DOI:10.13878/j.cnki.jnuist.20221012003

刘航[1]　李明[1]　李莉[2]　付登豪[3]　徐昌莉[1]

基于生成对抗网络的图像风格迁移

摘要

生成对抗网络(Generative Adversarial Network,GAN)可以生成和真实图像较接近的生成图像.作为深度学习中较新的一种图像生成模型,GAN在图像风格迁移中发挥着重要作用.针对当前生成对抗网络模型中存在的生成图像质量较低、模型较难训练等问题,提出了新的风格迁移方法,有效改进了BicycleGAN模型实现图像风格迁移.为了解决GAN在训练中容易出现的退化现象,将残差模块引入GAN的生成器,并引入自注意力机制,获得更多的图像特征,提高生成器的生成质量.为了解决GAN在训练过程中的梯度爆炸现象,在判别器每一个卷积层后面加入谱归一化.为了解决训练不够稳定、生成图像质量低的现象,引入感知损失.在Facades和AerialPhoto&Map数据集上的实验结果表明,该方法的生成图像的PSNR值和SSIM值高于同类比较方法.

关键词

生成对抗网络;风格迁移;自注意力机制;谱归一化;感知损失

中图分类号 TP391.4

文献标志码 A

收稿日期 2022-10-12

资助项目 国家自然科学基金(61877051,61170192);重庆市科委重点项目(cstc2017zdcy-zdyf0366);重庆市教委项目(113143);重庆市研究生教改重点项目(yjg182022)

作者简介

刘航,女,硕士生,研究方向为深度学习、计算机视觉.406369177@qq.com

李明(通信作者),男,博士,教授,研究方向为机器学习、计算机视觉.20131052@cqnu.edu.cn

1 重庆师范大学 计算机与信息科学学院,重庆,401331

2 西南大学 计算机与信息科学学院,重庆,400715

3 电子科技大学 经济与管理学院,成都,611731

0 引言

图像风格迁移是目前计算机视觉领域的研究热点.风格迁移的目的是将一幅图像转换成另一幅或多幅具有特定目标的图像,例如:输入一张纯色马的图片,输出的是斑马的图片;输入一张油画的图片,输出的是中国画的图片;等等.图像风格迁移不仅可以显著降低获取相关数据集的成本,而且还可以创建源数据以外的新图像,所以利用生成模型扩大研究数据集可有效提高深度学习网络模型的训练质量.

Goodfellow等[1]提出一种生成对抗网络模型,该模型由生成网络和判别网络组成,这两个网络在互相博弈的过程中优化彼此.随着判别网络的辨伪能力不断增强,生成网络产生的数据将更接近真实数据,生成对抗网络与其他网络相比具有更好的数据生成能力,因此生成对抗网络在图像生成和风格迁移领域得到了广泛应用.Isola等[2]提出的Pix2Pix模型作为有监督图像风格迁移的代表作,它通过有监督训练成对图像,得到的是一对一的风格迁移图像.虽然Pix2Pix模型得到的生成图像接近真实图像,但是该模型的训练需要大量的成对图像数据,而现实中收集成对数据集较困难,限制了其推广应用.Zhu等[3]提出的CycleGAN不需要训练成对的数据集,它是无监督风格迁移任务的经典模型,并且仅使用生成器和判别器完成图像域的风格迁移,首次实现了不成对图像之间的变换,在风格迁移领域得到了广泛的应用,但该模型并不能生成多样的结果.针对CycleGAN生成结果单一的问题,Zhu等[4]提出的BicycleGAN是一种基于条件遗传算法的混合模型,结合了cVAE-GAN[5]和cLR-GAN[6]的优点,学习两个图像域之间的多模态映射,有助于产生更多样化的结果.Park等[7]提出的CUT将对比学习应用到图像风格迁移,实现了一种轻量级的图像风格转换模型.尽管已有方法将一幅图像转换成另一幅或多幅图像时表现良好,但是由于输入生成器的生成图像与真实图像存在域差异,因此在模型收敛后得到的生成图像往往伴随着噪声和细节信息的丢失,使得图像风格迁移的质量仍有提升空间.在网络训练过程中数据之间的整体差异较小,存在较多极端值干扰的情况下,往往会导致模型训练变得不稳定.

本文针对上述问题,提出了以下几个方面的改进:

1)将残差模块引入BicycleGAN的生成器,解决GAN再训练中容易出现的退化现象;

2)将注意力机制引入 BicycleGAN 的生成器,在提取图像局部特征的同时,又注意全局特征,获得更多的图像特征,并使得图像风格转换过后的图像与真实图像保持特征一致性;

3)判别器每层卷积后面加入谱归一化,使得判别器和生成器在对抗训练中趋于稳定;

4)引入感知损失,能够稳定训练,提升生成图像的质量.

