计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900203

基于改进 CycleGan 模型和区域分割的表情动画合成

叶亚男[1,2]　迟　静[1,2]　于志平[1,2]　战玉丽[1,2]　张彩明[1,2,3,4]
1 山东财经大学计算机科学与技术学院　济南 250014
2 山东省数字媒体技术重点实验室　济南 250014
3 山东大学软件学院　济南 250101
4 未来智能计算协同创新中心　山东 烟台 264003
(1325809478@qq.com)

摘　要　针对现有人脸表情合成大多依赖于数据源驱动,且存在生成效率低、真实感差的问题,提出一种基于改进 CycleGan 模型和区域分割的表情动画合成新方法。新方法可实时地合成新表情动画,且具有较好的稳定性和鲁棒性。所提方法在传统 CycleGan 模型的循环一致损失函数中构造新的协方差约束条件,可有效避免新表情图像生成时出现的色彩异常和模糊不清等现象;提出分区域训练的思想,用 Dlib 人脸识别数据库对人脸图像进行关键点检测,通过检测到的关键特征点将源域和目标域的人脸分割成左眼、右眼、嘴部和剩余人脸部分共 4 个区域块,并利用改进的 CycleGan 模型对每块区域单独进行训练;最后将训练结果加权融合成最终的新表情图像。分区域训练进一步增强了表情合成的真实感。实验数据来自英国萨里大学的语音视觉情感(SAVEE)数据库,在 Tensorflow 框架下,用 python 3.4 软件进行实验结果的展示。实验表明,新方法无需数据源驱动,可直接在源人脸动画序列上实时地生成真实、自然的新表情序列,且对于语音视频可保证新面部表情序列与源音频同步。

关键词:表情合成;区域分割;CycleGan;协方差约束;深度学习
中图法分类号　TP391.41

Expression Animation Synthesis Based on Improved CycleGan Model and Region Segmentation

YE Ya-nan[1,2],CHI Jing[1,2],YU Zhi-ping[1,2],ZHAN Yu-li[1,2] and ZHANG Cai-ming[1,2,3,4]
1 School of Computer Science and Technology,Shandong University of Finance and Economics,Jinan 250014,China
2 Shandong Provincial Key Laboratory of Digital Media Technology,Jinan 250014,China
3 School of Software,Shandong University,Jinan 250101,China
4 Future Intelligent Computing Collaborative Innovation Center,Yantai,Shandong 264003,China

Abstract　Aiming at the problems of mostly relying on data source driver,low generation efficiency and poor authenticity of the existing facial expression synthesis methods,this paper proposes a new method for expression animation synthesis based on the improved CycleGan model and region segmentation. This new method can synthesize new expression in real time and has good stability and robustness. The proposed method constructs a new covariance constraint in the cycle consistent loss function of the traditional CycleGan model,which can effectively avoid color anomalies and image blurring in generation of new expression images. The idea of zonal training is put forward. The Dlib face recognition database is used to detect the key points of the face images. The detected key feature points are used to segment the face in domain source and target domain into four zones:left eye,right eye,mouth and the rest of the face. The improved CycleGan model is used to train each region separately,and finally the training results are weighted and fused into the final new expression image. The zonal training further enhances the authenticity of expression synthesis. The experimental data comes from the SAVEE database,and the experimental results are presented with python 3.4 software under the Tensorflow framework. Experiments show that the new method can directly generate real and natural new expression sequences in real time on the original facial expression sequence without data source driver. Furthermore,for

到稿日期:2019-06-16　**返修日期**:2019-09-30　**本文已加入开放科学计划**(OSID),请扫描上方二维码获取补充信息。
基金项目:山东省省属优青项目(ZR2018JL022);国家自然科学基金(61772309,61602273);山东省重点研发计划(2019GSF109112);山东省教育厅科技计划项目(J18RA272);山东省高等学校优势学科人才团队培育计划
This work was supported by the Natural Science Foundation of Shandong Province for Excellent Young Scholars in Provincial Universities (ZR2018JL022),National Natural Science Foundation of China (61772309,61602273),Shandong Provincial Key R&D Program (2019GSF109112),Science and Technology Program of Shandong Education Department (J18RA272) and Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.
通信作者:迟静(peace_world_cj@126.com)

the voice video，it can effectively ensure the synchronization between the generated facial expression sequence and the source audio.
Keywords Facial expression synthesis，Region segmentation，CycleGan，Covariance constraint，Deep learning

1 引言

具有真实感的人脸表情动画合成一直是计算机图形学和计算机视觉领域的研究热点和难点之一，且被广泛应用于数字娱乐、视频会议、医疗、辅助教育等领域。目前，表情合成的主要方式有：1）手工编辑人脸模型，生成一帧帧的新表情；2）将源表情传递到目标人脸，在目标人脸上重现该表情；3）融合已有的表情样本，生成新表情。第一种方式允许对已有的源表情数据进行任意编辑，但耗时、耗力，且对操作人员的专业技术要求较高。第二种和第三种方式需要借助表情数据源来驱动，合成的新表情数目和质量受限于已有的源表情数据规模，且合成的表情真实感通常不高，尤其在处理语音视频时，往往难以实现表情重现与源视频中音频的同步。

针对上述问题，本文提出了一种基于改进 CycleGan[1] 模型和区域分割思想的表情动画合成新方法。该方法无需借助数据源驱动，直接将人脸动画序列中的源表情转换成任意的新表情，如将中性表情下的演讲过程转变成在惊讶表情下的演讲过程（见图 1），且生成的新表情下的动画序列连贯、真实、自然。该方法在处理语音视频时，可很好地实现新面部表情序列与源音频的同步。

图 1 源语音视频中的中性表情（上）被转换成惊讶表情（下）

Fig. 1 Neutral expression (above) in the source voice video converted into a surprised expression (below)

