基于 CycleGAN 的人脸素描图像生成

徐志鹏,卢官明,罗燕晴
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:CycleGAN 是一种基于生成对抗网络的衍生模型,可以在缺少成对训练图像的条件下实现两个具有不同风格的图像域之间的相互转换。由于收集大量成对的人脸图像和素描图像存在较大的难度,并且针对人脸素描图像生成任务中存在的图像细节模糊和低清晰度的问题,提出一种改进的 CycleGAN 模型。通过引入基于注意力机制的残差模块,让 CycleGAN 的生成器模型可以更加有效地学习不同通道特征和人脸图像中不同区域的重要程度,降低人脸图像中无用信息对生成模型的影响,从而提升生成的人脸素描图像的质量。通过对比实验发现,使用基于注意力机制的 CycleGAN 模型生成的素描人脸图像质量较好,且更完整清晰地保留了较丰富的面部特征信息,优于 CycleGAN 和 DualGAN 模型,充分证明了基于注意力机制的改进 CycleGAN 模型的有效性。

关键词:CycleGAN;生成对抗网络;风格转换;人脸素描;注意力机制;残差模块
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2021)08-0063-06
doi:10.3969/j.issn.1673-629X.2021.08.011

Face Sketch Image Generation Based on CycleGAN

XU Zhi-peng,LU Guan-ming,LUO Yan-qing
(School of Telecommunication & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:CycleGAN is a derivative model based on generative adversarial network,which can realize the mutual conversion between two image domains with different styles in the absence of paired training images. Because it is difficult to collect a large number of pairs of face images and sketch images,in order to solve the problem of blurred image details and low definition in the task of generating face sketch images,an improved CycleGAN model is proposed. By introducing the residual module based on the attention mechanism,the CycleGAN generator model can learn the importance of different channel features and different regions in the face image more effectively,reducing the impact of useless information in the face image on the generation model,thereby improving the quality of the generated face sketch image. Through comparative experiments,it is found that the sketched face image generated by the CycleGAN model based on the attention mechanism is of better quality,and retains more complete and richer facial feature information,which is better than the CycleGAN and DualGAN models. It is fully proved that the CycleGAN model based on attention mechanism is effective.

Key words:CycleGAN;generative adversarial networks;style transfer;face sketch;attention mechanism;residual block

0 引 言

图像风格转换是指在保留原图像内容信息的基础上生成具有新风格的图像的一种技术,在艺术创作和社交娱乐等方面有着潜在的应用前景,受到学术界和工业界的高度关注。Gatys L A 等人在 2016 年提出了一种基于卷积神经网络的风格转换方法[1],通过预训练模型 VGG-19[2]对图像的内容特征和风格特征进行剥离,实现图像风格转换。通过实验,Gatys 等人发现卷积神经网络可以实现图像内容和风格的分离,图像风格转换可以取得较好的效果,但是生成图像的过程非常耗时,并且训练好的生成模型无法应用在其他风格转换的任务上,推广应用受限制。Goodfellow I 等人开创性地提出生成对抗网络[3](generative adversarial networks,GAN),对图像风格转换领域有着重大的意义,相继出现了基于 GAN 的风格转换模型,主要包括 Pix2Pix[4]、CycleGAN[5]和 StarGAN[6]等模型。其中,Pix2Pix 和 CycleGAN 模型适用于两个不同风格图像域之间的转换,而 StarGAN 模型则可以实现多个图像域之间的风格转换。Pix2Pix 使用 U-Net[7]模型,有效地保留不同尺度的特征信息,提升生成图像的细节效

收稿日期:2020-10-26 **修回日期**:2021-02-26
基金项目:江苏省研究生科研与实践创新计划(SJCX19_0245)
作者简介:徐志鹏(1996-),男,硕士研究生,研究方向为深度学习;卢官明,博士,教授,研究方向为计算机视觉。