1 相关工作

1.1 BicycleGAN

生成对抗网络的思想是利用博弈不断优化生成器和判别器从而使得生成的图像更加逼近真实样本,在图像风格迁移领域已经证明了生成对抗网络在图像合成中的巨大潜力.传统的图像风格迁移模型都是输入一张图片只能产生一种风格,缺乏多样性.为了避免输出的单一性,由 Zhu 等[3]提出的 BicycleGAN 网络模型,首次尝试输入图像可以得到多种对应样式的输出图像,强制生成器不得忽略噪声,使用噪声来获得样化的图片.BicycleGAN 可以找到潜在编码 Z 与目标图像 B 之间的关系,因此生成器可以在给定不同的潜在编码 Z 时学会生成不同的风格图像 $\hat{B}$. BicycleGAN 通过组合 cVAE-GAN 和 cLR-GAN 这两种模型来提高性能,在大量的图像风格迁移问题中产生多样化和视觉上吸引人的图像结果.BicycleGAN 的模型结构如图 1 所示.由于直接用随机噪声来产生多样性结果,会存在模式崩溃、训练不稳定问题,本文对其模型进行优化并得到了更好效果.

1.2 残差块

在深度学习中,网络的层数越多,意味着能够提取到的特征越丰富,并且越深的网络提取的特征越抽象、越具有语义信息[8],但如果简单地增加深度,会导致退化问题.随着网络层数增加,在训练集上的准确率趋于饱和甚至下降.为了解决这种退化现象,ResNet 被 He 等[8]提出,其结构如图 2 所示.残差网络的思想就是将网络学习的映射从 X 到 Y 转为学习从 X 到 $Y-X$ 的差,然后把学习到的残差信息加到原来的输出上即可.即便在某些极端情况下,这个残差为0,那么网络就是一个 X 到 Y 的恒等映射.残差块一个通用的表示方式是:

$$y_l = h(x_l) + \mathcal{F}(x_l, W_l), \quad x_{l+1} = f(y_l), \tag{1}$$

其中,x_l 和 x_{l+1} 分别是第 l 层的输入和输出,F 是一个残差函数,$h(\cdot)$ 是恒等映射,$f(\cdot)$ 是激活函数.残差网络跳过了一些网络层直接与后面某一层的输出结果进行连接.随着深度的增加,可以获取更高的精度,因为其学习的残差越准确.本文通过在生成器中引入残差模块,使得网络训练过程更加简单,有效缓解了网络退化的问题.

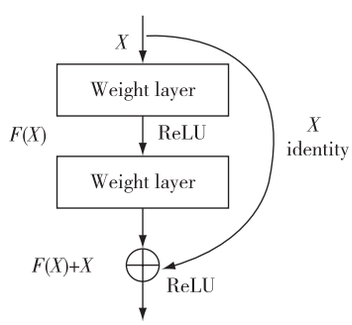

图 2 残差网络结构

Fig. 2 Residual network structure

1.3 注意力机制

人的眼睛可以有选择性地看自己关注的事物,

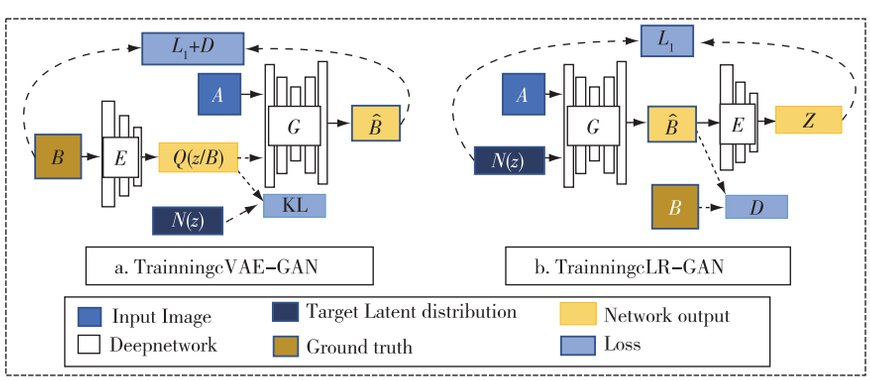

图 1 BicycleGAN 模型结构

Fig. 1 BicycleGAN model structure

从而忽略一部分不感兴趣的信号,重点聚焦自己感兴趣的事物,这就是注意力机制.注意力机制能够使得模型在生成图片过程中可以自己关注不同特征区域.如图 3 所示,输入数据通过卷积初始化,使用数量为 C/K ($k=8$) 且大小为 1×1 的卷积核来对输入数据执行卷积运算,以获得特征图 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$,使用一个数量为 C 且大小为 1×1 的卷积核对输入数据执行卷积运算以获得特征图 $h(\boldsymbol{x})$.经过局部自我注意特征图计算,重新调整特征图尺寸.然后计算 $f(\boldsymbol{x})$ 与 $g(\boldsymbol{x})$ 转置相乘,通过 softmax 归一化后得到 Attention Map.再让 Attention Map 与 $h(\boldsymbol{x})$ 每个像素点相乘,得到自注意力 feature maps.

