ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 1,Jan. 2021,pp. 22－34
DOI:10. 16337/j. 1004-9037. 2021. 01. 002
© 2021 by Journal of Data Acquisition and Processing
http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

RD-GAN:一种结合残差密集网络的高清动漫人脸生成方法

叶继华，刘　凯，祝锦泰，江爱文

(江西师范大学计算机信息工程学院,南昌 330022)

摘　要: 随着动漫产业的快速发展,动漫人脸的生成成为一项关键技术。由于动漫人脸具有的高度简化和抽象的独特风格以及倾向于具有清晰的边缘、平滑的阴影和相对简单的纹理,现有方法中的损失函数面临很大的挑战,同时绘画的风格迁移技术无法获得满意的动漫结果。因此,本文提出了一种新颖的适用于动漫图像的损失函数,该函数的语义损失表示为VGG网络高级特征图中的正则化形式,以应对真实图像和动漫图像之间不同的风格,具有的边缘增强的边缘清晰损失可以保留动漫图像的边缘清晰度。4个公开数据集上的实验表明:通过本文提出的损失函数可以生成清晰生动的动漫人脸图像;在CK＋数据集中,本文方法相比于现有的方法识别率提高了0.43%(宫崎骏风格)和3.29%(新海诚风格);在RAF数据集中,本文方法识别率提高了0.85%(宫崎骏风格)和2.42%(新海诚风格);在SFEW数据集中,本文方法识别率提高了0.71%(宫崎骏风格)和3.14%(新海诚风格);在Celeba数据集中也显示了本文方法优异的生成效果。实验结果说明本文方法结合了深度学习模型的优点,使检测结果更加准确。

关键词: 动漫;人脸生成;风格迁移;残差密集网络;生成对抗网络

中图分类号: TP391.41　　**文献标志码:** A

RD-GAN: A High Definition Animation Face Generation Method Combined with Residual Dense Network

YE Jihua, LIU Kai, ZHU Jintai, JIANG Aiwen

(School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022,China)

Abstract: With the rapid development of the animation industry, the face generation with animation characters becomes a key technology. The existing style transfer technology of painting style cannot obtain satisfactory animation results due to the following characteristics of animation: (1) Animation has a highly simplified and abstract unique style, and (2) animation tends to have clear edges and smooth shadows and relatively simple textures, which poses great challenges to the loss function in existing methods. This paper proposes a novel loss function suitable for animation. In the loss function, the semantic loss is expressed as a regularized form in the high-level feature map of the VGG network to deal with the different styles between real and animation images, and the edge sharpness loss with edge enhancement can preserve the edge sharpness of animation images. Experiments on the four public data sets show that through the proposed loss function, clear and vivid animation images can be generated. Moreover, in the CK＋ data

基金项目: 国家自然科学基金(61462042, 61966018)资助项目。
收稿日期: 2020-10-27;**修订日期:** 2021-01-11

set, the recognition rate of the proposed method is increased by 0.43% (Miyazaki Hayao style) and 3.29% (Makoto Shinkai style) compared with the existing method, increased by 0.85% (Miyazaki Hayao style) and 2.42% (Makoto Shinkai style) in the RAF data set, and increased by 0.71% (Miyazaki Hayao style) and 3.14% (Makoto Shinkai style) in the SFEW data set, respectively. The generation effect in the Celeba data set is also demonstrated. The above results show that the proposed method combines the advantages of the deep learning model to make the detection result more accurate.

Key words: anime; face generation; style transfer; residual dense network; generative adversarial network

引　言

人类的很多情感大多数借助于面部进行传递,随着深度学习和人机交互的发展,从人脸图像中了解人的情感变得越来越重要。为了实现更加精准的人脸表情识别,研究人员已经进行了大量的研究,但是目前的人脸表情识别方法大多是对真实人脸图像进行识别,如果将这些模型应用到动漫人物的面部表情识别当中则很难有好的效果。风格迁移是解决此类问题的常用技术,传统方法是为特定的风格开发专用算法,但需要付出巨大的努力才能产生细腻的特定风格。近年来生成对抗网络[1]得到了快速的发展,它以循环方式设计的生成对抗网络[2]可以实现高质量的风格迁移,它最突出的特征是使用未配对的图片和风格化的图片训练模型。

因此,本文提出了基于生成对抗网络的有效的损失函数。在生成器网络中,为了应对图像和动漫之间的风格变化,在VGG网络的高级特征图中引入语义损失,定义为L_1稀疏正则化。同时本文提出具有边缘增强的边缘清晰损失,可以保留动漫图像的边缘清晰度,生成具有清晰边缘的动漫图像。最后将风格迁移应用到人脸生成当中,可以增加对动漫人物表情识别的精度,并且在CK+、SFEW、Celeba和RAF公开数据集的实验中也有很好的效果。

1　相关工作

本文提出的方法主要涉及人脸表情识别、生成对抗网络以及风格迁移技术,因此本节主要讨论这3个方面的研究工作。

1.1　生成对抗网络

生成对抗网络是2014年由Goodfellow等提出来的一种生成模型[1]。传统的生成对抗网络最大的缺点是训练过程非常不稳定,导致生成的图像没有好的效果。为了解决这个问题,其中的一种方法是深度卷积生成对抗网络[3],采用全卷积神经网络。Ariovsky等[4]提出的Wasserstein生成对抗网络解决了训练不稳定的问题。Gulrajani等[5]提出在Wasserstein生成对抗网络的改进训练中使用梯度惩罚。Lipschitz限制是要求判别器的梯度不超过K,梯度惩罚则是设置一个额外的损失项来实现梯度与K之间的联系,这就是梯度惩罚的核心所在。近年来,虽然基于条件生成对抗网络相关研究工作很多,但是很难直接用于动漫人脸的生成。