果,适合应用于特定的图像风格转换任务。CycleGAN 模型通过添加循环一致性损失函数,成功地解决了缺少成对的训练图像的问题。StarGAN 模型解决了多个图像域间的风格转换的难题,只需要训练一个生成器模型就可以实现多个图像域间的风格转换。

由于人脸图像具有较多的细节信息,采用原 CycleGAN 模型很难很好地处理人脸图像的细节信息,导致生成图像的视觉效果较差。文中针对人脸素描图像风格转换任务,在 CycleGAN 的基础上,通过改进生成器的网络结构,更好地保留人脸图像的细节信息,生成高质量的图像。实验结果表明,使用改进 CycleGAN 模型可以得到更高质量的图像,验证了该方法的有效性。

1 相关理论

1.1 GAN

GAN 是由生成器 G(generator)和鉴别器 D(discriminator)共同构成的深度学习模型,生成器 G 负责学习训练图像集的概率分布规律并生成具有相似概率分布规律的图像;鉴别器 D 负责判别输入图像是生成的图像还是训练图像。通过让生成器 G 和鉴别器 D 进行对抗训练,使生成器 G 生成的图像具有与训练图像相似的风格,鉴别器 D 判别生成的图像和训练图像的能力也得到不断提高,最终使得生成器 G 和鉴别器 D 达到一种稳定平衡状态,又称纳什均衡。GAN 的网络结构如图 1 所示。

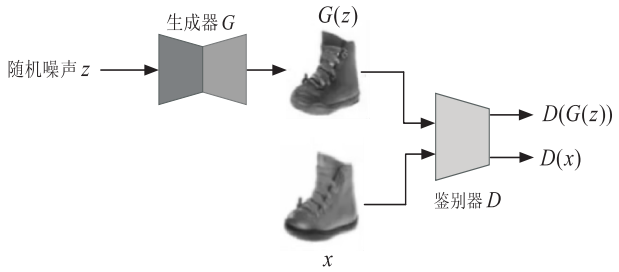


图 1 GAN 网络结构

随机噪声 z 是生成器 G 的输入,x 是训练图像,$G(z)$ 表示生成图像,$D(G(z))$ 表示鉴别器 D 判定生成的图像 $G(z)$ 是训练图像的概率,$D(x)$ 表示鉴别器 D 判定图像 x 是训练图像的概率。

目前,GAN 越来越受到学术界和工业界的重视,许多基于 GAN 的衍生模型已经被广泛应用于图像风格转换[8]、超分辨率[9]、图像修复[10,11]等领域,并不断向着其他领域继续延伸,具有广阔的发展前景[12]。

1.2 CycleGAN

CycleGAN 是由 Zhu J Y 等人提出的风格转换模型,该模型包含两个生成器和两个鉴别器,通过引入循环一致性损失函数,可以在缺少成对训练图像的条件下实现两个不同风格的图像域之间的转换。CycleGAN 模型结构如图 2 所示。

