文章编号：2095-2163(2022)04-0121-06 中图分类号：TP317.4 文献标志码：A

基于自注意力机制和谱归一化的 GAN 表情合成

苏梦晶[1]，王 波[2,3]，刘本永[1]
（1 贵州大学 大数据与信息工程学院，贵阳 550025；2 贵州大学 计算机科学与技术学院，贵阳 550025；
3 贵阳学院 数学与信息科学学院，贵阳 550005）

摘 要：为实现更具真实感的表情图像合成，探讨一种基于自注意力机制和谱归一化的生成式对抗网络（GAN）表情合成方法。通过在生成器中引入 2 层自注意力模块，使生成器能够在局部建立丰富的上下文关系，输出更加真实的表情细节；同时，在鉴别器中引入谱归一化，使鉴别器的训练更加稳定。实验结果表明，该模型在主观视觉和 FID 图像评价指标上均优于其他典型算法，图像质量和表情细节有明显提高。

关键词：表情合成；生成对抗网络；自注意力机制；谱归一化

Expression image synthesis using GAN by introducing self-attention mechanism and spectral normalization

SU Mengjing[1], WANG Bo[2,3], LIU Benyong[1]
（1 College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China;
2 College of Computer Science and Technology, Guizhou University, Guiyang 550025, China;
3 College of Mathematics and Information Science, Guiyang University, Guiyang 550005, China）

【Abstract】 In order to synthesize a more realistic expression image, a method introducing self-attention mechanism and spectral normalization into generative adversarial network (GAN) is proposed. By introducing two layers of self-attention modules in the generator of a GAN, the generator can locally establish a rich contextual relationship and output more realistic expression details; at the same time, spectral normalization is introduced in the discriminator of the GAN to make the training of the discriminator more stable. Experimental results show that the proposed method outperforms other typical algorithms with respected to objective vision and Frechet Inception Distance score image evaluation indicators, and the image quality and expression details are significantly improved.

【Key words】 expression image synthesis; generative adversarial networks; self-attention mechanism; spectral normalization

0 引 言

面部表情合成指在改变特定对象的面部表情的同时保留该对象的身份信息和面部特征。近年来，表情合成技术在电影特效、计算机动画[1]、交互界面[2]、面部手术规划[3]等方面得到广泛应用。此外，表情合成也可用于扩展表情识别训练数据，进一步提升识别性能[4]。由于人脸面部结构较为复杂，合成具有真实感的面部表情仍是一个难题。

目前，实现表情合成的方法主要分为传统方法和深度学习方法。传统方法可分为基于映射的方法和基于建模学习的方法。前者利用不同表情的面部特征矢量差实现表情变化，仅考虑了图像像素的差异，细节处理的能力较弱。如 2001 年，Liu 等人[5]提出的基于表情比率图的合成方法能较好地合成表情的细节，但对于光线、背景、图像质量等因素鲁棒性不足；2004 年，姜大龙等人[6]提出的基于局部表情比率图的合成方法选择额头、嘴角等具有表情细节的区域进行合成计算，但仅考虑了图像之间的像素差；2013 年，王晓慧等人[7]提出的基于小波图像融合的合成方法擅长提取纹理特征，但提取时会产生非表情特征，合成结果不佳。

基于建模学习的方法主要通过对人脸表情建立特定的模型进行表情合成，考虑的脸部变形因素较多，但搭建的模型较为复杂，难以在实际中运用。如 1982 年，Parke[8]提出的参数化模型设计了一套用

基金项目：国家自然科学基金（60862003）。

作者简介：苏梦晶（1997-），女，硕士研究生，主要研究方向：图像处理、模式识别；王 波（1979-），男，博士研究生，讲师，主要研究方向：机器学习、图像处理；刘本永（1966-），男，博士，教授，博士生导师，主要研究方向：机器学习、图像处理、模式识别。
通讯作者：刘本永 Email：byliu667200@163.com

收稿日期：2021-11-07

于人脸表情控制的参数。1996 年,Lei 等人[9]提出的肌肉模型以人脸的解剖学为基础,考虑了脸部肌肉活动的影响。2004 年,Abboud 等人[10]提出的主动外观模型方法能合成不同强度和类型的表情图像。

近年来,深度学习方法在计算机视觉、图像处理、计算机图形学等领域发展迅速,尤其是 2014 年 Goodfellow 等人[11]提出的生成对抗网络(generative adversarial network, GAN),极大地提高了合成图像的质量。该网络及其改进(pix2pix[12]、CycleGAN[13]、AttGAN[14]、StarGAN[15])在人脸合成问题上取得了很大成功,这些方法都能改变人脸的面部表情,但在细节上还有待改进。