1.2　人脸表情识别

人脸的面部表情是人类传达感情的重要方式,随着人脸处理技术的不断完善,利用计算机进行人脸表情分析成为了可能。深度学习的不断发展促进了人脸表情识别技术的快速进步,例如,叶继华等[6]对表情识别进行了概述;Vemulapalli等[7]提出了一个贴近人的视觉偏好的简约空间来描述面部表情;Li等[8]提出了一种Twin-Cycle autoencoder来学习头部姿势的相关和面部运动单元相关的运动信息。

1.3 图像风格迁移

现有很多算法可以用来模仿动漫的艺术风格,其中一些算法用简单的阴影渲染生成类似于动漫的效果,这种着色技术可以为艺术家节省大量的时间,并已用于创建游戏以及视频和电影。Kotovenko等[9]提出了关注图像的内容感知和样式感知的样式化,而且提出了一个规划层来提高风格化后的分辨率。Zhu等[10]将直接式的姿态迁移替换为渐进式的姿态迁移,提出了级联的姿态迁移模型。Cho等[11]提出解码之前,在网络结构上把风格编码加进内容编码中以使得风格迁移效果更好。Lin等[12]提出了多属性语义级人脸编辑,可以对人脸的多个部位进行编辑。

图像到图像的转换问题通常被表示为按像素分类或回归,这个想法可以追溯到Image analogies[13],在单个输入输出训练图像对上采用了非参数纹理模型[14]。最近的方法使用输入输出样例的数据集来学习卷积神经网络(Convolutional neural network, CNN)[15]的参数转换功能。跨模态场景网络[16]使用权重贡献策略来学习跨域的通用表示。无监督图像到图像的转换网络[17],通过变分自动编码器[18]和生成对抗网络的组合扩展了框架。这些方法还使用对抗网络,并带有附加项以在预定的度量空间(例如类标签空间[19],图像像素空间[20]和图像特征空间[21])中使输出接近输入。

2 高清动漫人脸生成对抗网络

本文设计生成器和鉴别器网络以适应动漫图像的特殊性。设计的学习过程将实现真实人脸图像转换成动漫图像,作为映射功能将真实图像域R映射到动漫域C。通过训练数据$T_{p(r)}=\{r_i|i=1,\cdots,M\}\subset R$和$T_{p(c)}=\{c_i|i=1,\cdots,N\}\subset C$来学习映射函数,其中$M$和$N$为训练数据集中真实图像和动漫图像的数量,和其他的生成对抗网络框架一样,训练判别器D通过将动漫域与真实图像域区分开并且为生成器G提供对抗损失来推动生成器,使其可以达到生成动漫人脸的目的。令$L(D,G)$为损失函数,G^*和D^*为生成器G和判别器D的网络权重,c为动漫图像,r为真实图像。本文的目的是解决最小-最大问题,即

$$\begin{aligned}&(G^*,D^*)=\arg\min_G\max_D L(G,D)\\&\min_G\max_D L(D,G)=\log D(c)+\log(1-D(G(r)))\end{aligned}\tag{1}$$

2.1 网络框架

本文的网络架构如图1所示。生成器G用于将输入的人脸图像映射到动漫域,当模型训练好之后就会生成动漫风格。输入图像经过下采样对图像进行空间压缩和编码,在此阶段提取有用的局部信息以进行后续的转换。当损失函数从判别器开始反向传播到生成器时,实际上经过了很多层,由于越深的网络参数越多,在反向传播的过程中也越容易梯度弥散,因此本文采用残差密集网络[22]保证了梯度信息能够有效传递以增强生成对抗网络的鲁棒性。最后通过上采样重构输出动漫图像。与生成器G互补的是判别器D,用于判断输入图像是否为动漫图像。由于该判断是一项要求较低的任务,而不是常规的全图像判别器,因此本文判别器中使用带有很少参数的简单的patch-level判别器。与目标分类不同,判断动漫风格依赖图像的局部特征,所以判别器D被设计得较浅。在经过平滑层之后,判别器使用了两个步长卷积块以降低分辨率和编码重要的局部特征用于分类。然后,使用特征构造块和3×3卷积层获得分类。

2.2 损失函数

损失函数$L(G,D)$在式(2)中包含两个部分:(1)对抗损失$L_{\mathrm{adv}}(G,D)$使生成器网络实现目标域的转换;(2)语义损失$L_{\mathrm{sem}}(G,D)$保留图像风格化后的语义内容。本文损失函数表达式为

$$\begin{aligned}&L(G,D)=L_{\mathrm{adv}}(G,D)+\mu L_{\mathrm{sem}}(G,D)\\&L_{\mathrm{adv}}(G,D)=L_{\mathrm{a}}(G,D)+L_{\mathrm{cl}}(G,D)\\&L_{\mathrm{sem}}(G,D)=\|\mathrm{VGG}(x_1)-\mathrm{VGG}(x_2)\|_1\end{aligned}\tag{2}$$