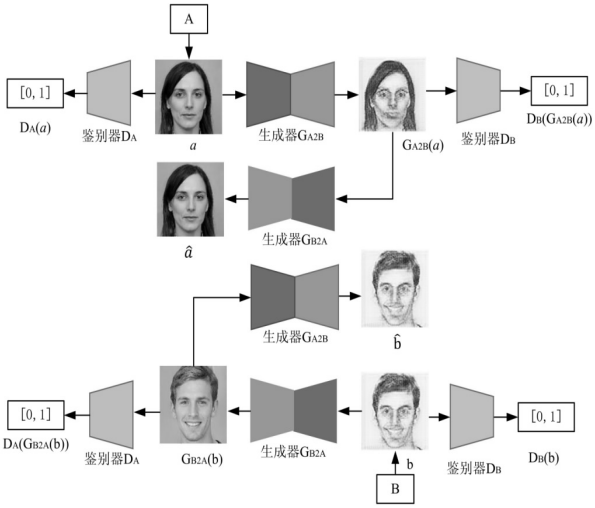


图 2 CycleGAN 模型结构

图 2 中的 A 和 B 分别表示两个不同风格的图像域,a 表示 A 图像域中的一幅训练图像,b 表示 B 图像域中的一幅训练图像。CycleGAN 模型先将图像 a 输入到生成器 G_{A2B},得到具有 B 图像域风格的生成图像 $G_{A2B}(a)$,再将生成图像 $G_{A2B}(a)$ 输入到生成器 G_{B2A},得到重建图像 $\hat{a}$。同理,可以得到具有 A 图像域风格的生成图像 $G_{B2A}(b)$ 以及重建图像 $\hat{b}$。鉴别器 D_A 计算出图像 a 和生成图像 $G_{B2A}(b)$ 属于 A 图像域的概率 $D_A(a)$ 和 $D_A(G_{B2A}(b))$,鉴别器 D_B 负责计算出图像 b 和生成图像 $G_{A2B}(a)$ 属于 B 图像域的概率 $D_B(b)$ 和 $D_B(G_{A2B}(a))$。鉴别器 D_A 和 D_B 的计算概率用于定义 CycleGAN 的对抗性损失,重建图像 $\hat{a}$ 和 $\hat{b}$ 用于构建 CycleGAN 的循环一致性损失。

1.2.1 损失函数

CycleGAN 的损失函数是由对抗性损失和循环一致性损失两部分共同组成。

CycleGAN 模型拥有两个生成器和两个鉴别器,分别实现 A 图像域→ B 图像域的风格转换和 B 图像域→ A 图像域的风格转换,所以 CycleGAN 的对抗性损失将由两部分构成,将 A 图像域→ B 图像域的风格转换的对抗性损失记作 L_{A2B},B 图像域→ A 图像域的风格转换的对抗性损失记作 L_{B2A}。

$$L_{A2B} = E_{b\sim p_{\text{data}(b)}}[\log D_B(b)] + E_{a\sim p_{\text{data}(a)}}[\log(1 - D_B(G_{A2B}(a)))] \tag{1}$$

$$L_{B2A} = E_{a\sim p_{\text{data}(a)}}[\log D_A(a)] + E_{b\sim p_{\text{data}(b)}}[\log(1 - D_A(G_{B2A}(b)))] \tag{2}$$

式中,$p_{\text{data}(a)}$ 和 $p_{\text{data}(b)}$ 分别表示 A 图像域的概率分布和 B 图像域的概率分布。

CycleGAN 的对抗性损失 L_{GAN} 如下:

$$L_{\text{GAN}} = L_{A2B} + L_{B2A} \tag{3}$$

LSGAN[13]证明采用最小二乘损失函数可以加速模型收敛速度,提高生成图像的质量,因此在 CycleGAN 的实际训练中,将其对抗性损失 L_{GAN} 中的对数运算优化成平方运算:

$$L_{A2B} = E_{b\sim p_{\text{data}(b)}}[D_B(b)]^2 + E_{a\sim p_{\text{data}(a)}}[1 - D_B(G_{A2B}(a))]^2 \tag{4}$$

$$L_{B2A} = E_{a\sim p_{\text{data}(a)}}[D_A(a)]^2 + E_{b\sim p_{\text{data}(b)}}[1 - D_A(G_{B2A}(b))]^2 \tag{5}$$

通过引入对抗性损失函数,使得鉴别器 D_A 无法区分生成图像 $G_{A2B}(a)$ 和 B 图像域的概率分布,鉴别器 D_B 无法区分生成图像 $G_{B2A}(b)$ 和 A 图像域的概率分布。

引入循环一致性损失函数 L_{Cycle},通过不断地迭代训练,使得 a 与 $\hat{a}$ 越来越相似,b 与 $\hat{b}$ 越来越相似,从而避免 CycleGAN 模型把所有源图像域中图像都转换为目标图像域中的同一幅图像,导致对抗性损失无效的问题出现。

$$L_{\text{Cycle}} = E_{a\sim p_{\text{data}(a)}}[\|\hat{a} - a\|_1] + E_{b\sim p_{\text{data}(b)}}[\|\hat{b} - b\|_1] \tag{6}$$