2 相关工作

目前已有的人脸表情合成方法很多，主要分为两大类：基于三维网格模型的人脸表情合成和基于二维图像的人脸表情合成。

（1）基于三维网格的方法：通过跟踪一系列人脸表情变化的曲线和基于全局统计的模型生成所需的目标图像，实现表情的合成。Pighin 等[2] 提出了具有逼真纹理的三维人脸建模的表情合成系统。Blanz 等[3] 提出了从单张图片或视频中恢复人脸的三维模型，尝试从带纹理的三维模型中进行表情合成，但创建一个好的模型相当困难，因为必须对面部的所有细节进行建模，如眼睛、头发、牙齿等。Vlasic 等[4] 提出的多线性模型（Multilinear Models）将分离参数化不同的属性（如顶点信息、形状、视位、表情等）建立在同一个数据张量空间中，这些属性之间用笛卡尔积来构造并且相互独立。将这些独立属性参数进行任意组合，最终得到不同的人脸表情。Lv 等[5] 提出一种面向同一人脸表情转移的方法，即对目标人脸进行三维建模，生成特定的 blendshape 模型，利用该模型生成与输入人脸图像匹配的三维人脸模型，并对图像进行扭曲融合，生成所需的目标人脸图像。Pasquariello 等[6] 在三维人脸模型中加入皱纹等细节因素，将人脸模型进行网格化，并按照人的生理结构将人脸分成嘴、眼睛、眉毛、额头等区域，使得每一区域的网格数量和拓扑等都不同，进而实现表情的模拟。Zhang 等[7] 在三维形变模型中也加入了皱纹等细节，将其划分为 14 个子区域，避免了表情皱纹超过分区的边界。Joshi 等[8] 提出的基于物理的分割方法可以自动将人脸分割成多个区域，每一个区域都表示成混合形状（blendshape）的线性组合，从混合形状中学习约束条件和参数。Park 等[9] 将给定的每一个源关键模型分成 3 个子区域，每一个子区域都包含人脸的关键特征，实现了面部表情的合成。Joshi[10] 把每一个表情看成其他表情的线性组合，通过改变这些线性组合的权重来合成比较完整的面部表情。Garrido[11] 提出了一种基于图像的人脸视频再现方法，对于输入的两个不同人脸的面部表情视频，将源序列的表情传递给目标序列，同时尽可能地保留灯光、背景等因素。该方法的缺点是：需要一些特定的人脸数据库，且依赖于输入图像的一些表情信息，如是否是中性人脸或者有无标记点等等。文献[12-13]首先对人脸进行建模，然后得到特定用户的 blendshape 模型，进而求解出 blendshape 系数，最后通过对系数的改变来合成目标表情。Huang 等[14] 提出了基于非联合学习的人脸表情合成方法：通过一种无监督回归的算法，将具有相同属性的三维人脸模型映射到同一个低维空间，对其进行重建，实现人脸表情的合成。

（2）基于二维图像的方法：用已知的表情数据来合成新的表情，或者直接将已有的表情传递到目标图像上。Williams[15] 提出通过表情映射的方法来合成新的目标表情，首先提取两幅图像不同的面部特征，然后计算特征之间的矢量差，最后利用特征向量来进行图像的扭曲。该方法虽然实现了不同表情之间的转换，但是不适用于戴眼镜和头部位姿有较大变化的情况。Yang 等[16] 提出了基于表情流的方法：首先提取两幅图像的特征点并分别进行三维人脸重建；然后计算两个三维模型之间的差异，将差异映射到二维图像上得到表情流，再利用所得表情流进行图像扭曲；最后进行图像融合[17]。表情流的计算方法往往比较复杂，鲁棒性很差，得到的效果不是很逼真，而且只能在同一人脸之间进行，不具有普遍性。文献[18-19]提出了一种人脸图像自动替换系统：首先对输入图像进行人脸检测；然后将提取的每个人脸进行对齐，并从大量的人脸数据库中找到与其相近的人脸；最后通过图像融合实现目标输入图像的表情合成。这些方法可以有效、快捷地合成人脸表情，但由于人脸具有特异性，且表情复杂、丰富，在表情合成过程中合成具有真实感的表情比较困难。

本文提出一种基于改进 CycleGan 模型和区域分割思想的人脸表情合成方法。在 CycleGan 模型的循环一致损失函数中引入新的协方差约束条件，用来约束源图像（或目标图

像)和重建后的源图像(或目标图像)之间的误差。新约束条件既可避免在大数据样本下将全部源图像转换成同一目标图像,又可避免在转换过程中出现色彩异常和模糊不清等现象,从而有效提高了表情合成的精度。为进一步提高人脸表情转换模型的鲁棒性和真实感,本文引入了区域分割的思想,即根据人脸的几何结构及脸部不同区域的表情变化特点,将输入的源人脸图像分割成左眼、右眼、嘴部和剩余人脸部分共 4 块区域,单独对每块区域进行训练,将所得的分块结果图进行加权融合,最终得到完整、真实自然的目标人脸表情图像。

3 改进 CycleGan 模型和区域分割方法

针对传统 CycleGan 算法在人脸表情合成过程中所出现的生成目标人脸图像清晰度不高、局部细节信息丢失严重以及色彩失真等缺点,本文引入了协方差约束条件来学习一种新的双向匹配映射关系,以提高生成目标图像的质量。为使表情合成具有更强的真实感,本文将源图像进行区域划分,并利用上述学习的表情映射关系生成目标人脸区域结果图,然后将转换后的区域结果图加权融合成最终的目标人脸图像。

本文方法的基本流程如图 2 所示。

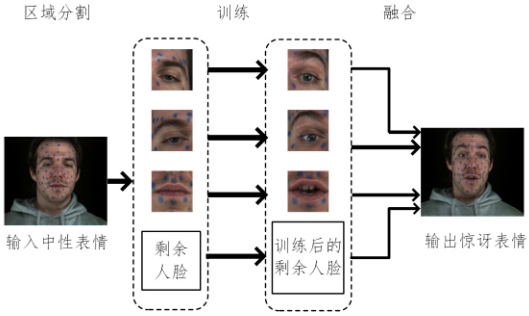


图 2 改进 CycleGan 模型和区域分割方法流程图

Fig. 2 Flow chart of improved CycleGan model and region segmentation method

(1)区域分割:输入一系列人脸表情图像,用检测到的关键特征点的位置对人脸图像进行区域划分。

(2)分区域训练:用改进的 CycleGan 模型对分割后的子区域进行单独训练,得到转换后每部分区域的结果图。

(3)图像融合:将转换后的区域结果图合成完整的人脸表情图像,用像素加权融合算法平滑合成的边界,使生成的目标人脸图像更加自然。

步骤(2)利用改进的 CycleGan 算法进行区域训练,其训练精度对表情合成结果有着重要影响。本文通过在传统 CycleGan 模型的循环一致损失函数中引入新的协方差约束项来提高其训练精度。