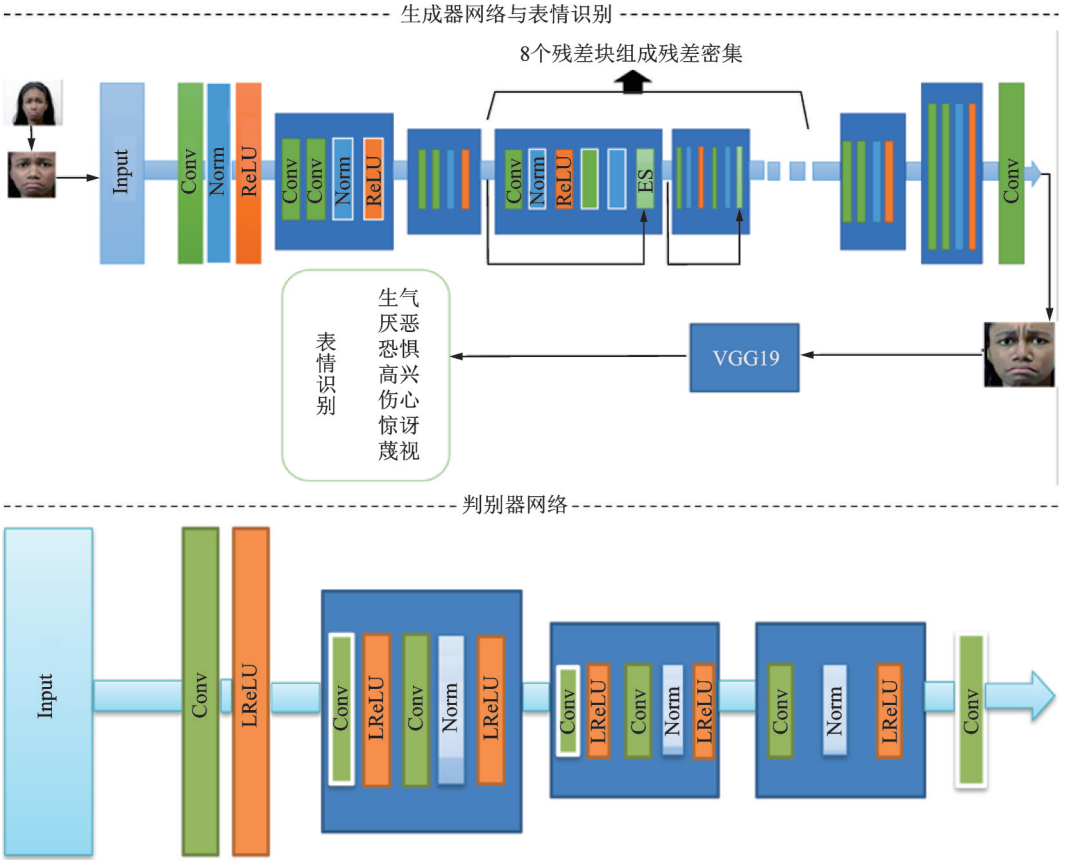


图1 生成对抗网络框架

Fig.1 Generative adversarial net framework

式中:μ平衡两个损失函数,G,D分别表示生成器和判别器。μ越大将保留更多来自输入图像的语义信息,因此具有更详细纹理的风格化图像。其中生成对抗损失由对抗损失$L_a(G,D)$和边缘清晰损失$L_{cl}(G,D)$组成,语义损失$L_{sem}(G,D)$由VGG损失组成。在实验中设置$\mu=10$使风格和语义之间达到良好的平衡,不同μ的生成效果如图2所示。

(a) $\mu=5$ (b) $\mu=8$ (c) $\mu=10$ (d) $\mu=13$ (e) $\mu=15$

图2 不同μ的生成效果

Fig.2 Generation effects of different μ

2.2.1 对抗损失函数

对抗损失同时应用到生成器网络G和判别器网络D,这会影响生成器网络G动漫风格化的过程。对抗损失$L_a(G,D)$的值是生成器网络G的输出图像在多大程度看起来像动漫图像。

$$L_{\mathrm{a}}(G,D)=E_{c_i\sim T_{p(c)}}[\log D(c_i)]+E_{r_k\sim T_{p(r)}}[\log(1-D(G(r_k)))] \tag{3}$$

式中：$T_{p(c)}$表示动漫域，$c_i\in T_{p(c)}$表示动漫图像；$T_{p(r)}$表示真实图像域，$r_k\in T_{p(r)}$表示真实图像。对于真实图像域R中的每个图像r_k，生成器网络G输出的图像为$G(r_k)$，判别器网络D的目标是最大程度将$G(r_k)$判定为正确的可能性。通过将输入图像输入到生成器当中，生成器生成的动漫化图像与真实图像放到判别器当中，判别器判定的参数传递到生成器，优化生成器的生成过程使其可以生成目标的动漫风格。但是在动漫化的过程中，需要使动漫图像具有清晰的边缘。

2.2.2 边缘清晰损失函数

在以前的生成对抗网络框架中，判别器网络D的任务是区分输入图像是来自生成器生成的图像还是真实的图像。但是对于本研究来说，仅仅训练判别器网络D区分生成图像和真实的动漫图像不足以将真实图像风格化为动漫图像，这是因为清晰的边缘呈现是动漫图像的重要特征，但是这些边缘通常在整个图像中比例很小。因此，没有重现清晰的边缘但是有正确纹理的输出图像会使判别器受到标准对抗损失的影响。为了解决这个问题，本文从训练的动漫图像$T_{p(c)}\in C$中，通过删除$T_{p(c)}$中的清晰边缘，自动生成一组图像$T_{p(a)}=\{a_j|j=1,\cdots,N\}\subset Q$，其中$C$和$Q$分别为动漫域和没有清晰边缘的动漫域。对于每个动漫图像$c_j\in T_{p(c)}$，使用标准Canny边缘检测器[23]检测边缘像素，使用高斯平滑扩张边缘区域。$a_j\in T_{p(a)}$为没有清晰边缘的动漫图像，模型的边缘清晰损失函数$L_{\mathrm{cl}}(G,D)$如下

