，PatchGAN 的输出是一个 $n \times n$ 的矩阵，矩阵中每一个元素的值代表对输入图像对应图像区块的判别结果，训练过程中，再通过将鉴别器产生的矩阵元素均值作为整幅图像的最终判别结果，PatchGAN 通过将鉴别器的注意力集中在图像各个子区块的方式，使鉴别器可以更好地处理图像高频部分，同时，采用 PatchGAN 结构的鉴别器相较于传统分类网络具有更少的参数，更短的训练周期，并且通过调整 n 的大小，PatchGAN 可以应用于任意尺寸的图像，并使生成的图像保持较高质量。

1.2　自注意力机制

注意力机制（Attention Mechanism，AM）是一种改进神经网络的方法，主要是通过添加权重的方式，强化重要程度高的特征并弱化重要程度较低的特征，从而改善神经网络模型的性能[16]，注意力机制得到的权重既可以应用在信道上[17-18]，也可以应用在特征图或其它方面[19-20]。

自注意力机制是由 Zhang 等人[21]提出的一种专门用于生成对抗网络中的注意力机制变体，其结构如图 3 所示。针对卷积层的信息感受能力会受到卷积核大小的影响而无法高效捕捉到各个图像中同类物体的具体特征（如某种动物的毛发纹理特征、人的肢体结构特点等）这一问题，自注意力机制通过计算输入特征图中每一个位置在整个特征图中的权重，使整个网络可以更快注意到不同输入图像中各物体的空间和纹理特征，从而针对输入图像的不同部位分配不同的权重，达到增强生成图像质量的效果。鉴于在人脸素描生成任务中，输入人脸照片和输出的人脸素描图像在结构上具有高度的关联性以及相似性，因此自注意力机制可以帮助神经网络更快地定位人脸细节丰富区域，并且更好地学习到各部分的统计特征，从而提高最终生成的人脸素描图像的质量。

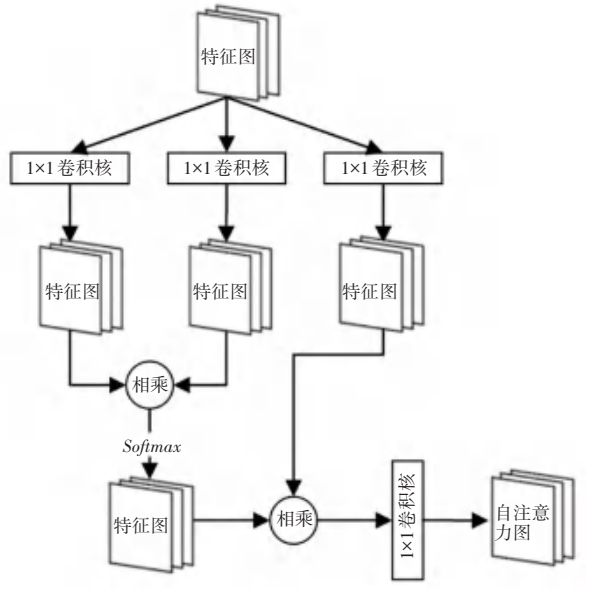


图 3　自注意力机制模块图

Fig. 3　Structure of self-attention mechanism

1.3　内容-风格损失函数

内容-风格损失函数（Content-Style loss Function）是由 Gatys 等人[2]在 2016 年提出的一种专门用于图像风格转换问题上的损失函数，其原理是使用预训练的神经网络分别对内容图像、风格图像和生成图像进行特征提取，通过计算提取到的特征图像之间的差异来衡量生成图像在内容和风格上与对应图像的差异。内容-风格损失函数由内容损失函数和风格损失函数两部分组成。其中，内容损失函数计算公式可表示为：

$$l_{content}(g,c)=\frac{1}{2}\sum_{l}\sum_{i,j}(\boldsymbol{F}^{l}_{i,j}-\boldsymbol{P}^{l}_{i,j}) \qquad (4)$$

其中，g 为生成图像；c 为内容图像；$\boldsymbol{F}^{l}$ 和 $\boldsymbol{P}^{l}$ 分别为预训练神经网络第 l 层提取的生成图像 g 和内