其数学表达式如下:

$$\beta_{j,i}=\frac{\exp(\boldsymbol{s}_{ij})}{\sum_{i=1}^{N}\exp(\boldsymbol{s}_{ij})},\quad \boldsymbol{s}_{ij}=f(\boldsymbol{x}_i)^{\mathrm{T}}g(\boldsymbol{x}_j). \tag{2}$$

在局部信息的基础上增加全局信息得到:

$$\boldsymbol{o}_j=\sum_{i=1}^{N}\beta_{j,i}h(\boldsymbol{x}_i). \tag{3}$$

注意力层的最终输出为

$$\boldsymbol{y}_i=\boldsymbol{\gamma}_{o_i}+\boldsymbol{x}_i. \tag{4}$$

前一个隐藏层的图像特征向量 $\boldsymbol{x}\in\mathbf{R}^{C\times N}$,转化为 3 个特征空间 $f(\boldsymbol{x})$、$g(\boldsymbol{x})$ 和 $h(\boldsymbol{x})$,$\beta_{j,i}$ 表示合成第 j 个区域时模型关注第 i 个区域的程度,然后输出注意力层是 $\boldsymbol{o}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\boldsymbol{o}_3,\cdots,\boldsymbol{o}_j,\cdots,\boldsymbol{o}_N)\in\mathbf{R}^{C\times N}$,再乘以一个初始化为 0 的可学习权重 γ,并且添加特征向量 $\boldsymbol{x}$,通过反向传播不断更新.本文在生成器网络中引入自注意力机制,使其能够充分发现图像内部表征的整体性和长期依赖性,有效地降低了训练的计算量,使得训练更加稳定.

1.4 谱归一化

谱归一化 (Spectral Normalization,SN)[9] 通过限制每一层的频谱范数来约束判别器的 Lipschitz 常数,从而提高生成对抗网络的稳定性.谱归一化与其他归一化方法相比计算成本较小,不需要额外的超参数调整.它通过约束判别器 D 中每层 f 的权重矩阵的 L2 矩阵范数来控制 Lipschitz 常数.对于线性层 $f(\boldsymbol{x})=\boldsymbol{W}\boldsymbol{x}$,给出了它的 Lipschitz 范数如式(6)所示.根据定义,其中 $\sigma(\boldsymbol{W})$ 代表矩阵 $\boldsymbol{W}$ 的 L2 矩阵标准,它也等于 $\boldsymbol{W}$ 的最大奇异值.

$$\|f\|_{\mathrm{Lip}}=\sup_{\boldsymbol{x}}(\nabla f(\boldsymbol{x}))=\sup_{\boldsymbol{x}}\sigma(\boldsymbol{W}), \tag{5}$$

$$\sigma(\boldsymbol{W})=\max_{\boldsymbol{x}\neq0}\frac{\|\boldsymbol{W}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}=\max_{\|\boldsymbol{x}\|_2\leqslant1}\|\boldsymbol{W}\boldsymbol{x}\|_2. \tag{6}$$

如果为每层选择的激活函数 a 的 Lipschitz 范数为 1,根据范数相容性,可以获得判别器 D 中的 Lipschitz 范数的边界,如不等式(7),其中 L 是 D 的层数.

$$\|D\|_{\mathrm{Lip}}\leqslant\prod_{l=1}^{L+1}\boldsymbol{W}^l\boldsymbol{x}_{l-1,\mathrm{Lip}}=\prod_{l=1}^{L+1}\sigma(\boldsymbol{W}^l). \tag{7}$$

因此,需要一种频谱归一化方法来确保 $\sigma(\boldsymbol{W})$ 始终等于 1,谱归一化如等式(8) 所示:

$$\overline{\boldsymbol{W}}_{\mathrm{SN}}(\boldsymbol{W})=\frac{\boldsymbol{W}}{\sigma(\boldsymbol{W})}. \tag{8}$$

式(8)用于归一化每层的权重矩阵 $\boldsymbol{W}^l$,从而得到 $\sigma(\overline{\boldsymbol{W}}_{\mathrm{SN}}(\boldsymbol{W}^l))=1$ 使得 D 可以满足 1-Lipschitz 约束.判别器的训练不稳定性问题转化为获取最大奇异值 $\sigma(\boldsymbol{W}^l)$ 的问题,$\sigma(\boldsymbol{W}^l)$ 可通过应用幂迭代法确定.