CycleGAN 总的损失函数 L_{CycleGAN} 如下:

$$L_{\text{CycleGAN}} = L_{\text{GAN}} + \lambda \times L_{\text{Cycle}} \tag{7}$$

式(7)中,参数 λ 为循环一致性损失的权重,控制着抗性损失和循环一致性损失的相对重要性。

1.2.2　生成器和鉴别器的网络结构

CycleGAN 的生成器采用编码-解码框架,编码器由两个步长为 2 的卷积层构成,解码器由两个步长为 $\frac{1}{2}$ 的反卷积层构成,通过在编码器和解码器之间添加几个残差块(residual block)[14],实现图像风格转换。在卷积运算和反卷运算之后都会进行实例正则化(instance normalization)[15],有效地降低网络训练的难度。然后将归一化的特征图进行 ReLU 激活,增加网络的稀疏性,防止梯度弥散。CycleGAN 的生成器网络结构如图 3 所示。

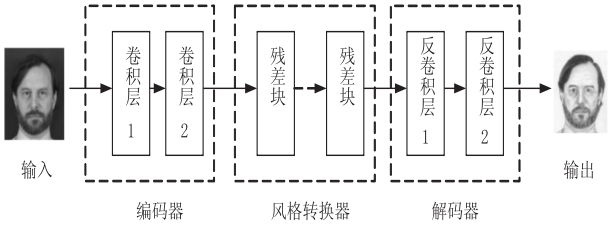


图 3　生成器网络结构

CycleGAN 的生成器网络采用残差块结构,通过在深层网络上添加一条直连路径,确保了梯度信息能够有效地在深层网络中进行传递,成功地解决深层网络中存在的梯度消失问题,改善深层网络的性能。反卷积(deconvolution)[16]又被称为转置卷积,是卷积的逆运算,用于图像生成。

CycleGAN 的鉴别器使用 PatchGAN[4]结构,其网络结构如图 4 所示。

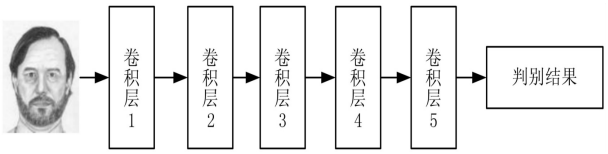


图 4　鉴别器网络结构

不同于普通鉴别器,PatchGAN 的输出是一个 $N * N$ 的矩阵,矩阵中的每一个元素值代表着鉴别器对输入图像中的每一个 patch 的判别结果,再将矩阵的均值作为整幅图像的最终判别结果。这种结构的鉴别器具有更少的参数,可以缩短训练时长,并且适用于任意尺寸的图像,有效地捕捉图像局部的高频特征,使生成的图像保持高分辨率和高细节。