容图像 c 的特征图矩阵。

风格损失函数计算公式可表示为：

$$l_{style}(g,s)=\sum\frac{w_l}{4N_l^2M_l^2}\sum_{i,j}(\boldsymbol{G}_{i,j}^l-\boldsymbol{A}_{i,j}^l) \quad (5)$$

其中，g 为生成图像；s 为风格图像；$\boldsymbol{G}^l$ 和 $\boldsymbol{A}^l$ 分别为生成图像和风格图像在预训练神经网络第 l 层的风格特征矩阵；N 和 M 为第 l 层风格特征矩阵的行数和列数。Gatys 等人[2]将图像在神经网络第 l 层的风格特征矩阵定义为该层特征图的格拉姆矩阵(Gram matrix)，其计算公式可表示为：

$$\boldsymbol{G}_{i,j}^l=\sum_k\boldsymbol{F}_{ik}^l\boldsymbol{F}_{kj}^l \quad (6)$$

最终，内容-风格损失函数计算公式可表示为：

$$l_{cs}=a\cdot l_{content}(g,c)+b\cdot l_{style}(g,s) \quad (7)$$

其中，a 和 b 分别为内容损失函数和风格损失函数的权重。

2 基于自注意力机制和风格迁移的 Pix2Pix

2.1 模型结构

鉴于自注意力机制能更好地发现图像中大范围特征间的依赖关系，所以，在空间尺寸越大的特征图上、自注意力机制发挥的作用也就越大，但与此同时更大尺寸的特征图也会显著增加模型训练的时间成本。因此本文将自注意力机制添加到生成器网络中最后 3 层之间，以达到在增强生成图像质量目的同时尽量减小网络训练成本。文中提出的改进 Pix2Pix 模型的生成器网络模型如图 4 所示。

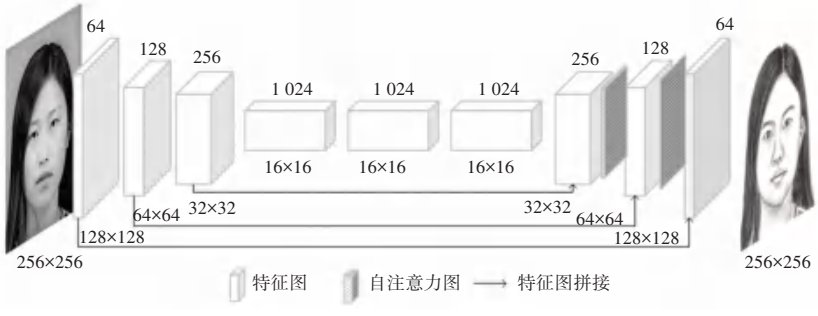


图 4 基于自注意力机制的 Pix2Pix 生成器网络结构

Fig. 4 Network structure of Pix2Pix generator based on self-attention mechanism

生成器的编码器卷积层参数设置均为：卷积核尺寸为 4×4，步长为 2，特征图边缘填充为 1，填充方式为镜像填充，激活函数使用 *LeakyRelu*，其参数设置为 0.2；解码器反卷积层参数设置为：卷积核大小为 4×4，步长为 2，特征图边缘填充为 1，填充方式为镜像填充，激活函数使用 *ReLu* 函数，前两层卷积网络使用 *Dropout*，概率设置为 0.5。鉴别器网络模型如图 5 所示。

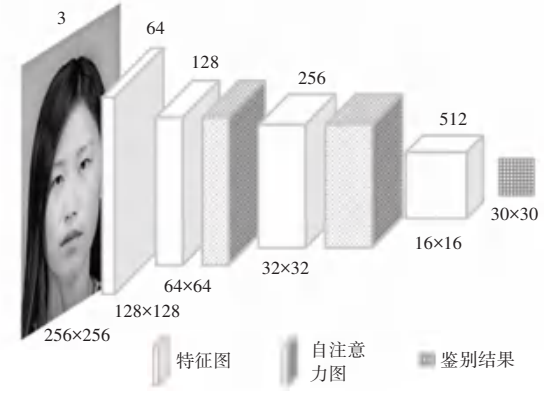


图 5 基于自注意力机制的 Pix2Pix 鉴别器网络结构

Fig. 5 Network structure of Pix2Pix discriminator based on self-attention mechanism

