班 级 _____

学 号 _____

西安电子科技大学

本科毕业设计论文

题　　目　基于单幅图像的面部表情生成算法研究

学　　院　人工智能学院

专　　业　智能科学与技术

学生姓名 _____

导师姓名 _____

摘 要

人脸表情识别算法一般需要大量的训练数据，而现有的数据库表情种类和数量较少。针对此问题，本文提出了基于单幅图像的面部表情生成算法，对面部表情的种类和数量进行扩充。

受启发于不同生成模型的结构和特性，本文将 GANimation 和 SinGAN 两种生成模型相结合，设计了一种新的无监督表情生成算法 SinGANimation，实现了基于单幅表情图像的面部表情数据生成过程。该算法首先通过 GANimation 进行单个 AU 变换、多个 AU 连续变换、多个 AU 离散变换等操作，对图像的表情种类进行扩充，得到初步结果再输入 SinGAN，进行再生成操作，增加图像的数量。其中，本文提出的算法将图像下采样后再输入，解决了 SinGAN 原有模型人脸生成失真的问题，保证了人脸的高度结构性。随后，本文对提出算法的生成结果进行了定性和定量分析。通过与其他经典模型对比，发现本文提出的算法既可以生成连续自然的表情也可以生成离散情绪的表情，其画质更加真实清晰。在多个数据集上训练，均达到不错的效果，证明算法的鲁棒性良好。此外，测试实验中也进行了 AMT 真伪用户测试和单幅图像 FID 测量，得到的混淆率接近 50%，生成图像与真实图像的深度特征分布之间的偏差接近 0.1，表明两种图像高度相似。最后，本文对提出的算法模型的优缺点进行分析，计划将算法应用于扩充人脸表情数据库，视频序列等商业和科研工作中。

关键词：单幅图像　表情生成　GANimation　SinGAN

摘　要

ABSTRACT

Generally, facial expression recognition algorithms need a large amount of training data, but the expression types and quantities of the existing databases are limited. To solve the problem, a facial expression generation algorithm based on a single image is proposed in this paper, which effectively expands the types and quantity of expressions.

Inspired by the structure and characteristic of various generation models, this paper combines the GANimation and SinGAN models together to develop a new fully unsupervised generation algorithm expression, called SinGANimation, which achieves the generation of facial expression images via utilizing only one expression image. The proposed method operates the single AU transformation, multiple AU continuous transformation, multiple AU discrete transformation of GANimation, et al, and it expands the expression types of the image and inputs the results to SinGAN for regeneration operation to increase the number of images. In order to solve the problem that generated face may be distorted by SinGAN, the proposed method adds the downsampling strategy for input images, which effectively improves the problem of face distortion in SinGAN and ensures the high facial structure information. Then, this paper makes qualitative and quantitative analyses for the generated results obtainted by the proposed method. Comparing with other classical generation models, it is found that the proposed method can generate both continuous, natural expressions and discrete emotional expressions, and the quality of pictures are more real and clear. The training on multiple data sets has achieved good results and proved the robustness of the algorithm. Additionally, this paper conducts the AMT true and false user test and FID measurement of a single image. The confusion rate are close to 50%. The deviation between the depth feature distribution of the generated image and the real image are close to 0.1, which indicates that the two images are highly similar. Finally, this paper analyzes the advantages and disadvantages of the algorithm, and plan to apply the algorithm to expand the facial expression database, video sequence and other commercial and scientific works.

Key words: single image facial expression generation GANimation SinGAN

ABSTRACT

目 录

第一章 绪 论

1.1 人脸表情概述

人的面部表情在社交中极为重要。通常，交流涉及言语和非言语。非语言交流是指人与动物之间通过眼神交流，手势，面部表情，肢体语言和非语言进行的交流。非语言交流是通过面部表情表达的。面部表情是更大范围交流的微妙信号，它能够有效地传达非语言信息及情感的交流，从而辅助听者推断说话人的意图。

人脸表情是传播人类情感信息与协调人际关系的重要方式，据心理学家A.Mehrabia 的研究表明[1]，在人类的日常交流中，通过语言传递的信息仅占信息总量的 7%，而通过人脸表情传递的信息却达到信息总量的 55%。尤其，当说话人在试图掩盖内在情绪时，面部表情的细微变化是无法隐藏和无法抑制的，它所传达的信息暗含了潜在的个体行为信息，是与人类情感、精神状态、健康状态等诸多因素相关的一种复杂的表达方式。