$$L_{\mathrm{cl}}(G,D)=E_{a_j\sim T_{p(a)}}[\log(1-D(a_j))] \tag{4}$$

从图3、4可以看出在有无边缘清晰损失时生成的动漫图像效果的差别。图3、图4分别是生成宫崎骏风格及新海诚风格的动漫图像。图3(a)及图4(a)是在有边缘清晰损失的情况下生成的动漫图像，具有清晰的边缘；图3(b)及图4(b)是在没有边缘清晰损失的情况下生成的动漫图像，可以看出动漫图像是模糊的，边缘不清晰，说明本文提出的边缘清晰损失对生成清晰动漫图像具有重要的作用。本文总的生成对抗损失函数$L_{\mathrm{adv}}(G,D)$由上述两部分组成，即

(a) With loss of edge sharpness (b) Without loss of edge sharpness

图3 有无边缘清晰损失的生成动漫图像(宫崎骏风格)

Fig.3 Generate animation images with and without loss of edge sharpness (Miyazaki Hayao style)

(a) With loss of sharp edges (b) Without loss of sharp edges

图4 有无边缘清晰损失的生成动漫图像(新海诚风格)

Fig.4 Generate animation images with and without loss of sharp edges (Makoto Shinkai style)

$$L_{\mathrm{adv}}(G,D)=E_{c_i\sim T_{p(c)}}[\log D(c_i)]+E_{a_j\sim T_{p(a)}}[\log(1-D(a_j))]+E_{r_k\sim T_{p(r)}}[\log(1-D(G(r_k)))] \quad (5)$$

判别器可以通过将输入转换为正确域来引导生成器网络G。在生成具有清晰边缘的动漫图像时,需要保证图像的语义特征不能改变。

2.2.3 语义损失函数

除了在正确域之间进行风格迁移,动漫风格化中的另一个重要目标是确保生成的动漫图像保留输入图像的语义内容。本文将语义损失$L_{\mathrm{sem}}(G,D)$定义为

$$L_{\mathrm{sem}}(G,D)=E_{r_i\sim T_{p(r)}}[\|\mathrm{VGG}_l(G(r_i))-\mathrm{VGG}_l(r_i)\|_1] \quad (6)$$

式中:l为特定VGG层的特征图,$r_i\in T_{p(r)}$表示真实图像。本文使用真实图像和生成的动漫图像之间VGG特征图的L_1稀疏正则化来定义语义损失。图5(a)是输入图像,图5(b)是在没有语义损失的情况下生成的图像,由于缺少语义损失,模型无法生成与输入图像语义相同的图像,并且可以保证在进行风格化之后图像的前后表情保持一致,这表明模型的语义损失在保留输入图像语义内容时有重要的作用。

(a) Input images (b) Generated images

图5 无语义损失生成的图像

Fig.5 Generated images without semantic loss

由于生成对抗网络模型具有非线性的具有随机初始化的功能,因此可以很容易地在次优局部最小值捕获优化。为了提高收敛性,本文开始一个新的初始化,生成器网络的目标是保留语义内容的同时用动漫风格重构输入图像。模型从生成器网络开始对抗学习,生成器仅重建输入图像的语义内容。本文使用语义损失$L_{\mathrm{sem}}(G,D)$来训练生成器网络。实验结果表明,初始化可以帮助本文模型快速收敛而不会过早收敛。文献[24]有类似的过程,它使用内容图像初始化结果图像来提高风格迁移的质量。

3 实验结果与分析

3.1 人脸表情识别结果分析

本文根据图1结构在pytorch环境下进行实验。首先通过“face_recognition”方法进行人脸检测,可以对人脸图像进行裁剪。表情识别的实验数据集使用CK+[25],RAF[26],SFEW[27]和Celeba。CK+数据集包含表情和动作单元的标记,以及123个目标图像和593个图像序列。每个图像序列的最后一帧都有动作单元的标记,而593个图像序列中有327个序列有表情标记。RAF数据集包含了接近3万张带有表情分布的图片。根据对应的表情概率分布向量,将数据集划分为7类基本表情。SFEW数据集从AFEW中开发,涵盖了不受约束的面部表情、各种头部姿势以及接近真实世界的光照。SFEW包含700张图像,使用独立的方式分为两组:第一组有346张图像,为训练数据集;第二组有354张图像,为测试数据集。Celeba是一个大规模的人脸属性数据集,拥有超过20万张名人图像,每个图像有40个标注属性。Celeba数据集有很大的多样性、较大的数量和丰富的注释,包括10 177个实体和202 599张人脸图像,

实验中为了获得一组具有相同样式的动漫图像,本文使用由同一位艺术家绘制和导演的动漫电影

的关键帧作为训练数据，在实验中使用3 617张“千与千寻”的卡通图片训练宫崎骏的模型，使用4 573张“你的名字”的卡通图片来训练新海诚的模型。验证集使用3 011张动画图像来验证本文实验结果，本文模型(宫崎骏/新海诚)的测试集使用的是经过模型风格迁移后的图像。