鉴别器卷积层参数设置为：卷积核尺寸为 4×4，特征图边缘填充为 1，填充方式为镜像填充，前三层卷积核步长为 2，后两层卷积步长为 1。所有卷积层都采用 *LeakyReLu* 为激活函数，其参数设置为 0.2。

由于在图像风格转换问题中，生成图像的风格主要依赖于风格图像的对比度信息，因此生成器在生成图像时应该尽量屏蔽内容图像中的对比度信息，而批归一化(Batch Normalization)[22]并不能很好地消除来自内容图像中的对比度信息，因此在改进的 Pix2Pix 模型的生成器网络和鉴别器网络中，使用实例归一化(Instance Normalization)[23]代替了批归一化。对于输入的一组特征图，IN 对每一特征图的每一信道进行归一化处理，从而更好地消除了每个特征图中包含的特殊信息，减少了图像生成过程中的干扰，并加快了生成器网络的收敛过程。

2.2 损失函数

改进 Pix2Pix 模型的损失函数的具体表达式为：

$$G=\arg\min_G\max_D l_{cGAN}(G,D)+\alpha\cdot l_{L_1}(G)+\beta\cdot l_{cs} \quad (8)$$

其中，$l_{cGAN}(G,D)$ 为 Pix2Pix 模型中生成器和鉴别器的对抗损失函数；l_{L_1} 为 Pix2Pix 生成器生成图像和手绘人脸图像的 L_1 损失；l_{cs} 为内容-风格损失函数，这里 a 设为 1，b 设为 0.1；α 和 β 分别为控制 l_{L_1} 损失函数和内容风格损失函数的权重，α 设为 100，β 设为 1。在计算内容损失函数 $l_{content}(g,c)$ 时，本文选择 VGG16 网络第二层中的第二个卷积层来提取生成素描和人脸照片的内容特征；而在计算风格损失函数 $l_{style}(g,s)$ 时，则选择 VGG16 网络中第四和第五层中的第一个卷积层来提取生成素描和对应手绘素描的风格特征。

2.3　改进训练方法

GAN 的训练是一个生成器和鉴别器互相博弈的过程，在这个过程中生成器试图生成与实际数据尽量相似的数据骗过鉴别器，而鉴别器则试图区分输入数据是否是真实数据，理论上，随着训练的进行，二者性能逐渐提高，并最终达到一种稳定状态。但在实际训练过程中，由于生成器和鉴别器网络训练难度不同、所采用的优化算法、学习率设置和数据集等因素影响，很难使 2 个网络同时收敛或达到纳什均衡，造成生成器部分或完全崩溃，以及某一模型收敛过快导致另一模型梯度消失等问题。因此，为了使 GAN 训练过程更稳定，文章采用的策略可做阐释论述如下。

（1）在生成器网络和鉴别器网络中使用谱归一化（Spectral Normalization）。根据 Ulyanov 等人[23]的研究，在生成器和鉴别器网络中使用谱归一化可以约束每层网络参数的谱范数，从而使网络参数在更新过程中变化更平滑，整个训练过程更加稳定。

（2）生成器和鉴别器采用不同的初始学习率及学习率调整策略。由于鉴别器的训练难度比生成器低，导致其损失很快收敛到一个非常低的值，无法为生成器梯度更新提供有效信息。因此，为了使生成器和鉴别器能够在训练过程中保持一种较为平衡的状态，让两者能够互相学习，在训练开始时分别为两者设置不同的学习率，并在随后的训练过程中根据具体训练效果采用不同的学习率更新策略。