为了改善表情合成图像中的细节,本文探讨一种基于自注意力机制和谱归一化的生成对抗网络表情合成方法,该方法以 StarGAN 为基础框架,在生成器中引入自注意力模块,通过计算卷积层中像素块位置的相互作用,捕捉图像之间的依赖关系,利用图像特征的位置线索生成细节,使合成的面部细节更具真实感;另外,在鉴别器中添加谱归一化来约束权重的 Lipschitz 常数,以稳定鉴别器的训练。本文模型与 pix2pix 和 StarGAN 的实验结果相比更具真实感,纹理细节更加丰富,图像质量得到了进一步提升。

1 相关工作

GAN 是一种基于博弈论的深度学习框架[11]。该框架基于随机噪声 z 的输入,让生成模型和鉴别模型交替进行对抗学习:生成模型尽可能欺骗鉴别模型,生成接近于真实数据分布的图像;鉴别模型相当于分类器,对生成的假样本和真样本进行区分和判断,当训练达到最优时,鉴别模型将无法正确区别生成样本和真样本,达到纳什平衡[16]。

以 GAN 为基础进行改进并可用于表情合成的方法主要有以下几种:

(1)pix2pix[12]:该模型的生成器使用 U-Net 结构,与原始的编解码结构相比,更好地共享了网络的输入与输出之间不同分辨率层次的信息;鉴别器采用 patchGAN 结构,将图片按照规定大小切割之后进行判别,其输出为所有切割块判别结果的平均值。该模型要求采用成对的数据集进行训练,即输入和输出有严格的对应关系。

(2)CycleGAN[13]:该模型不局限于 pix2pix 网络中训练集需要一对一的限制,其采用双向循环生成的结构,包含 2 个映射函数,实验时不需要成对的数据集,即可学习 2 个域之间的映射关系,但每次训练只能对单一属性进行改变。

(3)AttGAN[14]:该模型可实现人脸的多属性编辑,其架构主要包括属性分类约束、重构学习和对抗学习三个部分,其中引入属性分类约束确保了生成图片时对合适的属性进行编辑;引入重构学习保证了生成图像能够保留原始图像的身份特征。

(4)StarGAN[15]:该模型使用一个生成器同时训练多个不同域的数据集,实现多域之间的图像编辑。其鉴别器除了能判断图像真假之外,还能将生成图像归类到所属表情域。该模型通过重建原始图像,以保证生成图像仅改变不同域之间存在差异的部分,其余特征保持不变,但表情细节存在一定程度上的缺失。

2 相关原理

2.1 自注意力机制

自注意力(Self-Attention, SA)[17-18]机制是一种将内部关联性和外部信息结合从而提升局部区域的精细度的机制,能够学习某一像素点和其他所有位置像素点之间的关系,可以使生成器和鉴别器对广泛的空间区域进行建模,并将某个位置的注意力计算为局部区域内的像素特征加权求和,在保持全局依赖信息少量损失的前提下,大大降低计算量。

自注意力机制的网络框架如图 1 所示,特征图像 x 通过线性映射转换为 f、g 和 h,其中 $f(x)=W_f x$,$g(x)=W_g x$,利用转置后的 f 和 g 计算相似性和关注度:

$$s_{i,j}=f(x_i)^{\mathrm{T}}g(x_j) \tag{1}$$

$$\beta_{j,i}=\frac{\exp(s_{(i,j)})}{\sum_{i=1}^{N}\exp(s_{(i,j)})} \tag{2}$$

其中,$\beta_{j,i}$ 表示在合成第 j 个像素位置时,模型对第 i 个位置的关注度。那么 SA 映射(o)的输出是 $o=(o_1,o_2,\cdots,o_j,\cdots,o_N)$,这里的计算公式可写为:

$$o_j=\sum_{i=1}^{N}\beta_{j,i}h(x_i),\quad h(x_i)=W_h x_i \tag{3}$$

其中,W_g、W_f 和 W_h 是学习区域内各像素特征的注意力权重,可通过 1×1 卷积来实现。由式(3)的结果乘以一个比例参数,并加上输入的特征图,最终输出为:

$$y_i=\gamma o_i+x_i \tag{4}$$

其中,γ 是一个通过学习得到的标量,初值为

0。通过引入 γ 使网络先学习局部领域的线索，再转向全局的线索，逐渐增加任务的复杂度。

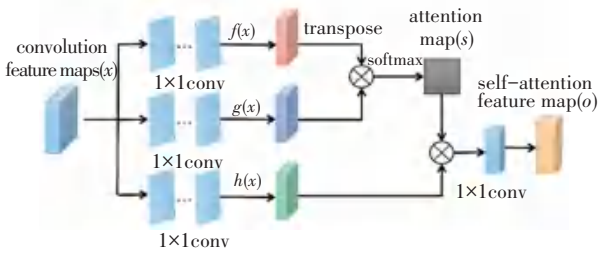

图 1　自注意力机制网络框架

Fig. 1　The framework of self-attention mechanism

2.2　谱归一化

谱归一化（Spectral Normalization，SN）[19] 通过限制训练时函数变化的剧烈程度，使鉴别器更加稳定。实现过程需要让每层网络的网络参数除以该层参数矩阵的谱范数，达到归一化的目的，即可满足 Lipschitz 约束（限制函数的局部变动幅度不能超过某常量）。为获得每层参数矩阵的谱范数，采用幂迭代法来近似求取参数矩阵的谱范数的最大奇异值以减少计算量。谱归一化如下：

$$\bar{W}_{SN}(W)=\frac{W}{\sigma(W)} \tag{5}$$

其中，W 为网络参数的权重，$\sigma(W)$ 为 W 的最大奇异值。

3　模型搭建

3.1　网络框架和结构

本文探讨的表情合成方法以 StarGAN 为基础框架进行改进，在其生成器中引入 2 层自注意力机制模块，丰富上下文联系，使合成表情更具真实感。模型训练时先向生成器提供从训练数据中随机抽取的表情图像和目标表情标签，使生成器能够对表情图像中的细节进行建模，调节表情细节变化，最终通过生成器得到生成图像；下一步，将生成图像输入鉴别器进行判别，鉴别器输出为图像的真假鉴别结果以及图像所属表情域的类别。另外，生成图像与输入图像的表情域标签会再次送入生成器重构原始表情，目的是使生成器能够保持原有图像的身份信息。为稳定鉴别器的训练，在鉴别模型的每一层都引入谱归一化，以确保其映射函数满足 Lipschitz 约束。

本文方法的网络结构如图 2 所示。图 2 中，上半部分的生成器由输入层、输出层、下采样层、瓶颈层、上采样层以及 2 个自注意力机制模块组成，虚线箭头表示生成器重构输入图像的过程，实线表示箭头对抗学习的过程。下半部分的鉴别器由输入层、输出层和隐藏层组成，每一层之间均有谱归一化层和 Leaky-ReLU 激活函数，除输出层以外卷积深度均为前一层的 2 倍，最终经过全连接层映射为 2 个输出，分别用于判别输入生成图像真假和生成表情域。