在 20 世纪，Ekman 和 Friesen 对人脸表情进行研究[2]，得出人类的七个基本情感:幸福，惊讶，愤怒，悲伤，恐惧，厌恶和中立。他们建立了不同种类表情的人类面部表情数据库，详细记录每种表情的面部变化，比如皱鼻、嘴角拉伸等动作变化，这便是 1976 年创造的“面部运动编码系统”（FACS，Facial Action Coding System）。FACS 包含 44 个面部动作单元（AU，Action Unit），例如抬高眉毛、眼睛变化等用作描述人面部局部表情的变化。AU 可以精确细致的描述人的面部表情，但其标注的成本高，耗时长，例如：标注一个人眼部 AU，需要标注员长达 30 分钟的时间。所以，现在的人脸表情数据库的采集对象和面部表情都相对有限。

1.2 研究意义与目的

如今，人脸识别技术发展迅猛，应用市场和用户需求大，而人脸面部表情识别作为人脸识别技术中的一个重要组成部分，对公共安全、安全驾驶、智能医疗、测谎技术、智慧课堂等领域具有非常重要的商业贡献，对学术界也有很大的研究意义。比如，在安全驾驶中，通过识别司机的眼部表情，判断司机是否为疲倦状态，若是便发出安全警告，减少安全隐患；在智能医疗场景中，根据患者的微表情，评估患者的精神健康状态； 在智慧课堂上，老师可以根据智能摄像头采集学生们的面部

表情，提醒走神和不注意听讲的同学集中注意力，为产生困惑表情的同学答疑解惑等。

如上文所述，现有的人脸表情数据库有以下两个不足：第一，因表情均为人为收集，并非真实环境下的自然表情，而且面部表情受许多因素的限制，例如年龄，性别，肤色等，故致表情种类单一，幅度夸张。第二，表情数据量小，难以满足人脸表情识别算法的的数据需求。如今的深度学习发展迅猛，许多人脸识别算法也纷纷采用深度学习框架。深度学习训练模型需要大规模的数据，而现在的表情数据库容量有限，不足以支撑基于深度学习框架的人脸表情识别算法。

基于以上人脸表情数据库种类和数量有限的缺点，本文将研究在基于单幅的人脸表情数据下，生成数量庞大，种类繁多，面部自然的表情数据用于人脸表情识别算法研究。

1.3 内容安排

第一章为绪论部分，主要介绍面部表情的相关知识和本文的研究意义与目的。

第二章主要介绍生成对抗网络 GAN 的基础内容，并给出一些常见的 GAN 模型。

第三章详细介绍 SinGAN 相关基础，原理方法，实验结果并总结 SinGAN 的优缺点。

第四章从 GANimation 相关基础，模型架构和方法以及优缺点分析等几个部分进行描述，为下章的实验奠定理论基础。

第五章主要讲解基于 SinGAN 和 GANimation 的 SinGANimation 面部表情生成算法实验，得出了多种实验结果，并将此模型与其他模型对比，进行定性和定量分析。