2 基于生成对抗网络的风格迁移模型

2.1 生成器模型

本文改进和优化了原始 GAN 的生成器部分,引入残差块和自注意力机制,使得在图像风格迁移过程中生成图像的真实性有较大提高,改善了生成图像的质量.所设计的生成器由编码器、转换器、解码

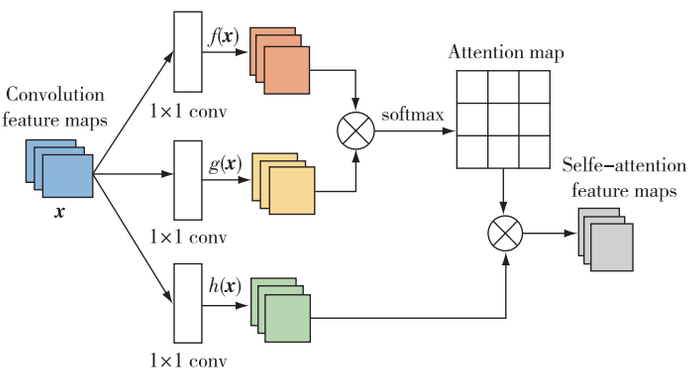

图 3 自注意力机制结构

Fig. 3 Self-attention mechanism structure

器三部分组成,编码器由卷积神经网络组成、转换器由自注意力机制网络层与残差网络块结构组成、解码器由反卷积神经网络组成,生成器网络如图 4 所示.

整个生成器的网络结构参数配置如表 1 所示,编码器和转换器使用实例归一化 (Instance Normalization,IN)[10] 和 ReLU 激活函数,解码器最后一层使用 Tanh 激活函数.自注意力机制有助于在图像的相邻区域之间建立长期的、多层次的依赖关系.本文通过在生成器中引入自注意力机制,更好地在局部依赖基础上增加全局依赖,这有助于生成器合成更逼真的风格迁移图像.为了避免因网络层数的增加而导致梯度消失的问题,本文在生成器中引入残差块,模型的训练速度得到提升,残差网络层与层之间的跳跃连接可以保留原图一部分没有进行风格迁移的完整信息,从而提高了图像风格迁移后的视觉效果.

在生成器中对真实图像 B 进行编码,以提供潜在矢量的真实样本并从中进行采样.首先使用生成器从随机噪声中生成伪图像 $\hat{B}$,然后对 $\hat{B}$ 进行编码,最后计算其与输入随机噪声的差异.向前计算步骤为首先随机产生一些噪声,然后串联图 A 以生成伪图像 $\hat{B}$,将真实图像 B 编码为多元高斯分布的潜在编码,然后从它们中采样以创建噪声输入,再用同样的编码器将伪图像 $\hat{B}$ 编码为潜矢量.最后,从编码的潜矢量中采样,并用输入噪声 Z 计算损失.损失函数为式(10) 和式(11).

$$L_1(G,E)=\mathbb{E}_{A\sim P(A),Z\sim P(Z)}\|z-E(G(A,z))\|_1, \tag{10}$$

$$G^*,E^*=\arg\min_{G,E}\max_{D}L_{1\mathrm{GAN}}(G,D)+\lambda_1 L_1(G,E), \tag{11}$$

G 是生成器,D 是判别器,E 是编码器,λ 为设置的超参数,嵌入 Z 的潜在编码被生成器网络用来保持接近实际测试时间分布 $p(z)$,L_1 损失可以体现重构后的图像轮廓,GAN 能更好地锐化图像的清晰度.

2.2 判别器模型

本文判别器网络采用 Wang 等[11] 提出的 PatchGAN 作为模型的判别器,用来对输入的生成图像与真实图像进行真伪判别.判别器网络如图 5 所示.

PatchGAN 将输入的图像划分为 70×70 的多个小块,小块代表输入图像的感受野,然后对每个图像块进行真伪判断,其输出二维矩阵中每一个元素的值表示每个图像块是真实样本的概率,每个图像块真实概率的平均值作为最终整体图像的判定结果.该判别器可以很大程度上保持图像的高分辨率和细节.为了缓解梯度消失从而增加模型的稳定性,本文在判别器每层卷积后面加入谱归一化.判别器的网络结构如表 2 所示.

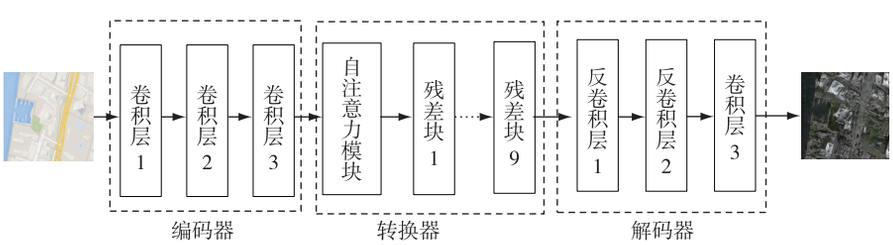

图 4 生成器网络

Fig. 4 Generator network

表 1 生成器网络结构参数配置

Table 1 Parameter configuration for generator network structure

模块	序号	层类型	数量	核尺寸	步长	深度	归一化	激活函数
编码器	0	Convolution	1	7×7	1	64	IN	ReLU
编码器	1	Convolution	1	4×4	2	128	IN	ReLU
编码器	2	Convolution	1	4×4	2	256	IN	ReLU
转换器	3	Self-Attention	1					
转换器	4	Residual Block	9	4×4	1	256	IN	ReLU
解码器	5	Deconvlution	1	4×4	2	128	IN	ReLU
解码器	6	Deconvlution	1	4×4	2	64	IN	ReLU
解码器	7	Convolution	1	7×7	1	3		Tanh