表1给出了VGG16、Resnet18、文献[28]以及本文模型在CK＋数据集上的人脸表情识别率，比较动漫场景的实现效果。文献[28]网络最终任务是进行表情识别，同样使用了GAN作为生成模型，并且在训练时考虑了图像内容，具有较高的表情识别率，但是该网络无法生成具有边缘清晰的图像。而本文网络可以生成清晰的图像，这有助于提高识别率，在训练时可以有效地避免梯度弥散现象，并且可以更好地适用于动漫场景下的表情识别。

表1 CK+数据集上动漫图像不同表情的识别结果

Table 1 Expression recognition results of anime images with different expressions on CK+ dataset %

识别模型	愤怒	厌恶	恐惧	高兴	伤心	惊讶	蔑视	平均
VGG16	88	90.0	70.0	90	70	83.0	71.0	80.28
Resnet18	90	93.0	70.0	93	73	85.0	74.0	82.57
文献[28]模型	95	92.5	72.5	95	80	92.5	80.0	86.78
本文模型(宫崎骏)	95	92.0	73.0	96	82	92.0	80.5	87.21
本文模型(新海诚)	97	93.5	78.5	98	86	91.5	86.0	90.07

从表1可见，本文模型均获得了较高的识别率，其中新海诚风格的动漫更加接近现实，所以识别结果更好。这是因为本文提出的边缘清晰损失可以使生成器生成具有清晰边缘的动漫图像，增加了识别的精度。从表1还可以看出：在7种表情中“高兴”的表情识别结果最高，其次是“愤怒”，因为这两种表情的面部特征最容易区分，不容易与其他表情混淆在一起；所有表情中识别结果最低的是“恐惧”，和“高兴”与“愤怒”表情相反，“恐惧”的表情容易被分类成其他表情。图6展示了本文模型在CK＋数据集上表情识别结果的混淆矩阵。由图6可以看出，“恐惧”表情识别率最低，主要是因为其与“伤心”“惊讶”表情之间存在较高的混淆率，有8.5%的“恐惧”表情被识别为“伤心”的表情，7.85%被识别为“惊讶”的表情。

图6 本文模型(新海诚风格)在CK＋数据集中的表情识别结果混淆矩阵

Fig.6 Confusion matrix of facial expression recognition results of CK＋ dataset by the proposed model (Makoto Shinkai style)

表2给出了VGG16、Resnet18、文献[28]以及本文模型在RAF数据集上的人脸表情识别结果。从

表2 RAF数据集上动漫图像不同表情的识别结果

Table 2 Expression recognition results of anime images with different expressions on RAF dataset %

识别模型	愤怒	厌恶	恐惧	高兴	伤心	惊讶	中立	平均
VGG16	86.0	88.0	70.0	92.0	76.5	85.0	60.5	79.71
Resnet18	91.0	90.0	71.0	93.0	77.0	86.0	63.0	81.57
文献[28]模型	95.0	92.5	72.5	95.0	80.0	92.5	67.5	85.00
本文模型(宫崎骏)	94.5	94.0	73.0	95.0	81.0	94.0	69.5	85.85
本文模型(新海诚)	97.0	94.0	78.0	95.5	82.0	93.5	72.0	87.42

表2中可以得出,在7种表情中“愤怒”的表情识别率最高,其次是“高兴”表情,因为这两种表情的面部特征最容易区分,不容易与其他的表情混淆在一起;而所有表情中识别率最低的是“中立”,因为“中立”的表情中面部肌肉变化最小。图7展示了本文模型在RAF数据集上表情识别结果的混淆矩阵。由图7可以看出,“中立”表情识别率最低主要是因为其与“伤心”“恐惧”表情之间存在较高的混淆率,有12%的“恐惧”表情被识别为“伤心”,7%被识别为“恐惧”。

表3为VGG16、Resnet18、文献[28]以及本文模型在SFEW数据集上的人脸表情识别率,比较在动漫场景下的实现效果。从表3可见,新海诚风格的动漫更加接近现实,所以识别结果更高;在7种表情中“愤怒”的表情识别率最高,其次是“高兴”,因为这两种表情的面部特征最容易区分,不容易与其他表情混淆在一起;所有表情中识别结果最低的是“中立”,因为“中立”表情中面部肌肉变化最小。图8展示了本文模型在SFEW数据集上的表情识别结果混淆矩阵。由图8可以看出,“中立”表情识别结果最低主要是因为其与“伤心”“厌恶”“恐惧”表情之间存在较高的混淆率,有5%的“恐惧”表情被识别为“伤心”,14%被识别为“恐惧”,5.5%被识别为“厌恶”。

	愤怒	厌恶	恐惧	高兴	伤心	惊讶	中立
愤怒	97.00	0.00	0.50	0.00	4.90	0.00	0.30
厌恶	1.50	94.00	0.30	4.50	0.00	0.00	1.90
恐惧	0.00	0.00	78.00	0.00	6.20	6.00	9.10
高兴	0.00	2.20	0.00	95.50	0.00	0.30	2.00
伤心	1.70	0.00	7.20	0.00	82.00	0.00	9.10
惊讶	0.00	0.00	5.50	0.52	0.00	93.50	2.48
中立	0.63	4.23	12.00	0.43	7.00	3.30	72.00

图7 本文模型(新海诚风格)在RAF数据集中的表情识别结果混淆矩阵

Fig.7 Confusion matrix of facial expression recognition results of RAF dataset by the proposed model (Makoto Shinkai style)