3　实验结果与分析

实验的硬件平台为 QEMU Virtual CPU Version 2.5+，使用 NVIDIA Tesla V100-SXM2-32 GB 进行加速处理。数据集使用 CUFS（CUHK Face Sketch Database），该数据集共包含 606 对人脸-素描图像。实验选取 CUFS 数据集中 594 张素描人脸图像作为训练数据集；选取 CUFS 数据集中 12 张学生人脸图像作为测试图像；将所有训练图像和测试图像的大小缩放为 256 * 256 像素，并通过以 50% 的概率对人脸图像-素描对进行水平翻转和亮度随机调整的方式对数据集进行增强。生成器和鉴别器的优化器采用 Adam 算法，用于计算梯度以及梯度平方的运行平均值的参数 *beta*1 和 *beta*2 分别设置为 0.5 和 0.99，生成器的初始学习率设置为 1e-3，鉴别器的初始学习率设置为 1e-4。训练过程中，当生成器的损失函数无法下降、并超过 10 个 *epoch* 时，其学习率下降 10 倍；当鉴别器的损失函数无法下降、并超过 30 个 *epoch* 时，其学习率下降 10 倍。训练共进行 200 个 *epoch*，训练结束时生成器的学习率为 1e-8，鉴别器的学习率为 1e-8。

为更好地展示改进 Pix2Pix 模型在人脸素描图像生成任务上的有效性，本文将改进模型的生成人脸素描图像与 Pix2Pix 模型和 CycleGAN 模型生成的人脸素描图像进行对比，上述所有模型在相同实验平台上训练了 200 个 *epoch*。

3.1　改进训练方法效果比较

为验证本文提出的改进 GAN 训练方法的有效性，将原 Pix2Pix、分别采用谱归一化和不同学习率更新策略的 Pix2Pix 以及采用本文训练方法的 Pix2Pix 在实验数据集下分别训练 150 个 *epoch*，并观察在每个 *epoch* 后生成器损失函数值变化情况。最终结果如图 6 所示。

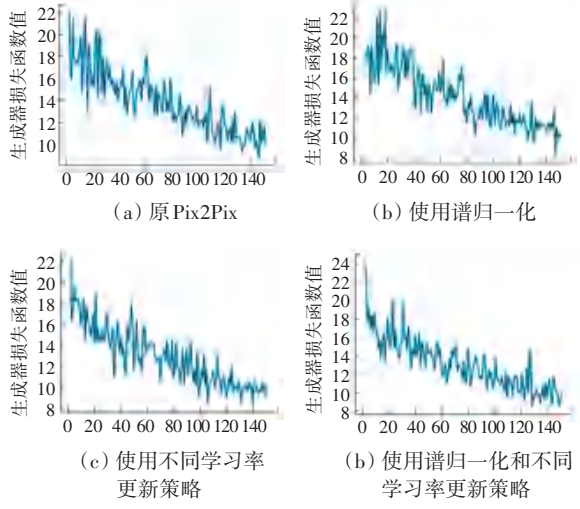


图 6　原 Pix2Pix 和采用不同训练方法后的 Pix2Pix 在 150 个 *epoch* 内损失函数变化对比

Fig. 6　Comparison of loss function changes of original Pix2Pix and Pix2Pix after using different training methods within 150 *epochs*

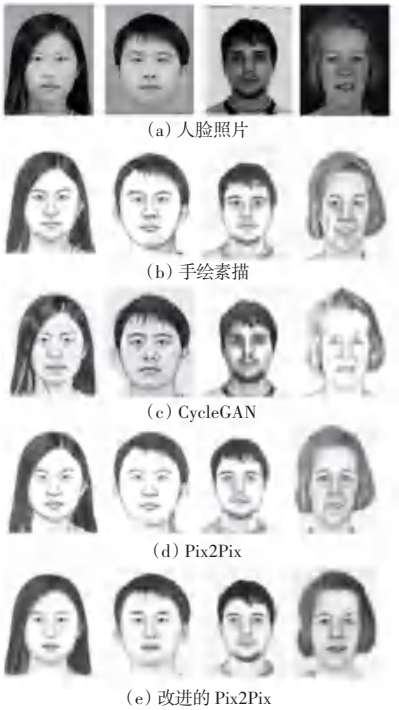
从图 6 可以看出，采用谱归一化和不同学习率

更新策略的 Pix2Pix 相比于原 Pix2Pix 生成器在训练过程中损失函数下降更快,但下降过程中仍然波动较大,而采用本文训练方法的 Pix2Pix 生成器在训练过程中不仅损失函数下降相比原 Pix2Pix 更快,下降过程中其波动也比其它 3 种更小,从而证明本文改进 GAN 训练方法的有效性。