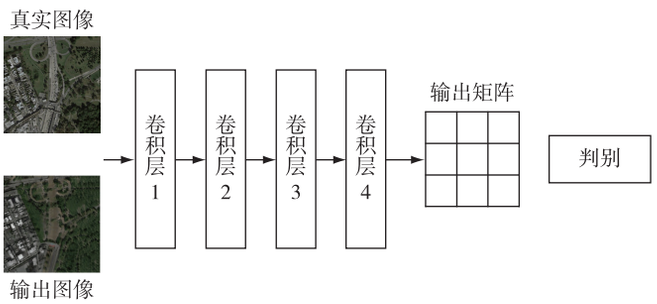

图 5　判别器网络

Fig. 5　Discriminator network

表 2　判别器结构参数配置

Table 2　Parameter configuration of discriminator structure

序号	层类型	核尺寸	步长	深度	归一化	激活函数
0	Convolution	4×4	2	64	SN	LeakyReLU
1	Convolution	4×4	2	128	SN	LeakyReLU
2	Convolution	4×4	2	256	SN	LeakyReLU
3	Convolution	4×4	2	512	SN	LeakyReLU
4	Convolution	4×4	1	1	SN	

2.3　感知损失

Johnson 等[12]利用感知损失增强图像细节.为了生成图像的真实视觉效果,本文使用感知损失来优化生成网络.感知损失依赖训练的 VGG16 模型提取图像高级特征.先提取生成图像和原始图像的特征,然后计算它们之间的差异.为了最大限度地减少细节的丢失,应用感知损失来提高细节保护能力,如等式(12)所示:

$$L_{\mathrm{pl}}=\frac{1}{dwh}\left\|\varphi(G(\boldsymbol{x},c))-\varphi(\boldsymbol{x})\right\|_2^2, \tag{12}$$

其中,$\boldsymbol{x}$ 是输入数据,c 是输入目标属性标签,$G(\boldsymbol{x},c)$ 是生成的数据,φ 是特征提取函数,d、w 和 h 分别表示深度、宽度和高度.损失函数将通过真实图像卷积获得的特征与通过生成图像卷积获得的特征进行比较,以使内容信息和全局结构信息接近.

3　实验与结果分析

3.1　实验平台与数据集

本文实验环境如表 3 所示.

本文使用 Facades 数据集和 AerialPhoto&Map 数据集作为实验数据集.Facades 数据集包含不同领域风格的建筑物图像,AerialPhoto&Map 数据集包含 Google Maps 网站上获取的纽约市及其附近的卫星航拍图与导航路网图匹配图像.本研究工作分别取 Facades 数据集中的 400 幅图像和 AerialPhoto&Map 中的 1 096 幅图像作为实验数据集.Facades 数据集中 320 幅图像用作训练集,80 幅图像用做测试集,AerialPhoto&Map 数据集中 600 幅图像用做训练集,496 幅图像用作测试集.

表 3　实验环境

Table 3　Experimental environment

实验环境	版本
操作系统	Ubuntu 16.04 LTS
CPU	Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20 GHz
GPU	GTX 1080 Ti * 1
Pytorch	1.4.0
CUDA	11.3

3.2　实验结果

在 Facades 和 AerialPhoto&Map 数据集上分别进行了图像风格迁移实验.为了验证本文方法的有效性,将本文方法的实验结果与 Pix2Pix、Cyclegan、CUT、BicycleGAN 进行了对比.

3.2.1　在 Facades 数据上实验结果

Facades 数据集上的实验结果如图 6 所示,目的是将输入语义图还原为真实建筑图像.图 6 中第 1 列是输入的建筑语义图像,第 2、第 3、第 4 和第 5 列分别为 Pix2Pix 模型、Cyclegan 模型、CUT 模型和 BicycleGAN 模型的图像风格迁移实验结果,第 6 列为本文方法的实验结果.

从图 6 中可以看出:

图 6　Facades 数据集实验结果

Fig. 6　Experimental results on Facades dataset

1)第 1 行,Pix2Pix 没有完整转换出语义图像的建筑屋顶信息,风格迁移的建筑图像屋顶有缺失;Cyclegan 的颜色不一致,建筑上方颜色偏淡,无法显示均匀一致的外墙颜色,处理的色彩不够真实;CUT 风格迁移的建筑图像发生了大量缺失,建筑的墙体四周残缺,不能很好地填充完整的建筑图像;BicycleGAN 对广告牌的转换能力欠佳,建筑下方的广告牌出现了黑影;本文方法能够完整提取语义图像信息,补全建筑整体外观,显示均匀一致的建筑外墙颜色,下方广告牌的转换没有严重形变和黑影出现,优于前 4 种风格迁移模型.