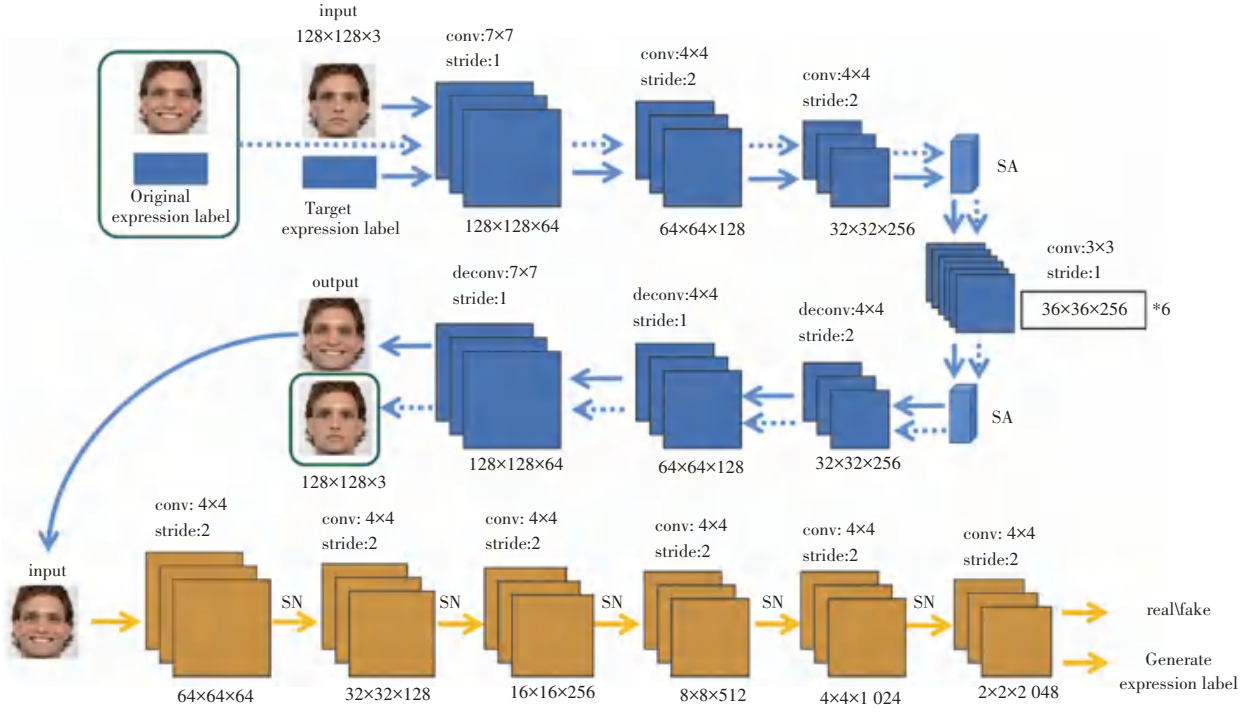

图 2　本文方法的网络结构

Fig. 2　The framework of the proposed method

3.2 损失函数

本文模型的损失函数包括对抗损失函数、分类损失函数和重构损失函数。对此拟做分述如下。

(1)对抗损失函数。为生成与真实图像难以区分的面部表情图像,引入对抗损失函数:

$$L_{adv} = E[\log D(x)] + E_{x,c}[\log(1 - D(G(x,c)))] \tag{6}$$

其中,x,c 分别为原始图像和目标表情域标签,$G(x,c)$ 为生成图像,该图像的表情特征尽可能接近目标表情,鉴别器需要判断生成图像的真实性。

(2)分类损失函数。为使生成器生成具有目标表情特征的假图像,同时鉴别器能够将合成的表情正确归类,提出分类损失函数,分别对生成器和鉴别器进行优化。分类鉴别器损失函数为:

$$L_{cls}^{r} = E_{x,c'}[-\log D_{cls}(c' \mid x)] \tag{7}$$

其中,c' 为输入图像原始表情域标签;$D_{cls}(c' \mid x)$ 为鉴别器将输入图像辨别为原始表情的概率,通过训练使得鉴别器 D 能够将输入图像 x 分类为对应的表情 c'。分类生成器损失函数为:

$$L_{cls}^{f} = E_{x,c}[-\log D_{cls}(c \mid G(x,c))] \tag{8}$$

其中,$D_{cls}(c \mid G(x,c))$ 为生成器将生成图像判别为目标表情 c 的概率,通过训练使生成器尽可能生成符合目标表情特征的表情图像,让鉴别器将表情图像归类到目标表情域 c。

(3)重构损失函数。为保持人脸原有身份信息,引入重构损失函数,利用生成图像重建原始图像:

$$L_{rec} = E_{x,c,c'}[\| x - G(G(x,c),c') \|_1] \tag{9}$$

最终鉴别器和生成器的目标函数分别为:

$$L_D = -L_{adv} + \lambda_{cls} L_{cls}^{r} \tag{10}$$

$$L_G = L_{adv} + \lambda_{cls} L_{cls}^{f} + \lambda_{rec} L_{rec} \tag{11}$$

其中,λ_{cls} 和 λ_{rec} 是超参数,其值大于等于 0,用于控制域分类和重构损失的比重。

4 实验及结果分析

4.1 实验结果

选取 RaFD 数据集[20]的 1 608 张正面表情图像作为训练数据,将图片剪裁为 128×128 进行训练,每更新 5 次生成器后更新 1 次鉴别器,共迭代 200 万次。

本文将 pix2pix 和 StarGAN 作为对比以验证所提出模型的有效性,为保证实验结果的公平性,将 2 种模型的图像分辨率参数均调整至 128×128 进行实验,实验结果如图 3 所示。

图 3 原始图像及不同方法生成的合成图像

Fig. 3 Original images and composite images generated by different methods

从图 3 可看出,pix2pix 合成的表情图像五官较模糊,缺少表情细节,嘴部缺失较为明显,合成效果不理想;StarGAN 合成的表情图像五官较为清晰,但表情细节不够丰富。本文的方法可增强上下文联系,使合成图像更具真实感、质量更高。图 4 是对原始图像、StarGAN 以及本文方法实验结果的局部细节进行比较。由图 4 可看出本文方法所得的表情更接近原始图像,细节更丰富、清晰度更高。