3.1 新的循环一致损失函数

传统的 CycleGan 模型包括两个生成器和两个判别器,生成器和判别器的网络结构分别如图 3 和图 4 所示。其中,图 3 所示生成器 $G:X\to Y$ 是由 1 个输入层、3 个隐藏层和 1 个输出层组成。输入层和输出层的维数 $m=46$,每 1 个隐藏层的维数 $n=100$。图 4 所示判别器 D 是由 1 个输入层、3 个隐藏层和 1 个输出层组成。输入层的维数 $m=46$,每个隐藏层的维数 $n=100$,输出层的 p 为概率值。模型通过学习源域到目标域的表情映射关系,实现两个表情序列间的转换。

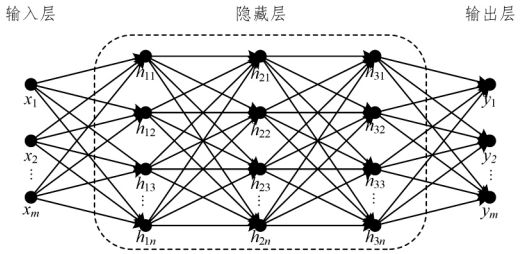


图 3 生成器

Fig. 3 Generator

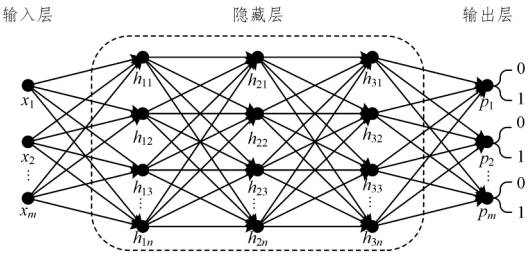


图 4 判别器

Fig. 4 Discriminator

模型由循环一致损失函数和对抗性损失函数组成,如图 5 所示。其中,$G:X\to Y$ 和 $F:Y\to X$ 是源域和目标域之间的两个匹配函数;D_X 和 D_Y 是两个判别器,用来区分真实的样本数据与转换后的样本数据;图中虚线代表两个循环一致性损失。循环一致损失函数实现了图像的转换,避免了大数据量下将所有的源表情序列转化成同一目标表情序列;对抗性损失函数判别转换后的图片是否为真实数据库的图片,提高了图像转换的精度。用传统的 CycleGan 模型进行训练,会导致生成的目标域图像出现模糊不清和色彩不一致等现象。为了解决这个问题,本文用协方差约束项来构建新的循环一致损失函数,使得其在表情转换过程中可以生成较高质量的图片。

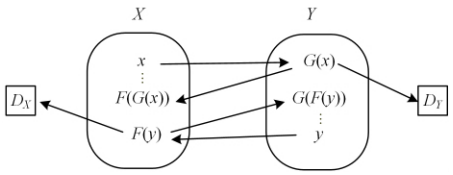


图 5 源域和目标域之间的两个匹配函数

Fig. 5 Two mapping functions between two expressions

设源表情序列为 X 域,目标表情序列为 Y 域。训练样本为 $\{x_i\}_{i=1}^{N}$ 和 $\{y_j\}_{j=1}^{M}$,其中 $x_i\in X$,$y_j\in Y$,记 X 域和 Y 域中的一个训练样本分别为 x 和 y。本文的模型学习了两个映射关系 $G:X\to Y$ 和 $F:Y\to X$。其中,G 的目的是让变换后的样本 $G(x)$ 更接近 Y 域中的真实样本;同理,F 的目的是让变换后的样本 $F(y)$ 更接近 X 域中的真实样本。

G 将 X 域的样本数据转换成 Y 域中的样本数据 $G(x)$,再通过 F 映射回 X 域中的样本数据 $F(G(x))$;同理,Y 域中的样本数据经过一个循环变换后,变成 $G(F(y))$。转换后的样本数据应尽可能地与真实样本数据接近,即 $F(G(x))\approx x$,$G(F(y))\approx y$。用欧氏距离对两个样本数据进行约束,以保证生成的 $G(x)$ 和 $F(y)$ 都可以分别映射回 X 域和 Y 域。原始的循环一致损失函数[20]为:

$$E_{cyc}(G,F)=\|F(G(x))-x\|_1+\|G(F(y))-y\|_1 \quad (1)$$

其中，$\|\ \|_1$ 为 1 范式，即欧氏距离。

本文提出的新的循环一致损失函数由欧氏距离约束项和协方差约束项构成。其中，欧氏距离约束项是 CycleGan 中常采用的约束项；而本文新提出的协方差约束项可进一步约束源（目标）图像和转换后源（目标）图像之间的相似程度。欧氏距离约束项作为 CycleGan 模型中常采用的约束项，衡量的是空间中各个像素点颜色间的绝对距离，也就是颜色的差异，可以在一定程度上反映两个图像间的相似性。但该约束项不能反映数据集中各个像素点颜色的分布差异，而像素颜色分布的相似性也是衡量两幅图像之间相似性的重要指标。显然，像素颜色的分布越相似，两幅图像的相似程度就越高。因此，本文提出新的协方差约束项，用于反映两幅图像间像素颜色分布的相似程度。

协方差表示图像数据集不同维度之间的相关程度。源（或目标）图像和循环转换回来的源（或目标）图像的协方差矩阵之差越小，表示相关程度越高，两幅图像的像素颜色分布越相似，图像自然也越相像。通过最小化真实数据和循环转换数据之间的协方差矩阵之差，可使生成的目标图像更加清晰、自然，且包含丰富的表情细节信息。

设样本图像为 $\boldsymbol{x}=[x_{-1}x_{-2}\cdots x_b]$，$\boldsymbol{x}_k$ 为图像的列像素，记为 $\boldsymbol{x}_{-k}=[x_{1k}x_{2k}\cdots x_{ak}]^{\mathrm{T}}$，$k=1,\cdots,b$。这里，$\boldsymbol{x}_{ij}$（$i=1,\cdots a$，$j=1,\cdots,b$）为图像的像素点，$a$ 是样本图像的宽度（行数），b 是样本图像的长度（列数）。样本图像所有的像素点可以表示为一个 $a\times b$ 的矩阵，则该样本图像的协方差矩阵 $\sum\boldsymbol{x}$ 可表示为：

$$\sum\boldsymbol{x}=\begin{bmatrix}q_{11}&q_{12}&\cdots&q_{1a}\\q_{21}&q_{22}&\cdots&q_{2a}\\\vdots&\vdots&&\vdots\\q_{a1}&q_{a2}&\cdots&q_{aa}\end{bmatrix}=\frac{1}{b}\sum_{k=1}^{b}(\boldsymbol{x}_{-k}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{-k}-\bar{\boldsymbol{x}})^{\mathrm{T}} \quad (2)$$