2)第 2 行,前 4 种模型风格迁移的建筑图像均出现了下方大门的形变,线条扭曲,外墙砖的显色模糊,本文方法风格迁移的细节比较前 4 种模型更好,大门的线条没有扭曲,外墙砖的纹理能够清晰显示.

3)第 3 行,CycleGAN 风格迁移的建筑图像中护栏和窗户兼容能力较差,有护栏的窗户均出现了形变;BicycleGAN 风格迁移的建筑图像下方出现了大量阴影,色彩偏暗;本文方法可以更好地显示建筑图像中窗户下方的护栏,色彩明亮,更接近真实建筑的颜色.

3.2.2　AerialPhoto&Map 数据集实验结果

AerialPhoto&Map 数据集上的实验结果如图 7 所示,目的是将输入语义图还原为真实航拍卫星图像.图 7 中第 1 列是输入的地图语义图像,第 2、第 3、第 4 和第 5 列分别为 Pix2Pix 模型、Cyclegan 模型、

图 7　AerialPhoto&Map 数据集实验结果

Fig. 7　Experimental results on AerialPhoto&Map dataset

CUT 模型和 BicycleGAN 模型的航拍卫星图像风格迁移实验结果,第 6 列为本文方法的实验结果.

从实验结果可以看出,前 4 种模型对水系的语义无法有效提取,本文方法对水系的提取能力较好.第 1 行:Pix2Pix 风格迁移的航拍卫星图像中将水系生成草坪,道路图像也失真,纹理不清晰;CycleGAN 对复杂建筑的航拍卫星图像转换效果较差,出现了移位和交错的模糊图像;本文方法更接近真实效果,道路图像纹理清晰.第 2 行:CUT 对林地和道路的色调信息提取较差,林地的颜色偏暗,风格迁移的质量不能满足人眼主观认知;BicycleGAN 风格迁移的航拍卫星图中右下方道路被草坪遮挡,道路提取效果较差;本文方法可以更好地展示边界位置,草坪没有遮挡道路.第 3 行:前 4 种模型方法风格迁移的航拍卫星图中建筑屋顶未能清晰显示,本文方法可以有效显示建筑屋顶,很好地提取了边界位置的道路.

3.3 评价指标

为了客观地反映不同模型的生成图像的效果,采用峰值信噪比(PSNR)和结构相似性(SSIM)指标来评价生成图像.这两个指标常用作图像处理的评价指标.两幅图像之间较高的 PSNR 值表示生成图像和原始图像之间失真较小,即生成图像质量较高.SSIM 反映了生成图像在亮度、对比度和结构方面与真实图像的相似性.SSIM 越接近 1,两幅图像之间的相似性越高,表明生成的图像更符合公众的视觉感知效果.

PSNR 是评价彩色图像质量的客观标准.计算公式如下:

$$\mathrm{PSNR}=10\log_{10}\frac{(2^n-1)^2}{\mathrm{MSE}},\tag{13}$$

$$\mathrm{MSE}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}[X(i,j)-Y(i,j)]^2,\tag{14}$$

其中,H 和 W 分别代表图像的宽度和高度,(i,j) 代表每个像素点,n 代表像素的位数,X 和 Y 分别代表两幅图像.由于 PSNR 指数也有其局限性,不能完全反映图像质量和人的视觉效果的一致性,所以采用 SSIM 指数做进一步的比较.SSIM 是一种度量两个图像相似性的标准.通过将模型绘制的图像与原始彩色图像进行比较,可以验证该算法的有效性和准确性.计算公式如下:

$$\mathrm{SSIM}=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2\mu_y^2+c_1)(\sigma_x^2\sigma_y^2+c_2)},\tag{15}$$

其中,μ_x 和 μ_y 分别表示真实图像和生成图像的平均值,σ_x^2 和 σ_y^2 分别表示真实图像和生成图像的方差,σ_{xy} 表示真实图像和生成图像的协方差,$c_1=(k_1,L)^2$ 和 $c_2=(k_2,L)^2$ 是保持稳定的常数,L 是像素值的动态范围,$k_1=0.01$,$k_2=0.03$. 在 Facades 数据集和 AerialPhoto&Map 数据集上的实验数据 PSNR 分数如表 4 所示,SSIM 分数如表 5 所示.