第六章是本文的总结与展望。对本文已做的工作进行梳理总结，并提出该算法之后需要改进优化之处。

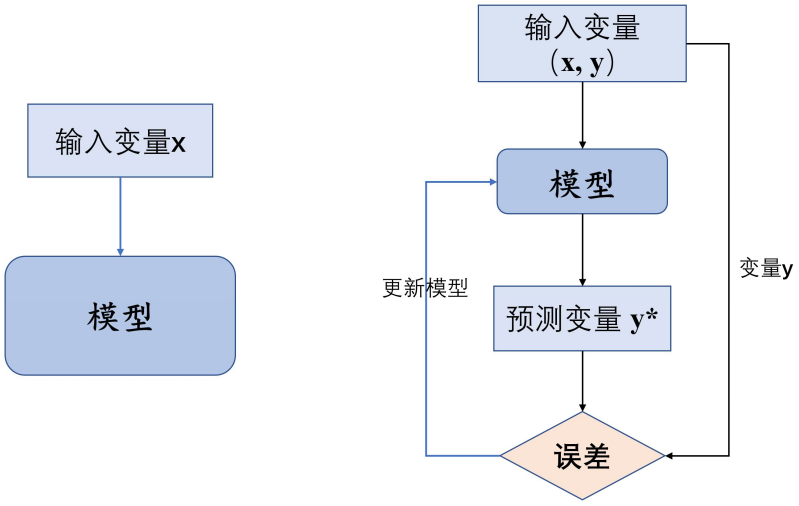
第二章　生成对抗网络（GAN）

生成对抗网络，简称 GAN，是一种使用深度学习进行生成式建模的方法。所以，本章在介绍生成对抗网络模型的同时，需提及其前身生成建模，便于读者理解，深入浅出地介绍生成对抗网络的理论基础及其延伸拓展模型。

2.1　生成对抗网络的理论基础

2.1.1　生成式模型

生成式模型（Generative Model）是机器学习中的无监督学习任务，可用于生成或输出可能从原始数据集中得出的新示例。图2.1展示了无监督学习和有监督学习的流程图。从图2.1中可看出：无监督学习形式是仅给模型输入变量 x，没有任何输出变量 y。而有监督学习形式的训练数据集，每个样本均具有输入变量 x 和输出类别标签 y。通过预测输出并校正模型以使输出更像预期输出来训练模型。



（a）无监督学习　　（b）有监督学习

图 2.1 无监督学习与有监督学习流程图

生成式模型则会对 x 和 y 的联合分布 $p(x,y)$ 建模，然后通过贝叶斯公式来求得 $p(y_i|x)$，然后选择特定的 y_i 使 $p(y_i|x)$ 最大，即

$$\begin{aligned}\arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y)\end{aligned} \tag{2-1}$$

常用的生成式模型有：朴素贝叶斯模型，高斯混合模型GMM，新马尔可夫模型HMM等。

2.1.2 生成对抗网络模型

生成对抗网络（GAN）是基于深度学习的生成模型。一般而言，GAN 是用于训练生成模型的模型架构，最常见的是在该架构中使用深度学习模型。2014 年,Ian Goodfellow 等人首次给出了 GAN 架构[3]。GAN 模型结构包括两个子模型：生成网络G（Generator）和判别网络D（Discriminator），它们的功能如下所示：

- 生成网络G负责接收一个随机的多维向量噪声z，由其生成的图像为G（z)。
- 判别网络D负责判别输入图像的真伪。假设m为输入图像，则输出D(m)是输入图像m为真的概率。若$D(x)$为 1，则m为真实图像；若D(m)为 0，则m为不真实图像。

在 GAN 的训练过程中，生成网络G试图生成真实的图像来欺骗判别网络D，判别网络D试图将生成网络G生成的图像和真实的图像区分开来。如此，两者便形成了动态的零和博弈。当 GAN 训练到最好状态时，生成图像G(z)足以骗过D，而D难以分辨生成图像G（z)的真伪，所以，生成图像为真实图像的概率D(G(z))$=0.5$。GAN 的数学原理为

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{2-2}$$

式中x表示真实图像，z表示多维向量噪声，而G（z)表示生成网络G的生成图像。判别函数D（x)表示判别网络D鉴别真实图像的概率，其值接近或等于 1。而D（$G(z)$)为判别网络D鉴别生成图像G（z)为真实的概率。

生成网络G的目标为D（$G(z)$)最大化，此时$V(D,G)$会变小，所以对于G求最小($\min_G$)。判别网络D的目标为D（x)最大化，D（$G(z)$)最小化，此时$V(D,G)$会变大，所以对于D求最大（$\max_D$）。具体 GAN 的训练过程，如图 2.2 所示。