（a）原始图像　　（b）StarGAN

（c）本文方法

图 4　目标表情原始图像以及 StarGAN 和本文方法所生成图像的细节

Fig. 4　Details of the original images of the target expression and the images generated by StarGAN and the proposed method

4.2　定量分析

本文采用 *FID*[21] 作为合成图像的评价指标，通过计算真实图像和生成图像的特征向量之间的距离，评价两者之间的相似度，分数越低表示合成图像越趋近于真实图像。*FID* 公式如下：

$$FID(x,g) = \|\mu_x - \mu_g\|_2^2 + Tr\left(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x\Sigma_g}\right) \tag{12}$$

其中，x 和 g 分别为真实图像和生成图像；μ_x，μ_g 为图像特征的均值；Σ_x，Σ_g 为图像特征的协方差矩阵；$\|\cdot\|_2^2$ 为 L_2 范数的平方；$Tr(\cdot)$ 为矩阵的迹。

本文分别计算了 pix2pix、StarGAN 和本文方法所合成的 8 种表情图像的 *FID*，评估结果见表 1。

根据表 1 可知，本文所提出的模型在愤怒、恐惧、幸福、悲伤、惊奇、蔑视和中立表情的 *FID* 分数相较于 pix2pix 和 StarGAN 均为最低的，合成图像质量相比于其他 2 种算法更佳，生成图像与原始图像更接近。

表 1　pix2pix、StarGAN 和本文方法的 *FID*

Tab. 1　*FID* of pix2pix, StarGAN and the proposed method

Model	Expressions							
	angry	contempt	disgust	fear	happy	neutral	sad	surprise
pix2pix	63.220 5	62.442 1	69.796 2	63.243 6	72.172 9	57.981 4	59.082 3	71.981 3
StarGAN	56.802 0	56.148 6	**53.053 1**	53.748 5	49.985 4	50.867 7	57.842 9	70.267 5
The proposed	**56.428 8**	**55.580 7**	56.674 9	**53.748 2**	**49.732 4**	**48.863 0**	**53.887 3**	**69.747 7**

5　结束语

本文提出一种基于自注意力机制和谱归一化的生成对抗网络表情合成方法，使用生成对抗网络来实现多域之间的表情合成，并引入自注意力机制，使生成器输出更具细节的表情图像，引入谱归一化来约束 Lipschitz 常数，使鉴别器的训练更加稳定。通过对比实验表明，本文模型的合成图像更具真实感，图像质量明显提高。

由于不同的表情数据集之间存在差异，难以用一个模型去泛化所有人的表情，将来希望针对不同背景下的表情合成进行研究。

参考文献

[1] NOH J, NEUMANN U. Expression cloning[C]// Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 2001: 277-288.

[2] MENDI E, BAYRAK C. Facial animation framework for web and mobile platforms[C]//2011 IEEE 13th International Conference on e-Health Networking, Applications and Services. Columbia, MO, USA: IEEE, 2011: 52-55.

[3] KEEVE E, GIROD S, KIKINIS R, et al. Deformable modeling of facial tissue for craniofacial surgery simulation[J]. Computer Aided Surgery, 1998, 3: 228-238.

[4] CHENG Yufang, LING Shuhui. 3D Animated facial expression and autism in Taiwan[C]//2008 Eighth IEEE International Conference on Advanced Learning Technologies. Santander, Spain: IEEE, 2008: 17-19.

[5] LIU Z, YING S, ZHANG Z. Expressive expression mapping with ratio images[C]//Proceedings of the 28th annual conference on Computer graphics and interactive techniques. New York: ACM Press, 2001: 271-276.

[6] 姜大龙，高文，王兆其，等. 面向纹理特征的真实感三维人脸动画方法[J]. 计算机学报，2004(06)：750-757.

[7] 王晓慧，贾珈，蔡莲红. 基于小波图像融合的表情细节合成[J]. 计算机研究与发展，2013，50(02)：387-393.

[8] PARKE F I. Parameterized model for facial animation[J]. IEEE Computer Graphics & Applications, 1982, 2(9): 61-68.

[9] LEI Y W, WU J L, MING O. A three-dimensional muscle-based facial expression synthesizer for model-based image coding[J]. Signal Processing Image Communication, 1996, 8(4): 353-363.

[10] ABBOUD B, DAVOINE F, DANG M. Facial expression recognition and synthesis based on an appearance model[J]. Signal Processing Image Communication, 2004, 19(8): 723-740.

（下转第 129 页）