3.2 生成图像质量比较

为更好地验证文中改进 Pix2Pix 模型在人脸素描生成任务中的有效性,除将其与原 Pix2Pix 模型进行对比外 ,还选择了 GycleGAN 模型与其进行对比分析。GycleGAN 模型作为图像翻译领域中另一经典模型,因其训练时不需要成对数据集、易于实现以及生成图像质量高等特点,一经提出便受到了广泛关注,因此选择将其作为参照对象可以使参照实验结果更具有代表性。

改进模型生成图像与其它模型生成图像对比如图 7 所示,通过对比发现,文中提出的改进 Pix2Pix 模型生成的人脸素描比 Pix2Pix 和 CycleGAN 生成的图像人脸轮廓更清晰,细节部分保留更完整,表情更明显,噪点更少,同时在整体观感上更接近人工绘制素描。



(a) 人脸照片

(b) 手绘素描

(c) CycleGAN

(d) Pix2Pix

(e) 改进的 Pix2Pix

图 7 生成图像质量对比

Fig. 7 Generated images quality comparison

3.3 生成图像量化比较

为量化评价改进 Pix2Pix 模型生成的图像质量,本文采用特征相似度 (Feature Similarity Index Measure, *FSIM*) 作为系统评价指标[24]。相较于 *SSIM*[25] 和 *MS - SSIM*[26],*FSIM* 充分考虑了图像视觉信息的冗余性和人类视觉系统主要通过低级特征来理解图像的特点,并且更偏向于清晰度较高的图像[27]。*FSIM* 通过计算 2 幅图像的相位一致区域和图像梯度幅值来评价这 2 幅图像在人类视觉系统中的相似度。其中,相位一致区域用来寻找一张数字图像在人类视觉系统中会被认为是“信息量丰富”的区域,而图像梯度幅值用来弥补相位一致性无法感知图像局部对比度变化对图像整体视觉效果产生影响的不足。在测试集上各模型所得 *FSIM* 分数见表 1。由表 1 数据可知,改进 Pix2Pix 模型在测试集上得分为0.648 3,相比原 Pix2Pix 模型和 CycleGAN 模型分别提高了0.020 6和0.027 6,从量化指标上进一步说明了文中提出的改进 Pix2Pix 模型在人脸素描生成任务中的有效性。此外,相比于原 Pix2Pix 和 CycleGAN 模型更低的分数方差也说明除生成的素描图像质量更好之外,改进 Pix2Pix 模型在稳定性上相较于其它对比模型也更有优势。

表 1 各模型在测试集上 *FSIM* 得分

Tab. 1 *FSIM* score of each model on the test set

算法	平均分	方差	标准差
CycleGAN	0.620 7	0.002 8	0.052 9
Pix2Pix	0.627 7	0.004 3	0.065 7
改进 Pix2Pix 模型	**0.648 3**	**0.002 1**	**0.045 8**

3.4 消融实验

本文通过消融实验对比分析,进一步验证了文中提出的改进 Pix2Pix 模型中各改进点在人脸素描生成任务中的优化效果,实验结果见表 2。从表 2 数据可知,原 Pix2Pix 在测试集上 *FSIM* 得分为 0.627 7,引入自注意力机制后,增强了原 Pix2Pix 模型细节特征提取能力,将测试集上 *FSIM* 分数提高了 0.108;而通过在生成器的损失函数中加入内容-风格损失函数,亦提高了模型在测试集上的表现。综合上述 2 种改进后,相较于原 Pix2Pix 模型,本文提出的改进 Pix2Pix 模型有效地提高了生成的人脸素描图像质量,说明了改进 Pix2Pix 模型在人脸素描生成任务中的有效性。