其中：

$$\begin{cases}q_{11}=\dfrac{1}{b}\sum\limits_{k=1}^{b}(\boldsymbol{x}_{1k}-\bar{\boldsymbol{x}_1})(\boldsymbol{x}_{1k}-\bar{\boldsymbol{x}_1})\\q_{12}=\dfrac{1}{b}\sum\limits_{k=1}^{b}(\boldsymbol{x}_{2k}-\bar{\boldsymbol{x}_1})(\boldsymbol{x}_{2k}-\bar{\boldsymbol{x}_2})\\\cdots\\q_{aa}=\dfrac{1}{b}\sum\limits_{k=1}^{b}(\boldsymbol{x}_{ak}-\bar{\boldsymbol{x}_a})(\boldsymbol{x}_{ak}-\bar{\boldsymbol{x}_a})\end{cases} \quad (3)$$

$\bar{\boldsymbol{x}}$ 是该样本图像的像素均值，x_{-k} 是该样本图像的第 k 列像素。

同理，计算转换后样本数据的协方差。针对转换后的协方差矩阵应尽可能与真实样本的协方差矩阵相似的问题，提出了协方差保持约束项：

$$E_{cov}(G,F)=\|\sum(F(G(x)))-\sum(x)\|_1+\|\sum(G(F(x)))-\sum(y)\|_1 \quad (4)$$

新的循环一致损失函数由式（1）和式（4）的约束项加权组合得到：

$$E_{ncyc}=\lambda E_{cyc}+\mu E_{cov} \quad (5)$$

其中，λ 和 μ 用于调节各约束项所占的比重。协方差和欧氏距离的共同约束，不仅提高了转换图像的清晰程度，而且有效地增强了模型的真实感。

本文模型包含两个判别器 D_X 和 D_Y，用来判别转换后的样本数据是否是真实的样本数据。具体来说，D_X 用来区分从 Y 域转换生成的样本数据 $F(y)$ 与 X 域中的真实样本数据 x；D_Y 用来区分从 X 域转换生成的样本数据 $G(x)$ 和 Y 域中的真实样本数据 y。为了使转换后的样本数据和目标样本数据尽可能地相近，本文采用传统 CycleGan 模型中的对抗性损失函数，其表达式如下：

$$E_{GAN}(G,D_Y)=E_{y\sim pY}[\log D_Y(y)]+E_{x\sim pX}[\log(1-D_Y(G(x)))] \quad (6)$$

$$E_{GAN}(F,D_X)=E_{x\sim pX}[\log D_X(x)]+E_{y\sim pY}[\log(1-D_X(F(y)))] \quad (7)$$

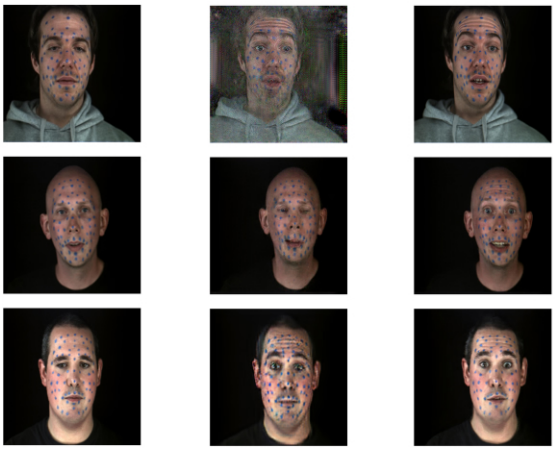
其中，D_X 和 D_Y 均为 0，1 二分类的损失函数。$x\sim X$ 和 $y\sim Y$ 表示 X 域和 Y 域样本数据的分布。X 域的样本数据 x 通过映射函数 G 生成 Y 域的样本 $G(x)$，判别器 D_Y 用于判断 $G(x)$ 是不是 Y 域本身的数据；而对于 G 来说，希望 $D_Y(G(x))$ 无限接近 Y 域本身的样本数据。同理，判别器 D_X 用来判断从 Y 域映射过来的样本数据 $F(y)$ 是不是 X 域本身的数据；而对于 F 来说，希望 $D_X(F(y))$ 无限接近 X 域本身的样本数据。

由式（5）—式（7）构成的新的总损失函数为：

$$E(G,F,D_X,D_Y)=E_{GAN}(G,D_Y)+E_{GAN}(F,D_X)+E_{ncyc} \quad (8)$$

其中，E_{ncyc}（式（5））是本文提出的新的基于协方差约束的循环一致损失函数，实现了源域到目标域的人脸表情转换，提高了目标域图像的质量。

图 6 比较了利用传统 CycleGan 算法和改进的 CycleGan 算法对不同人脸进行表情合成的效果。这里，输入一幅中性表情图片，训练一个从中性到惊讶的表情映射关系。



(a) Source images (b) Traditional CycleGan result images (c) Improved CycleGan result images

图 6 传统 CycleGan 和改进 CycleGan 结果图的比较

Fig. 6 Comparison of traditional CycleGan and improved CycleGan results

由图 6 可见，在传统 CycleGan 结果图中，人脸和背景部分都出现了明显的色彩异常，并且清晰度较低，丢失了大量的表情细节。本文提出的基于协方差约束的人脸表情合成方法，在传统 CycleGan 模型中引入协方差约束项来约束循环一致损失函数，在很大程度上改善了传统模型的训练结果，消除了转换后明显的色彩差异和局部细节丢失的情况。

利用改进的 CycleGan 模型进行表情转换，提高了目标图像的质量；但在某些情况下，仍存在目标图像的眼睛、嘴等局部位置细节信息丢失的情况，如图 7 所示。为进一步改善这种情况，提高人脸表情转换的鲁棒性和适应性，本文引入分区域训练的思想，提出新的表情转换框架和方法，以增强人脸局部位置的细节信息，进一步提高生成目标表情的质量。