由表 4 和表 5 可以看出,本文方法的 PSNR 分数和 SSIM 分数均高于前 4 种方法,说明本文方法能够生成更丰富、更生动的图像内容.在 Facades 数据集上,本文方法的 PSNR 值比第二高的 BicycleGAN 模型高 2.103 9 dB,SSIM 分数比第二高的 CUT 模型提高了 0.031 7.在 AerialPhoto&Map 数据集上,本文方法的 PSNR 值比第二高的 CUT 模型高 2.135 1 dB,SSIM 分数比第二高的 CUT 模型提高了 0.102 8.CUT 模型通过引入对比学习能够专注于两个域之间共性的部分,但忽略两个域之间的差异性部分,使得图像轮廓不清晰,本文方法引入自注意力机制增强了远距离像素之间的连接,能够使得风格迁移的图像获得更清晰的边缘.BicycleGAN 在多目标任务中表现良好,但模型捕捉局部和整体之间的内部映射关系的能力较弱,本文通过引入感知损失,使得细节

表 4 PSNR 分数对比

Table 4 PSNR score comparison dB

数据集	Pix2Pix	Cyclegan	CUT	BicycleGAN	本文方法
Facades	12.135 6	12.196 7	13.135 6	13.659 8	15.763 7
AerialPhoto&Map	13.560 8	13.683 0	14.943 4	14.682 1	17.078 5

表 5 SSIM 分数对比

Table 5 SSIM score comparison

数据集	Pix2Pix	Cyclegan	CUT	BicycleGAN	本文方法
Facades	0.209 5	0.236 4	0.307 2	0.225 9	0.338 9
AerialPhoto&Map	0.232 7	0.240 7	0.332 2	0.269 7	0.435 0

方面的表现优于其他模型.从 PSNR 来看,本文方法生成的图像质量更高,和原始图像之间失真较小.从 SSIM 来看,本文方法生成的图像在亮度、对比度和结构方面与真实图像更相似.

3.4 消融实验

为了验证自注意力机制、谱归一化和感受损失在风格迁移效果上的有效性,本文设计了一组消融实验.本文方法是在原 BicycleGAN 的基础上加入了自注意力机制、谱归一化和感知损失改进而来.实验采用控制变量法进行,将自注意力机制、谱归一化和感知损失 3 种改进方案分别命名为 A、B、C,设计了 4 组实验进行对比.实验所生成的风格迁移图像如图 8 所示.

从图 8 可以看出,自注意力机制可以优化全局细节,无论是建筑图像的窗户线条还是航拍卫星图像的道路线条都能完整显示,提升图像质量.谱归一化和感知损失能够使得图像风格迁移更稳定的同时提升图像信息提取能力,原图像在航拍卫星图像的建筑屋顶出现了被草地覆盖使得显示不够清晰,改进方案使得图像的建筑屋顶与草地分明,细节处理更优.实验结果的 PSNR 分数对比客观指标如表 6 所示,SSIM 分数对比如表 7 所示.

由表 6 和表 7 客观指标可以看出,在 Facades 和 AerialPhoto&Map 数据集上,自注意力机制、谱归一化和感知损失的改进均有助于提升 PSNR 和 SSIM 分数,图像风格迁移效果有明显提升.说明添加的模

图 8 消融实验结果

Fig. 8 Results of ablation experiment

表 6 消融实验 PSNR 分数对比

Table 6 Comparison of PSNR scores in ablation experiment dB

数据集	BicycleGAN	BicycleGAN+A	BicycleGAN+B	BicycleGAN+C
Facades	13.659 8	15.281 0	14.964 5	14.297 0
AerialPhoto&Map	14.682 1	16.784 1	16.910 3	16.304 2

表 7 消融 SSIM 分数对比

Table 7 Comparison of SSIM scores in ablation experiment

数据集	BicycleGAN	BicycleGAN+A	BicycleGAN+B	BicycleGAN+C
Facades	0.225 9	0.310 4	0.298 6	0.305 8
AerialPhoto&Map	0.269 7	0.349 0	0.362 7	0.387 2

块在提高图像风格迁移的质量和保真度方面是有效的,采用本文方法生成的图像更加真实,且风格迁移的图像细节更加丰富.

4 结论

传统的 BicycleGAN 网络模型在图像风格迁移过程中图像细节不清晰,训练不稳定,有时会出现梯度爆炸的现象.本文对 BicycleGAN 进行了改进,在生成器中引入残差块,改善模型训练的退化现象,利用自注意力机制获得更多的图像特征,使得生成图像更接近真实图像.在判别器中引入谱归一化,提高训练稳定性,提升判别能力.同时引入感知损失,提升了图像生成质量.实验结果表明,本文方法与传统的风格迁移模型 Pix2Pix、Cyclegan、CUT、BicycleGAN 相比,图像生成质量和视觉效果有较大提高,PSNR 分数和 SSIM 分数有较大提升.