	愤怒	厌恶	恐惧	高兴	伤心	惊讶	中立
愤怒	97.50	0.00	0.53	0.00	1.50	0.00	0.47
厌恶	1.10	94.00	0.90	2.64	0.00	0.00	1.33
恐惧	0.20	0.30	77.00	0.00	7.50	6.00	9.100
高兴	0.00	1.20	0.00	96.00	0.00	0.00	2.80
伤心	3.20	0.00	5.80	0.00	87.00	0.00	6.00
惊讶	0.00	0.00	4.50	0.50	0.00	94.00	1.00
中立	1.00	5.50	14.00	1.50	5.00	1.50	71.50

图8 本文模型(新海诚风格)在SFEW数据集中的表情识别结果混淆矩阵

Fig.8 Confusion matrix of facial expression recognition results of SFEW dataset by the proposed model (Makoto Shinkai style)

表3 SFEW数据集上动漫图像不同表情的识别结果

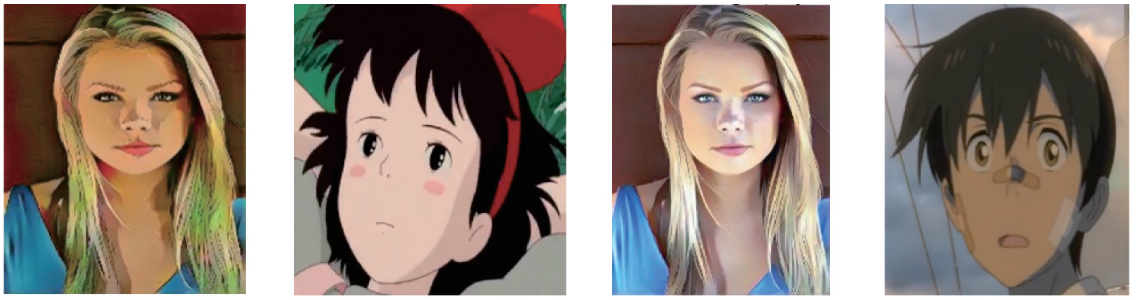
Table 3 Expression recognition results of anime images with different expressions on SFEW dataset %

识别模型	愤怒	厌恶	恐惧	高兴	伤心	惊讶	中立	平均
VGG16	85.5	88.0	70.5	75.5	86.0	84.0	62.0	78.78
Resnet18	92.0	90.5	71.0	93.0	77.5	87.0	63.5	82.07
文献[28]模型	95.0	92.5	72.5	95.0	80.0	92.5	67.5	85.00
本文模型(宫崎骏)	96.0	91.5	72.5	96.5	82.5	92.0	69.0	85.71
本文模型(新海诚)	97.5	94.0	77.0	96.0	87.0	94.0	71.5	88.14

3.2 风格迁移图像生成结果

为了验证本文模型在图像风格迁移方面的效果,本文在Celeba数据集中选取老人、成人和佩戴眼镜的人的人脸图像的生成效果,其中每种选择生成了宫崎骏风格和新海诚风格的动漫人脸图像,使用CycleGAN训练生成新海诚风格的图像。

图9展示了模型的动漫化效果。其中图9(a, c)分别是本文模型生成的宫崎骏风格和新海诚风格的图像,图9(b, d)分别是动漫域的宫崎骏电影图像和新海诚电影图像。从图9可以看出,本文模型可以很好地生成与之对应的风格图像,同时宫崎骏风格的图像在色彩上更加鲜艳,新海诚风格的图像色彩更加偏于适中,这也显示出本文模型良好的生成能力。

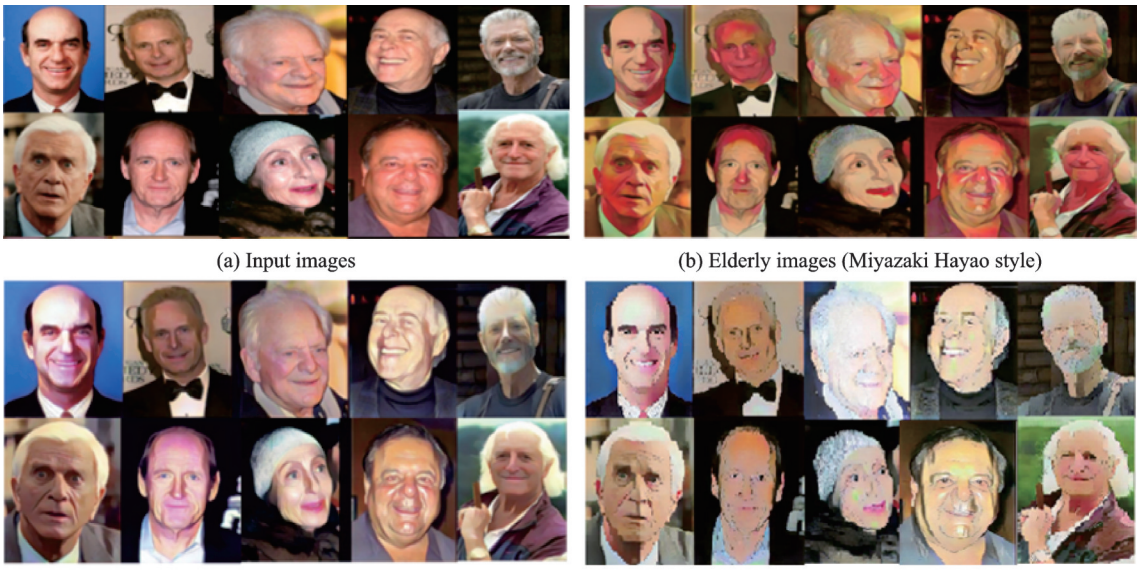


(a) Generated image (Miyazaki Hayao style) (b) Miyazaki Hayao movie image (c) Generated image (Makoto Shinkai style) (d) Makoto Shinkai movie image