(a)Source image (b)Improved CycleGan result image

图 7 改进 CycleGan 对 KL 进行表情转换的结果

Fig. 7 Expression conversion results of KL by the improved CycleGan

3.2 区域分割

通过检测到的关键特征点对源域和目标域的人脸进行区域分割。依据表情转换过程中人脸各个区域的易变换程度，将人脸划分成左眼、右眼、嘴部和剩余人脸部分共 4 块区域。利用基于协方差约束的 CycleGan 模型，将训练图像的训练区域从整张人脸缩小到一个子区域，去除了无关特征的影响，改善了转换效果并提高了转换后图像的质量，同时有效地减少了训练时间。

由于源域和目标域图片的训练数据量较大，且训练样本中的人脸的几何结构各不相同，直接利用检测到的关键特征点位置进行区域划分，会导致不同样本分割后其对应的同一区域(如左眼区域)大小不一致，使得分割结果无法用于后期的训练。对此，本文划分子区域时限定了分割窗口的大小，即在划分子区域时将同一子区域划分为统一尺寸的图片，然后再进行训练，生成子区域目标表情。

区域分割的步骤如算法 1 所示。

算法 1 区域分割算法

输入：训练样本 $\{x_i\}_{i=1}^{N}$ 和 $\{y_j\}_{j=1}^{M}$

输出：左眼区域 $\{xl_i\}_{i=1}^{N}$ 和 $\{yl_j\}_{j=1}^{M}$，右眼区域 $\{xr_i\}_{i=1}^{N}$ 和 $\{yr_j\}_{j=1}^{M}$，嘴部区域 $\{xm_i\}_{i=1}^{N}$ 和 $\{ym_j\}_{j=1}^{M}$，以及剩余人脸区域 $\{xc_i\}_{i=1}^{N}$ 和 $\{yc_j\}_{j=1}^{M}$

Step1 导入 Dlib 人脸识别数据库。

Step2 对每一个样本 $x_i(i=1,\cdots,N)$ 和每一个样本 $y_j(j=1,\cdots,N)$：

Step2.1 检测出 68 个人脸特征点，利用特征点标定左眼、右眼和嘴部区域；

Step2.2 计算每个子区域的中心点；

Step2.3 计算每个子区域的长和宽，并将这两个值中较大的一个暂记为该子区域的窗口大小。

Step3 对每一个子区域，取 Step2 中所有样本对应的该子区域的窗口值中最大的一个作为该子区域的最终窗口大小。

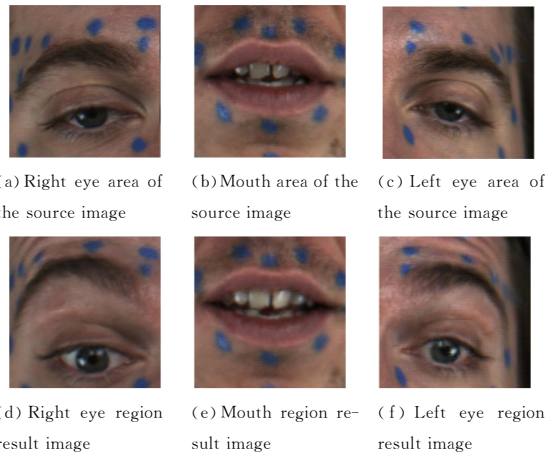
Step4 对每一个样本 $x_i, i=1,\cdots,N$ 和每一个样本 $y_j, j=1,\cdots,M$，根据 Step2.2 中所得每个子区域的中心点坐标和 Step3 中所得的每个子区域的最终窗口大小，分割出最终的左眼区域 xl_i 和 yl_j、右眼区域 xr_i 和 yr_j、嘴部区域 xm_i 和 ym_j，整张图片的剩余部分记为剩余人脸区域 xc_i 和 yc_j。

本文用人脸识别的一个数据库 Dlib 对人脸特征点进行检测，其计算量小，速度快且准确率高，具有良好的实时性和鲁棒性。该方法使用的是 Ensemble of Regression Tress 级联回归的算法[21]，简称 ERT 算法。将输入人脸图像 I 的所有特征点的形状记为 S，用 ERT 算法建立对应人脸的 ERT 模型，将上述模型不断地进行迭代，得到一个最优的模型。首先，初始化人脸特征点形状 S，计算出所有特征点对的像素差，用像素差特征进行训练，得到随机森林，其中叶子节点保存特征点模型残差，非叶子节点保存相应点和节点的分离阈值。对本层所有的树求残差并将其相加得到残差总和，将所得残差总和结果与前一次迭代的结果进行相加，经过多次迭代，输出最后拟合好的人脸检测模型。利用人脸检测模型识别人脸面部关键特征点，再依据特征点的位置将人脸表情划分为表情易变区域和表情相对不变区域，分开对其进行处理。这样的操作一方面克服了将整个人脸作为一个研究对象来处理所存在的复杂性；另一方面减少了背景和其他物体的干扰，提高了训练结果的准确性，减少了图像数据处理量，同时节省了处理时间，具有较强的适应性。

3.3 基于 CycleGan 的表情映射

利用本文提出的基于协方差约束的 CycleGan 新模型，分别对分割后的各个人脸子区域图像集进行训练，实现源表情序列到目标表情序列的转换。分区域转换可有效避免直接使用完整人脸图像进行表情转换时所产生的人脸扭曲、五官错位和图像模糊等问题，提高了表情转换的稳定性。该转换模型包含两个生成器和两个判别器。判别器是一个卷积神经网络，从输入的图像中提取特征，通过添加产生一维输出的卷积层，判断图像提取的特征是否属于给定的类别。生成器的网络结构与文献[22]的网络结构类似(见图 3)，是由两个步幅为 2 的卷积、两个步幅为 1/2 的卷积和几个残差块组成。其中，步幅为 2 的卷积进行下采样，步幅为 1/2 的卷积进行上采样，减少了参数的数量，从而提高了系统的性能。生成器包含 1 个输入层、3 个隐含层和 1 个输出层。隐含层使用线性整流函数(Rectified Linear Unit，ReLU)作为激活函数。除输出层以外，所有非残差卷积层后面都用批次归一化[23]和 ReLU 进行非线性处理；输出层使用缩放的双曲正切函数(tanh)来进行约束，确保输出的像素在[0，255]这个有效范围之内。输入层和输出层都用 46 个单位来表示表情向量，每一个隐藏层都有 100 个单位。判别器使用的是 70 * 70 的全卷积网络 PatchGans[24]，减少了网络参数数量(见图 4)。输出层仅有一个单元能产生一个概率，这个概率表示输入的样本是否为真实样本。系统采用 Adam 优化器[25]进行优化，将步长 *batch-size* 设置为 1。本文将网络学习的单次迭代次数设置为 200 epochs，前 100 个 epochs 将学习率设置为 0.000 2，后面 100 个 epochs 对学习率进行线性衰减，直至学习率衰减至 0 时，学习结束。