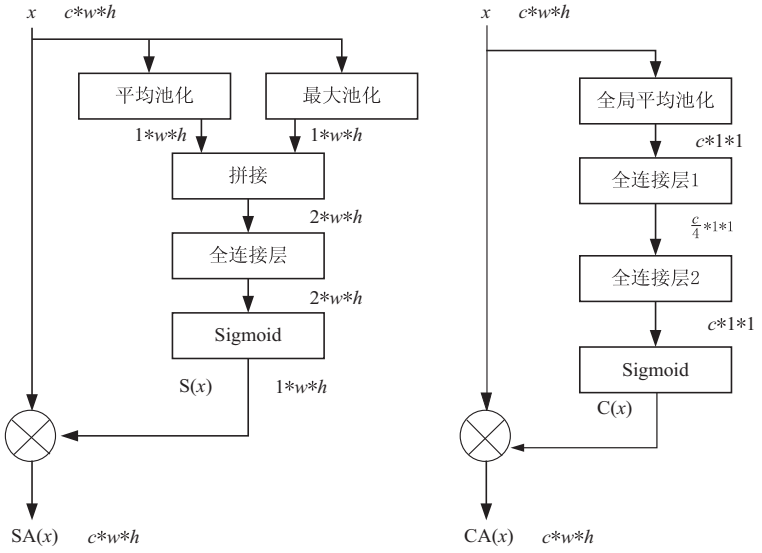
2　基于注意力机制的 CycleGAN

注意力机制(attention mechanism,AM)是一种改进神经网络的方法,在近些年得到迅速发展,出现了许多基于注意力机制的深度学习网络,极大地丰富了神经网络的表示能力。注意力机制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。注意力机制得到的权重可以作用在原始图上[17-18],也可以作用在特征图上[19]。目前,注意力机制已经在图像分类[20-21]和图像分割[22]等计算机视觉任务中取得较好的效果。

从注意力域的角度来分析,可以将注意力机制分为三类:空间域(spatial domain)、通道域(channel domain)和混合域(mixed domain)。空间域注意力机制的代表是 spatial transformer networks(STN)[23]。STN 通过图像进行空间变换,提取出关键信息,降低图像中无用信息对模型训练的干扰,从而提升网络的性能。空间域注意力机制主要适用于对输入图像的处理。通道域注意力机制的代表是 squeeze-and-excitation networks(SENet)[24]。SENet 能够计算出特征图的每一个特征通道的权重值,并实现特征通道的权重分配,从而增强重要特征对当前任务所起的作用。将空间域注意力机制和通道域注意力机制进行组合,即混合域注意力机制,可以对特征图中每个元素同时实现空间域和通道域的注意力机制。

文中提出的改进 CycleGAN 模型主要是将空间域和通道域注意力机制用于生成器网络中,所采用的空间域和通道域注意力机制的结构如图 5 所示。

假设空间域和通道域注意力机制输入的特征图的尺寸均为 $c * w * h$,中间各层的运算结果如表 1 和表 2 所示。



(a)空间域 (b)通道域

图 5 注意力机制的网络结构

表 1 空间域注意力机制的各层输出结果

运算方法	输出特征图的尺寸
平均池化	1 * w * h
最大池化	1 * w * h
拼接	2 * w * h
全连接层 1	1 * w * h
Sigmoid	1 * w * h

表 2 通道域注意力机制的各层输出结果

运算方法	输出特征图的尺寸
全局平均池化	c * 1 * 1
全连接层 1	$\frac{c}{4}$ * 1 * 1
全连接层 2	c * 1 * 1
Sigmoid	c * 1 * 1

空间注意力机制通过拼接运算实现对特征图的平均池化和最大池化运算结果的融合,再经过全连接层 1 和 Sigmoid 函数得到归一化的权重;通道域注意力机制先将特征图进行全局平均池化运算,借鉴 SENet,将全连接层 1 的卷积核个数设置为 $\frac{c}{4}$,再经过全连接层 2 和 Sigmoid 函数得到归一化的权重。

基于注意力机制的残差块(AM 残差块)模型结构如图 6 所示。AM 残差块模型结合残差模块和注意力机制的优点,既可以缓解在深度神经网络中增加网络深度带来的梯度消失问题,又可以在只需要增加较少的计算量的情况下减少无用信息对模型的干扰,提升网络的表现力,改善生成图像的质量。