图9 风格图像与真实动漫域图像

Fig.9 Style images and real anime domain images

图10展示了本文模型和CycleGAN生成老人图像的效果。其中图10(a)是输入的人脸图像,图10(b)是生成的宫崎骏风格的动漫图像,图10(c)是生成的新海诚风格的动漫图像,图10(d)是CycleGAN生成的图像。从图10可以看出,生成的动漫图像具有较清晰的边缘,这是由于本文的对抗损失函数使用了3个部分判断动漫域、没有清晰边缘的动漫域以及来自生成器生成的图像,其中一部分是判别器判别没有清晰边缘的动漫图像,这样可以推动生成器网络生成具有清晰边缘的动漫图像。



(a) Input images (b) Elderly images (Miyazaki Hayao style)

(c) Elderly images (Makoto Shinkai style) (d) Elderly images generated by CycleGAN

图10 Celeba数据集中生成的老人图像效果

Fig.10 Generated effect of elderly images in the Celeba dataset

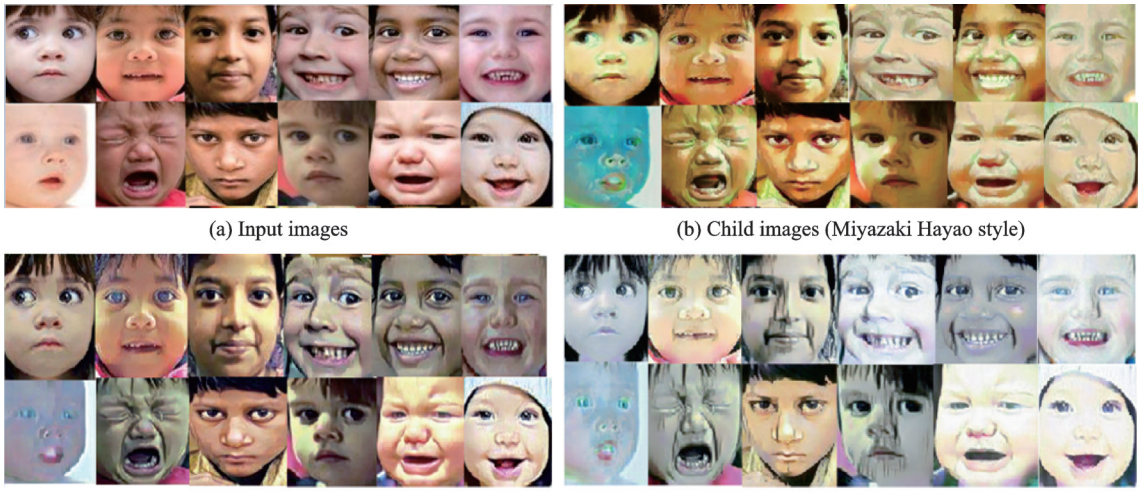
图11展示了本文模型和CycleGAN生成的成年人动漫图像。其中图11(a)是输入图像,图11(b,c)分别是生成的宫崎骏风格和新海诚风格的动漫图像,图11(d)是CycleGAN生成的动漫图像。从图11(a,b)可以看出,输入图像中都有光照的阴影部分,本文生成的动漫图像同样保留了光照的阴影,这是由于本文的语义损失函数可以保留输入图像中的语义内容,即使是光照部分同样会保留下来。不过本文的生成效果也存在缺点,例如图中第5列的"乔布斯"图像在通过风格迁移动漫化之后,生成宫崎骏风格的动漫图像时额头位置部分变成了黑色,这是由于宫崎骏的风格化偏向于使图像颜色更加鲜艳,输入图像的背景和人物的衣服都是黑色,风格化之后额头也变成黑色。

(a) Input images
(b) Adult images (Miyazaki Hayao style)
(c) Adult images (Makoto Shinkai style)
(d) Adult images generated by CycleGAN

图11 Celeba数据集中生成的成年人图像效果

Fig.11 Generated effect of adult images in the Celeba dataset

图12展现了本文模型和CycleGAN在RAF数据集上小孩动漫图像的生成效果。其中图12(a)是输入图像,图12(b)是生成的宫崎骏风格的动漫图像,图12(c)是生成新海诚风格的动漫图像,图12(d)是CycleGAN生成的动漫图像,图像没有清晰的边缘。从图12可以看出,本文模型对于小孩的表情有着很好的风格迁移效果,对哭着的小孩进行风格迁移可以保留眼泪,说明本文模型可以很好地保留语义特征;对于图12(a)第2行第1个输入图像,无论生成宫崎骏风格或者新海诚风格的动漫图像,效果都比较差,这是由于输入图像的小孩皮肤比较白,背景也是白色,这使得本文模型进行风格迁移时人脸颜色产生错误的风格化。



(a) Input images (b) Child images (Miyazaki Hayao style)

(c) Child images (Makoto Shinkai style) (d) Child images generated by CycleGAN

图12 RAF数据集中生成的小孩图像效果

Fig.12 Generated effect of child images in the RAF dataset

4 结束语

本文提出了一种基于生成对抗网络以及风格迁移的人脸生成方法，在生成对抗网络中将真实人脸进行风格迁移生成动漫风格的人脸，然后通过卷积神经网络对不同表情的动漫人脸进行表情识别。在4个公开的数据集上的实验表明，本文模型可以提升动漫人脸表情识别率，并且具有很好的生成动漫人脸的效果。

在未来的工作中，作者将考虑结合动漫人物身体各个部位的姿态提高表情识别精度，然后通过对动漫人脸的表情识别生成具有目标表情的动漫人脸，可以进一步推动人工智能技术自动生成视频动漫，为动漫产业节约成本。

参考文献：

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the Conference and Workshop on Advances in Neural Information Processing Systems. [S.l.]: ACM, 2014: 2672-2680.