图 8—图 10 给出了采用具有新的循环一致损失函数的 CycleGan 模型分别对 KL，JK 和 JE 进行分区域训练得到的结果。可以明显地看出，利用改进的 CycleGan 模型进行训练，较好地保持了每部分区域的细节信息，提高了训练结果的准确性，增强了目标图像的真实感，从而证明了所提方法具有很好的鲁棒性和适应性。

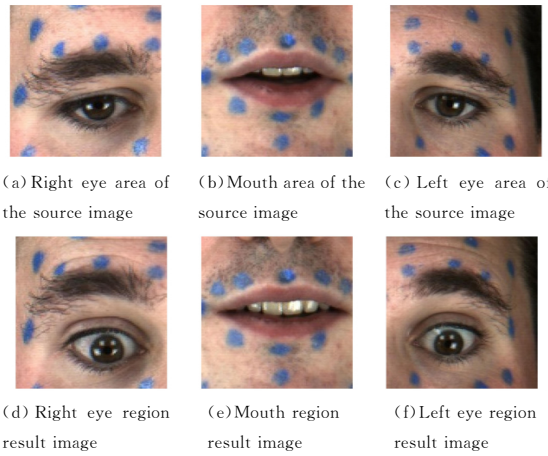


(a) Right eye area of the source image (b) Mouth area of the source image (c) Left eye area of the source image

(d) Right eye region result image (e) Mouth region result image (f) Left eye region result image

图 8 KL 分区域训练图

Fig. 8 Zonal training images of KL

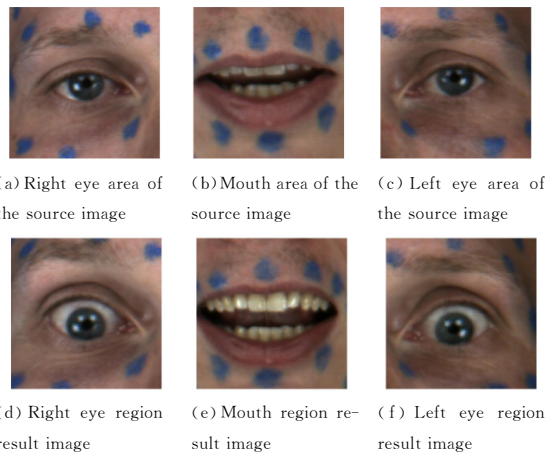


(a) Right eye area of the source image (b) Mouth area of the source image (c) Left eye area of the source image

(d) Right eye region result image (e) Mouth region result image (f) Left eye region result image

图 9 JK 分区域训练图

Fig. 9 Zonal training images of JK



(a) Right eye area of the source image (b) Mouth area of the source image (c) Left eye area of the source image

(d) Right eye region result image (e) Mouth region result image (f) Left eye region result image

图 10 JE 分区域训练图

Fig. 10 Zonal training images of JE

3.4 图像融合

将训练好的分区域结果图进行融合，形成完整的目标人脸表情图像。为避免在融合过程中出现区域边界过渡不自然的现象，本文采用加权融合的思想，将区域边界一定范围内的像素点进行加权融合。区域融合不仅提高了图像的质量和清晰度，还提高了图像的信噪比。具体实现过程如下。

设两幅区域图像分别为 $M(m,n)$ 和 $N(m,n)$，则对于两幅图像交界范围内的每一个像素点 (m_i,n_j)，其像素值 $F(m_i,n_j)$ 的融合方式如下。

1）若交界区域为纵向的，即图像 $M(m,n)$ 和 $N(m,n)$ 左右相邻，则：

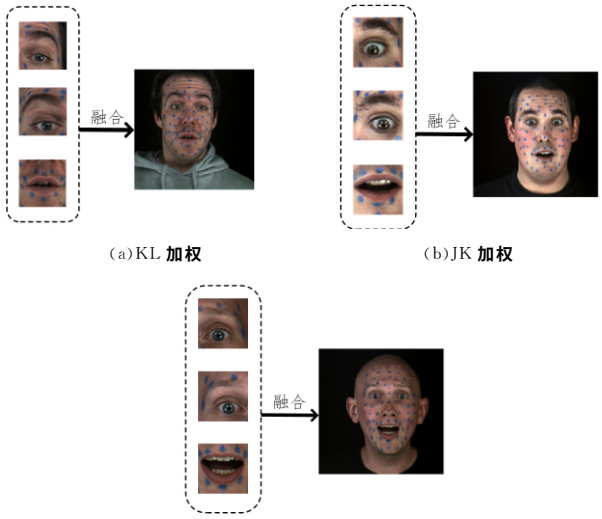
$$F(m_i,n_j)=\omega_1 M(m_{i-1},n_j)+(1-\omega_1)N(m_{i+1},n_j)+\omega_2 M(m_{i-2},n_j)+(1-\omega_2)N(m_{i+2},n_j)+\cdots+\omega_k M(m_{i-k},n_j)+(1-\omega_k)N(m_{i+k},n_j) \tag{9}$$

2）若交界区域为横向的，即图像 $M(m,n)$ 和 $N(m,n)$ 上下相邻，则：

$$F(m_i,n_j)=\omega_1 M(m_i,n_{j-1})+(1-\omega_1)N(m_i,n_{j+1})+\omega_2 M(m_i,n_{j-2})+(1-\omega_2)N(m_{i+2},n_{j+2})+\cdots+\omega_k M(m_i,n_{j-k})+(1-\omega_k)N(m_i,n_{j+k}) \tag{10}$$