表 2 消融实验

Tab. 2 Ablation experiments

模型	自注意力机制	内容-风格损失函数	测试集上 *FSIM* 平均得分
Pix2Pix	×	×	0.627 7
优化模型 1	√	×	0.638 5
优化模型 2	×	√	0.631 7
改进 Pix2Pix	√	√	**0.648 3**

4　结束语

文中主要对 Pix2Pix 的生成器模型进行改进，将自注意力机制用于生成器和鉴别器网络中，减小无用信息对生成器的影响，加强生成器对输入图像中的人脸重要部分的学习，提升生成的人脸素描图像的质量；并在生成器损失函数中引入了内容-风格损失函数，使生成网络在生成人脸素描图像时既保留人脸照片中的细节部分，又能使图像更接近素描风格。同时，量化比较实验表明，改进 Pix2Pix 在测试集上的 *FSIM* 得分比 Pix2Pix 和 CycleGAN 分别高出了 2%和 2.7% ，进一步说明了改进 Pix2Pix 在人脸素描生成任务中的有效性。但与此同时，该改进模型依然存在一些问题，如对非正面拍摄的人脸图像效果较差。因此今后的工作便是提出能针对各种不同场景下不同角度的人脸图像也能生成质量较高的人脸素描图像的生成方法。

参考文献

[1] ZHAO Changshen. A survey on image style transfer approaches using deep learning[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1453(1): 012129.

[2] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2414-2423.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv: 1409.1556, 2014.

[4] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, 2014: 2672-2680.

[5] ZHU Junyan, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2223-2232.

[6] 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. Salt Lake City: IEEE, 2018: 8789-8797.

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

[8] MIRZA M, OSINDERO S. Conditional generative adversarial nets [J]. arXiv preprint arXiv: 1411.1784, 2014.

[9] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-assisted Intervention. Cham: Springer, 2015: 234-241.

[10] LEDIG C, THEIS L, HUSZáR F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4681-4690.

[11] WANG Xintao, YU Ke, WU Shixiang, et al. Esrgan: Enhanced super-resolution generative adversarial networks [M]//LEAL-TAIXÉ L, ROTH S. Computer Vision-ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science ( ). Cham: Springer, 2018, 11133: 63-79.

[12] YU Xiaoli, QU Yanyun, HONG Ming. Underwater-GAN: Underwater image restoration via conditional generative adversarial network [C]//International Conference on Pattern Recognition. Cham: Springer, 2018: 66-75.

[13] PAN Jinshan, DONG Jiangxin, LIU Yang, et al. Physics-based generative adversarial models for image restoration and beyond [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(7): 2449-2462.

[14] WANG Zhengwei, SHE Qi, WARD T E. Generative adversarial networks in computer vision: A survey and taxonomy[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-38.

[15] 汪美琴，袁伟伟，张继业. 生成对抗网络 GAN 的研究综述[J]. 计算机工程与设计，2021，42(12)：3389-3395.

[16] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[17] HU Jie, SHEN Li, SAMUEL A. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.

[18] GAO Zilin, XIE Jiangtao, WANG Qilong, et al. Global second-order pooling convolutional networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 3024-3033.

[19] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, 2014: 2204-2212.

[20] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks [C]// Advances in Neural Information Processing Systems. Montréal, Canada: NIPS Foundation, 2015: 2017-2025.

[21] ZHANG Han, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks [C]//International Conference on Machine Learning. Jiangxi: PMLR, 2019: 7354-7363.

[22] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// International Conference on Machine Learning. Lile, France: PMLR, 2015: 448-456.

[23] ULYANOV D, VEDALDI A, LEMPITSKY V. Instance normalization: The missing ingredient for fast stylization [J]. arXiv preprint arXiv: 1607.08022, 2016.

[24] ZHANG Lin, ZHANG Lei, MOU Xuanqin, et al. FSIM: A feature similarity index for image quality assessment [J]. IEEE Transactions on Image Processing, 2011, 20(8): 2378-2386.

[25] HORÉ A, ZIOU D. Image quality metrics: PSNR vs. SSIM [C]//2010 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE, 2010: 2366-2369.

（下转第 15 页）