[2] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 17453078.

[3] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. (2015-11-19)[2020-12-20]. http:// arxiv.org/1511.06434.pdf.

[4] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[EB/OL]. (2017-01-27)[2020-12-20]. http://arxiv.org/abs/1701.07875.pdf.

[5] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]//Proceedings of the International Conference on Neural Information Processing Systems.[S.l.]: [s.n.], 2017: 5767-5777.

[6] 叶继华，祝锦泰，江爱文，等. 人脸表情识别综述[J]. 数据采集与处理，2020，35(1): 21-34.
YE Jihua, ZHU Jintai, JIANG Aiwen , et al. Facial expression recognition: A survey[J]. Journal of Data Acquisition and Processing, 2020, 35(1): 21-34.

[7] VEMULAPALLI R, AGARWALA A. A compact embedding for facial expression similarity[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5676-5685.

[8] LI Y, ZENG J, SHAN S, et al. Self-supervised representation learning from videos for facial action unit detection[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 10916-10925.

[9] KOTOVENKO D, SANAKOYEU A, MA P, et al. A content transformation block for image style transfer[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 10024-10033.

[10] ZHU Z, HUANG T, SHI B, et al. Progressive pose attention transfer for person image generation[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 2342-2351.

[11] CHO W, CHOI S, PARK D K, et al. Image-to-image translation via group-wise deep whitening-and-coloring transformation [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 10631-10639.

[12] LIN Y, WU P, CHANG C, et al. RelGAN: Multi-domain image-to-image translation via relative attributes[C]// Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 5913-5921.

[13] HERTZMANN A, JACOBS C E, OLIVER N, et al. Image analogies[C]//Proceeding of ACM SIGGRAPH. Los Angeles: ACM, 2001: 327-340.

[14] EFROS A A, LEUNG T K. Texture synthesis by non-parametric sampling[C]// Proceedings of the Seventh IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 1999: 1033-1038.

[15] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA: IEEE, 2015: 3431-3440.

[16] AYTAR Y, CASTREJON L, VONDRICK C, et al. Cross-modal scene networks[C]// Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence. [S,l.]: IEEE, 2016: 2303-2314.

[17] LI Shan, DENG Weihong, DU JunPing. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2584-2593.

[18] JHA A H, ANAND S, SINGH M, et al. Disentangling factors of variation with cycle-consistent variation auto-encoders[C]// Proceedings of European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer, 2018: 829-845.

[19] ROMERO A, ARBELÁEZ P, VAN GOOL L , et al. SMIT: Stochastic multi-label image-to-image translation[C]// Proceedings of IEEE/CVF International Conference on Computer Vision Workshop. Seoul, South Korea: IEEE, 2019: 3285-3294.

[20] BOUSMALIS K, SILBERMAN N, DOHAN D, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 95-104.

[21] TAIGMAN Y, POLYAK A, WOLF L. Unsupervised cross-domain image generation[EB/OL]. (2017-02-10)[2020-12-20]. https://arxiv.org/abs/1611.02200.pdf.

[22] YE Z, LYU F, LI L, et al. SR-GAN: Semantic rectifying generative adversarial network for zero-shot learning[C]// Proceedings of IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019: 85-90.

[23] CANNY J. A computational approach to edge detection[C]// Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence. [S.l.]: IEEE, 1986: 679-698.

[24] GATYS L, ECKER A, BETHGE M. Image style transfer using convolutional neural networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 2414-2423.

[25] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, CA, USA: IEEE, 2010: 94-101.

[26] MA Liqian, XU Jia, STAMATIOS G, et al. Exemplar guided unsupervised image-to-image translation with semantic consistency[C]// Proceedings of International Conference on Learning Representations (ICLR). New Orleans, LA, USA: ICLR, 2019: 882-895.

[27] DHALL A, GOECKE R, LUCEY S, et al. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark[C]//Proceedings of 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona: IEEE, 2011: 2106-2112.

[28] 张飞飞,张天柱,毛启容,等. 基于生成对抗网络的多姿态人脸表情识别[EB/OL]. (2019-12-05)[2020-12-20]. http://kns.cnki.net/kcms/detail/11.1826.TP.20191205.1151.002.html.
ZHANG Feifei, ZHANG Tianzhu, MAO Qirong, et al. Multi-pose facial expression recognition based on generative confrontation network[EB/OL]. (2019-12-05)[2020-12-20]. http://kns.cnki.net/kcms/detail/11.1826.TP.20191205.1151.002.html.

作者简介:

叶继华(1966-),**通信作者**,男,教授,博士生导师,研究方向:普适计算、物联网技术、数据融合、图像处理等, E-mail: yjhwcl@163.com。

刘凯(1993-),男,硕士研究生,研究方向:智能信息处理, E-mail:13455022996@163.com。

祝锦泰(1994-),男,硕士研究生,研究方向:图像处理、表情识别, E-mail: 2545000505@qq.com。

江爱文(1984-),男,博士,副教授,研究方向:智能信息处理、图像处理等, E-mail:flix_2001@163.com。

(编辑:张黄群)