其中，k 为步长，以点 (m_i,n_j) 为中心，左右（或上下）各取 k 个像素进行融合，这个过程相当于对交界处的像素进行了模糊处理；$\omega_1,\omega_2,\cdots,\omega_k$ 为融合系数，其值根据相应像素点与点 (m_i,n_j) 的距离来确定。利用式(9)和式(10)对左眼、右眼、嘴和剩余人脸 4 个区域的交界处进行模糊处理，并反复进行多次模糊，直到得到自然的融合图像。该融合方法简单直观、速度快，可以应用在实时性要求比较高的场合。

图 11 展示了将训练得到的各区域结果进行加权融合的效果图。融合后的图像提高了图像信息的利用率，形成了对目标人脸图像清晰、完整、准确的信息描述，同时消除了图像冗杂信息，提高了生成面部人脸图像的质量。



(a)KL 加权 (b)JK 加权

(c)JE 加权

图 11 融合效果图

Fig. 11 Weighted fusion effect images

4 实验结果

本文实验数据来自 SAVEE 数据库[26]，该数据库是 4 名中年男性演员的视频剪辑，分别为 DC，JK，JE 和 KL。其中包括 7 种情绪类别（愤怒、厌恶、恐惧、快乐、中性、悲伤、惊喜）。每个人用英语说了 120 句话。每段视频均以 60 帧/秒的速度录制，共产生大约 10 万张照片。本文训练了一个从中性到惊讶的表情映射关系，因为中性情绪有 30 个句子，惊讶情绪有 15 个句子，数据量庞大，所以对其进行等间隔采样，并选取一部分作为训练集，训练模型的输入图片库数量为 820 张。一个模型需要训练 8h，测试一张图片需要 97.48 ms。

实验环境：Tensorflow[28]**框架**，python3.4 **软件**，Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz **处理器和** NVIDIA Geforce GTX 1060 GPU。

传统的 CycleGan **网络多用于图像风格的迁移。本文把改进后的网络应用于人脸表情的合成中，将实验结果与基于传统** CycleGan **模型的表情合成和基于** StarGan[27] **模型的表情合成结果进行比较，在对比中使用相同**(SAVEE)**数据库中** KL，JK **和** JE **人脸表情图像。本文对整体的视觉效果进行对比，结果如图** 12—**图** 14 **所示。**

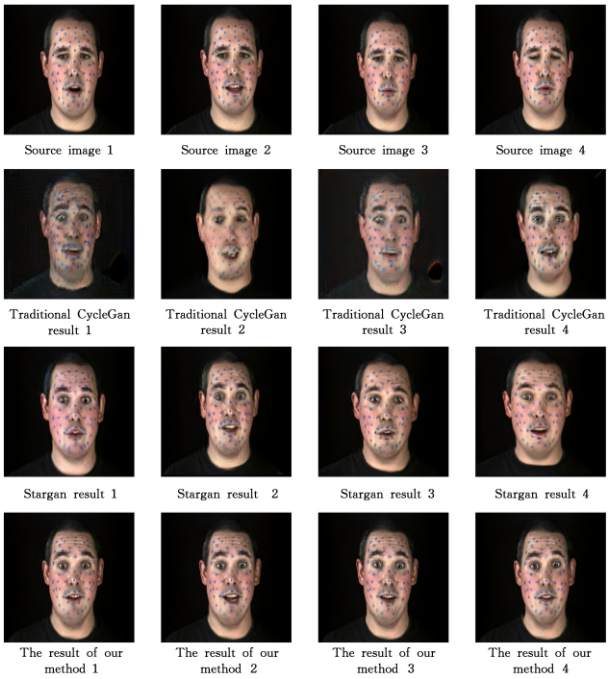


图 12 JK **表情转换结果图**

Fig. 12 Expression conversion result images of JK

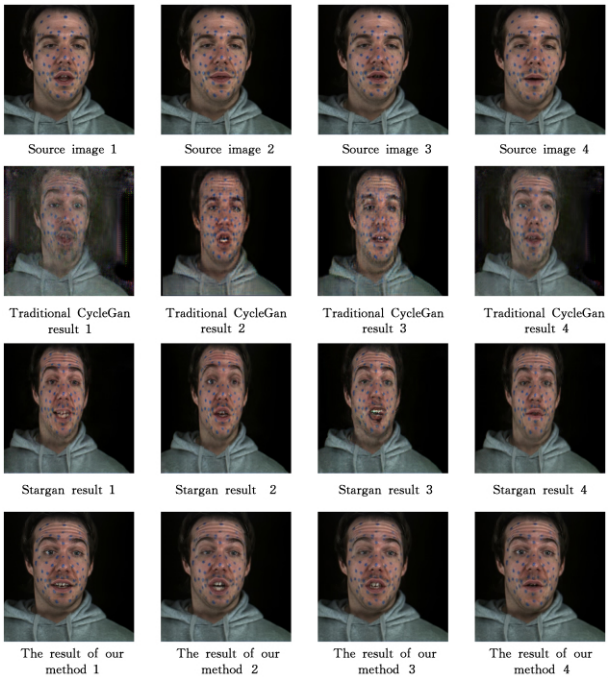


图 13 KL **表情转换结果图**

Fig. 13 Expression conversion result images of KL

图 14 JE **表情转换结果图**

Fig. 14 Expression conversion result images of JE

本文将循环一致损失的权重系数 λ **设为** 10，μ **设为** 1。μ **设为** 1 **避免了在表情转换过程中将所有表情转换成同一个表情而出现过拟合现象。在图像融合中，边界融合通常在** 10 **次迭代中收敛，达到了比较不错的效果。由图** 12—**图** 14 **可以看出，本文方法在视觉效果上优于传统的** CycleGan **算法和** StarGan **算法，较好地保持了转换图像的清晰程度和表情细节；且在处理语音视频时，可很好地实现新面部表情序列与源音频的同步。**

图 12 **中，对于传统** CycleGan **结果图** 1、**图** 3 **和图** 4 **来说，可以大致看出转换后的人脸五官，但是眼睛和嘴等关键位置的表情变化出现了模糊不清的现象；对于传统** CycleGan **结果图** 2 **来说，图像难以清晰显示局部特征，导致整张人脸清晰度过低。**StarGan **结果图** 1—**图** 4 **中的图像同样出现了局部细节模糊的现象，图像清晰度不高。本文方法有效保护了关键位置的图像清晰程度，保留了大量的表情细节。**