参考文献

References

[1] Goodfellow I,Pouget-Abadie J,Mirza M,et al.Generative adversarial networks[J].Communications of the ACM,2020,63(11):139-144

[2] Isola P,Zhu J Y,Zhou T H,et al.Image-to-image translation with conditional adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).July 21-26,2017,Honolulu,HI.IEEE,2017:1125-1134

[3] Zhu J Y,Park T,Isola P,et al.Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV).October 22-29,2017,Venice.IEEE,2017:2223-2232

[4] Zhu J Y,Zhang R,Pathak D,et al.Toward multimodal image-to-image translation[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.December 4-9,2017,Long Beach,California,USA.New York:ACM,2017:465-476

[5] Larsen A B L,Sønderby S K,Winther O.Autoencoding beyond pixels using a learned similarity metric[J].arXiv e-print,2015,arXiv:1512.09300

[6] Dumoulin V,Belghazi I,Poole B,et al.Adversarially learned inference [J].arXiv e-print,2016,arXiv:1606.00704

[7] Park T,Efros A A,Zhang R,et al.Contrastive learning for unpaired image-to-image translation[M]//Computer Vision—ECCV 2020.Cham:Springer International Publishing,2020:319-345

[8] He K M,Zhang X Y,Ren S Q,et al.Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).June 27-30,2016,Las Vegas,NV,USA.IEEE,2016:770-778

[9] Miyato T,Kataoka T,Koyama M,et al.Spectral normalization for generative adversarial networks [J].arXiv e-print,2018,arXiv:1802.05957

[10] Ulyanov D,Vedaldi A,Lempitsky V.Instance normalization:the missing ingredient for fast stylization[J].arXiv e-print,2016,arXiv:1607.08022

[11] Wang T C,Liu M Y,Zhu J Y,et al.High-resolution image synthesis and semantic manipulation with conditional GANs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18-23,2018,Salt Lake City,UT,USA.IEEE,2018:8798-8807

[12] Johnson J,Alahi A,Li F F.Perceptual losses for real-time style transfer and super-resolution[M]//Computer Vision—ECCV 2016.Cham:Springer International Publishing,2016:694-711

[13] Wang T,Wu L,Sun C Y.A coarse-to-fine approach for dynamic-to-static image translation[J].Pattern Recognition,2022,123:108373

[14] Yang X,Zhao J Y,Wei Z Y,et al.SAR-to-optical image translation based on improved CGAN[J].Pattern Recognition,2022,121:108208

[15] 黄菲,高飞,朱静洁,等.基于生成对抗网络的异质人脸图像合成:进展与挑战[J].南京信息工程大学学报(自然科学版),2019,11(6):660-681
HUANG Fei,GAO Fei,ZHU Jingjie,et al.Heterogeneous face synthesis via generative adversarial networks:progresses and challenges[J].Journal of Nanjing University of Information Science & Technology (Natural Science Edition),2019,11(6):660-681

[16] Llull P,Yuan X,Carin L,et al.Image translation for single-shot focal tomography[J].Optica,2015,2(9):822-825

[17] Armanious K,Jiang C M,Fischer M,et al.MedGAN:medical image translation using GANs[J].Computerized Medical Imaging and Graphics,2020,79:101684

[18] Gao N,Xue H,Shao W,et al.Generative adversarial networks for spatio-temporal data:a survey[J].ACM Transactions on Intelligent Systems and Technology,2022,13(2):1-25

[19] 张颖涛,张杰,张睿,等.全局信息引导的真实图像风格迁移[J].计算机科学,2022,49(7):100-105
ZHANG Yingtao,ZHANG Jie,ZHANG Rui,et al.Photo-realistic style transfer guided by global information[J].Computer Science,2022,49(7):100-105

[20] Gatys L A,Ecker A S,Bethge M,et al.Controlling perceptual factors in neural style transfer[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).July 21-26,2017.Honolulu,HI.IEEE,2017:3985-3993

南京信息工程大学学报(自然科学版),2023,15(5):514-523
Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2023,15(5):514-523

Image style transfer based on generative adversarial network

LIU Hang[1] LI Ming[1] LI Li[2] FU Denghao[3] XU Changli[1]

1 College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China

2 College of Computer and Information Science, Southwest University, Chongqing 400715, China

3 School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract Generative Adversarial Network (GAN) can generate images that are close to real images, thus plays an important role in image style transfer. However, the GAN-based image transfer is perplexed by problems of low quality of generated images and difficult training of models, herein a new style transfer approach based on BicycleGAN model is proposed. First, the residual module is introduced into the generator of GAN to solve the degradation of GAN in training, and the self-attention mechanism is employed to obtain more image features thus improve the generation quality of the generator. To solve the gradient explosion in the training of GAN, the spectral normalization is added behind each convolution layer of the discriminator. Then the perceptual loss is introduced to address the unstable training and low generated image quality. The experiments on Facades and AerialPhoto&Map datasets show that the proposed approach outperforms other image style transfer methods in the PSNR and SSIM values of the generated images.

Key words generative adversarial network (GAN); style transfer; self-attention mechanism; spectral normalization; perceptual loss