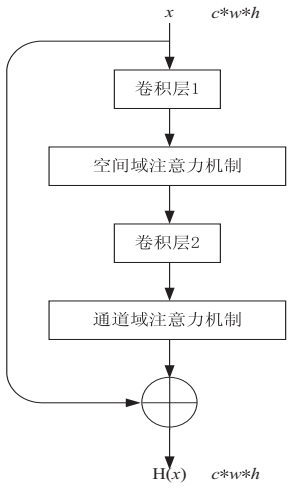


图 6 AM 残差块

基于注意力机制的 CycleGAN 生成器网络结构如图 7 所示。

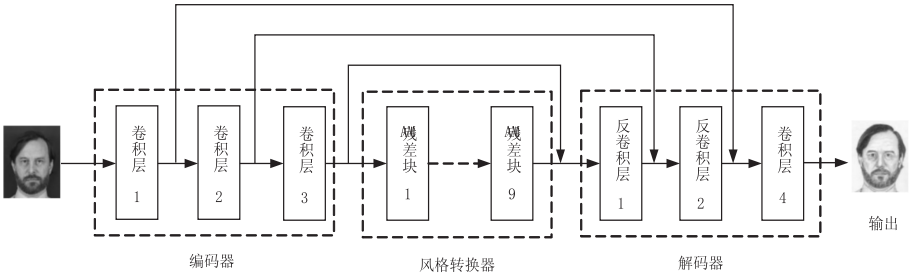


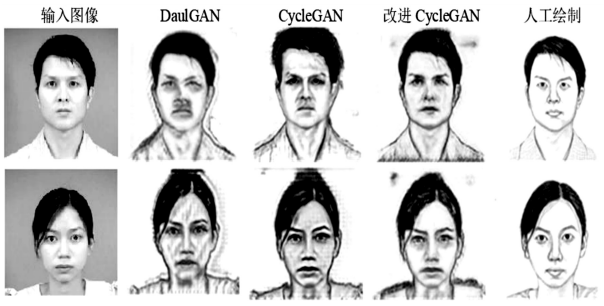
图 7　基于注意力机制的 CycleGAN 生成器网络结构

文中在 CycleGAN 的生成器网络中添加跳跃连接机制,实现将编码器中卷积运算得到的特征图传递到解码器中,使得解码器可以学习更多不同尺度的特征信息,改善生成的人脸素描图像质量。文中提出的基于注意力机制的 CycleGAN 的鉴别器采用图 4 所示的 PatchGAN 网络结构。

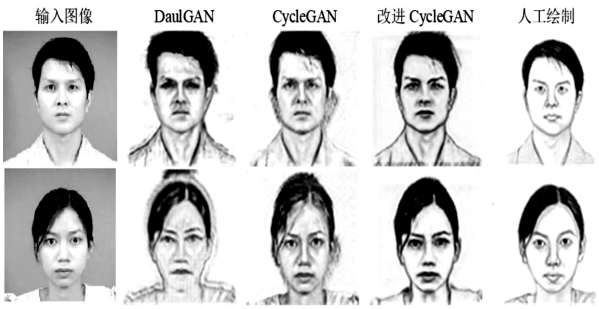
3　实验与结果分析

实验的硬件平台为 Intel Xeon CPU E5-2650 v4,使用 NVIDIA GTX 1080 GPU 进行加速处理。实验选取从网络上收集到的 300 张人脸图像和 CUFSF 数据集[25-26]中 1 194 张素描人脸图像作为训练数据集;选取 CUFS 数据集[26]中 88 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。优化器采用性能较好的 Adam 算法,参数 beta1 和 beta2 分别设置为 0.5 和 0.999,学习率 lr 设置为 0.002。

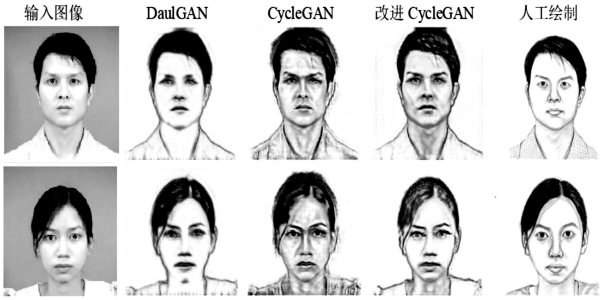
如图 8 所示,当网络模型进行 40 个 epoch 迭代训练之后,已经能够实现通过人脸彩色照片生成黑白人脸图像,初具素描风格,但是生成图像的边缘比较模糊;当进行 60 个 epoch 迭代训练之后,生成图像边缘逐渐清晰,具有较明显的素描风格;当进行 80 个 epoch 迭代训练之后,可以生成较为逼真的人脸素描图像。通过对比发现,文中提出的改进 CycleGAN 模型生成的图像比 CycleGAN 和 DualGAN 生成的图像更加清晰,在图像边缘处处理地更好,更好地保留了人脸五官特征和表情等有效信息。