图 13 **中，传统** CycleGan **结果图** 1 **和图** 4 **色彩异常，干扰噪声信息较多；而传统** CycleGan **结果图** 3 **中的人脸五官出现了较为严重的扭曲，视觉效果较差。**StarGan **结果图** 1 **和图** 3 **中嘴巴部分出现了明显的扭曲和模糊，而** StarGan **结果图** 2 **和图** 4 **同样在眼睛等局部位置出现了伪影。**

图 14 **中，传统** CycleGan **结果图同样出现了图像质量较低的情况，有明显的图像扭曲、伪影以及画面不清楚的问题。对于** StarGan **结果图来说，图像也出现了局部位置模糊不清的现象。本文方法合成的表情动画，较好地提高了目标图像序列的质量，解决了图像模糊和细节缺失的问题，保证了新表情的真实自然。**

结束语 **本文提出了一种改进的** CycleGan **表情映射模型，通过在原有的循环一致损失函数中引入协方差约束项，有效避免了传统** CycleGan **表情转换过程中出现的大量局部细节丢失、图像扭曲和模糊不清的问题，可生成高精度的目标人**

脸表情图像。本文提出了基于分区域训练、以改进的 CycleGan 模型为核心的表情合成方法。所提方法不仅增强了生成表情动画的真实感，而且具有较好的稳定性和鲁棒性。但是，本文方法无法处理输入图片中的极端头部姿态，如何进一步提高本文算法对各种不同条件下获取的人脸表情数据的普遍适应性将是我们今后工作的研究重点。

参 考 文 献

[1] ZHU J，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). 2017：2242-2251.

[2] PIGHIN F，HECKER J，LISCHINSKI D，et al. Synthesizing Realistic Facial Expressions from Photographs[C]// Proceedings of the ACM SIGGRAPH Conference on Computer Graphics. 1998：75-84.

[3] BLANZ V，BASSO C，VETEER T，et al. Reanimating Faces in Images and Video[C]// European Association for Computer Graphics. 2003：641-650.

[4] VLASIC D，BRAND M，PFISTER H，et al. Face Transfer with Multilinear Models[J]. ACM Transactions on Graphics，2006，24(3)：426-433.

[5] LV P，XU M L. Expression of Face Expre-ssions Unrelated to Expression Database[J]. Journal of Computer-Aided Design & Computer Graphics，2016，28(1).

[6] PASQUARIELLO S，PELACHAUD C. GRETA：A Simple Facial Animation Engine[M]// Soft Computing and Industry. London：Springer，2002.

[7] ZHANG Q，LIU Z，GUO B，et al. Geometry-Driven Photorealistic Facial Expression Synthesis[J]. IEEE Transactions on Visualization & Computer Graphics，2005，12(1)：48-60.

[8] JOSHI P，TIEN W C，DESBRUN M，et al. Learning controls for blendshape based realistic facial animation[C]// Proceedings of ACM SIGGRAPH Eurographics Symposium on Computer Aimation. 2003：187-192.

[9] PARK B，CHUNG H，NISHITA T，et al. A feature-based approach to facial expression cloning：Virtual Humans and Social Agents[J]. Computer Animation and Virtual Worlds，2005，16(3/4)：291-303.

[10] JOSHI P，TIEN W C，DESBRUN M，et al. Learing Cnotrols for Blend Shape Based Realistic Facial Animation[C]// ACM Transactions on Graphics. 2006：426-433.

[11] GARRIDO P，VALGAERTS L，REHMSEN O，et al. Automatic face reenactment[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos：IEEE Computer Society Press，2014：4217-4224.

[12] WESIE T，BOUAZIZ S，LI H，et al. Realtime performance-based facial animation[J]. ACM Transactions on Graphics，2011，30(4)：1.

[13] CAO C，WENG Y，LIN S，et al. 3D shape regression for real-time facial animation[J]. ACM Transactions on Graphics，2013，32(4)：1.

[14] HUANG X Q，LIN Y X，SONG M L. Three-dimensional facial expression synthesis method based on nonlinear joint learning [J]. Journal of Computer-Aided Design & Computer Graphics，2011，23(2).

[15] WILLIAMS L. Performance-driven facial animation [C]// ACM SIGGRAPH Computer Graphics. 1990：235-242.

[16] YANG F，WANG J，SHECHTMANE，et al. Expression flow for 3D-aware face component transfer[J]. ACM Transactions on Graphics，2011，30(4)：1.

[17] PEREZ P，GANGNET M，BLAKE A. Poisson image editing [J]. ACM Transactions on Graphics，2003，22(3)：313-318.

[18] BITOUK D. Face Swapping：Automatically Replacing Faces in Photographs[J]. ACM SIGGRAPH，2008，27(3)：1-8.

[19] DALE K，SUNKAVALLI K，JOHNSON M K，et al. Video face replacement[J]. ACM Transactions on Graphics，2011，30(6)：1.

[20] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative Adversarial Nets[C]// International Conference on Neural Information Processing Systems. MIT Press，2014：2672-2680.

[21] KAZEMI V，SULLIVAN J. One Millisecond Face Alignment with an Ensemble of Regression Trees[C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2014：1867-1874.

[22] JOHNSON J，ALAHI A，LI F F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[M]// Computer Vision-ECCV 2016. Springer International Publishing，2016：694-711.

[23] IOFFE S，SZEGEDY C. Batch normalization：accelerating deep network training by reducing internal covariate shift[C]// International Conference on International Conference on Machine Learning. JMLR. org，2015.

[24] ISOPLA P，ZHU J Y，ZHOU T，et al. Image-to-Image Translation with Conditional Adversarial Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2017.

[25] KINFMA D P，BA J. Adam：A Method for Stochastic Optimization[J]. arXiv：1412. 6980v8，2014.

[26] HAQ S，JACKSON P J. Multimodal emotion recognition[M]// Machine Audition：Principles，Algorithms and Systems，2010，17：398-423.

[27] CHOI Y，CHOI M，KIM M，et al. StarGAN：Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation[C]// CVPR. 2017.

[28] ABADI M，AGARWAL A，BARHAM P，et al. Tensorflow：Large-scale machine learning on heterogeneous distributed systems[J]. arXiv：1603. 04467，2016.

YE Ya-nan，born in 1994，master，postgraduate. Her main research interests include computer animation and digital image processing.

CHI Jing，born in 1980，Ph.D，associate professor，postgraduate supervisor. Her main research interests include computer animation，geometric shape，and medical image processing.