(a) epoch = 40



(b) epoch = 60



(c) epoch = 80

图 8　实验结果

4　结束语

当今网络社交媒体拥有着巨大的用户量,如果可以发布一些关于图像风格转换的手机端应用程序,让用户充分发挥艺术创造力,设计属于自己的特有风格的作品,这会让图像风格转换技术走进人们日常生活。但是,现阶段的图像风格转换领域仍然存在一些问题。首先,目前主流的基于深度学习的图像风格转换算法,存在模型的参数数量过多,训练耗时较长。其次,很难将需要进行风格转换的部分从原图中分割出来,无法实现局部图像风格化,导致一些生成图像质量较低。

文中主要对 CycleGAN 的生成器模型进行改进,将空间域和通道域注意力机制用于生成器网络中,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。

参考文献:

[1] 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, NV, USA: IEEE, 2016: 2414-2423.
[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2020-10-18]. https://arxiv.org/abs/1409.1556.
[3] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc., 2014: 2672-2680.
[4] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 1125-1134.
[5] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on 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. Las Vegas, NV, USA: IEEE, 2018: 8789-8797.
[7] RONNEBERGER O, FISCHER P, BROX T, et al. U-Net: convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention. Munich, Germany: Springer International Publishing, 2015: 234-241.
[8] ZHANG H, XU T, LI H, et al. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. Venice, Italy: IEEE, 2017: 5907-5915.
[9] WANG X, YU K, WU S, et al. ESRGAN: enhanced super-resolution generative adversarial networks[C]//Computer vision - ECCV 2018 workshops. Munich, Germany: Springer International Publishing, 2018: 63-79.
[10] 强振平,何丽波,陈 旭,等. 深度学习图像修复方法综述[J]. 中国图象图形学报,2019,24(3):447-463.
[11] KUPYN O, BUDZAN V, MYKHAILYCH M, et al. Deblurgan: blind motion deblurring using conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE, 2018: 8183-8192.
[12] 王坤峰,苟 超,段艳杰,等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报,2017,43(3):321-332.
[13] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. Venice, Italy: IEEE, 2017: 2794-2802.
[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.
[15] ULYANOV D, VEDALDI A, LEMPITSKY V. Instance normalization: themissing ingredient for fast stylization[EB/OL]. (2016-07-27)[2020-10-18]. https://arxiv.org/abs/1607.08022.
[16] NOH H, HONG S, HAN B, et al. Learning deconvolution network for semantic segmentation[C]//International conference on computer vision. Santiago, Chile: IEEE, 2015: 1520-1528.
[17] BA J, MNIH V, KAVUKCUOGLU K. Multiple object recognition with visual attention[EB/OL]. (2014-12-24)[2020-10-18]. https://arxiv.org/abs/1412.7755.
[18] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]//Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc., 2014: 2204-2212.
[19] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[C]//International conference on machine learning. Red Hook, NY, USA: Curran Associates, Inc., 2015: 2048-2057.
[20] FU J, ZHENG H, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 4438-4446.
[21] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 3156-3164.
[22] CHEN L C, YANG Y, WANG J, et al. Attention to scale: scale-aware semantic image segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 3640-3649.
[23] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc., 2015: 2017-2025.
[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2018: 7132-7141.
[25] ZHANG W, WANG X, TANG X. Coupled information-theoretic encoding for face photo-sketch recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2011: 513-520.
[26] WANG X, TANG X. Face photo-sketch synthesis and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(11): 1955-1967.