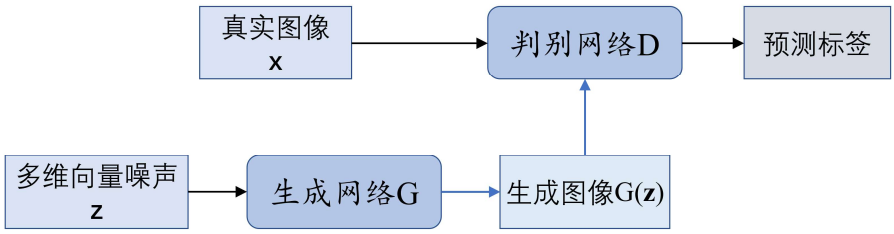


图 2.2 GAN 训练过程流程图

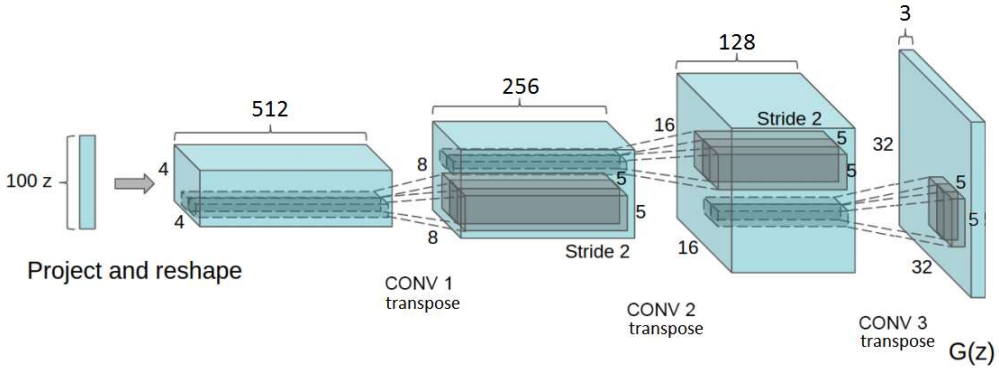


图 2.3 DCGAN 的生成网络结构图

2.2　生成对抗网络的常用模型

2.2.1　深度卷积生成对抗网络（DCGAN）

深度卷积生成对抗网络[4]（DCGAN）是由 Alec Radford 等人提出的 GAN 模型，该网络能很有效地将监督学习中的 CNN 和无监督学习中的 GAN 结合在一起。DCGAN 可跨一系列数据集进行稳定的训练，并允许训练更高分辨率和更深入的生成模型。DCGAN 的生成网络结构如图 2.3 所示：

具体地，DCGAN 在 GAN 的基础上做了如下几点变化：（1）将池化层的卷积进行替换，其中，在判别网络上用跨步卷积替换，在生成网络上用部分跨步卷积替换；（2）在判别网络和生成网络中都使用 batchnorm，这有助于解决初始化差的问题，帮助梯度传播到每一层，并防止生成网络把所有的样本都收敛到同一个点。直接将 BN 应用到所有层会导致样本震荡和模型不稳定，通过在生成网络输出层和判别网络输入层不采用 BN 可以防止这种现象；（3）删除完全连接的隐藏层以进行更深层次的体系结构；（4）在生成网络中的所有层上使用 ReLU 激活函数，但输出除外，后者使用 Tanh；（5）在判别网络的所有层上使用 LeakyReLU 激活。

基于这些改进，DCGAN 解决了 GAN 生成网络产生无意义输出的问题，具体贡献如下：

- 此模型对 GAN 的体系结构进行约束，可以通过稳定地训练使其更趋于收敛。
- DCGAN 训练大量没有标签的图像时，特征提取非常有效，这既来自于生成网络也有判别网络（主要是判别网络）。由于 DCGAN 出色的特征提取，它可用于更高级别的监督任务，例如图像分类。
- 对 GAN 学习到的 filter 进行了定性的分析。
- DCGAN 的生成网络具有很好的矢量计算特性，可以轻松操纵生成样本的许多语义质量。

2.2.2 条件生成对抗网络（CGAN）

条件生成对抗网络[5]（CGAN）是由 Goodfellow Ian 等人提出的一种带有条件约束的 GAN，在生成网络和判别网络中均引入条件变量 y，根据补充信息 y 对网络进行约束，指导生成过程。条件变量 y 可以基于多种信息，比如类别标签，用于图像修复的部分数据，来自不同模态的数据，这样可以看做 CGAN 是把纯无监督的 GAN 改进为有监督的网络。如图 2.4 所示，通过将补充信息 y 传送给生成网络和判别网络，作为输入层的一部分，从而实现条件 GAN。在生成网络中，随机噪声 z 和条件信息 y 联合组成了联合隐层表征。对抗训练框架在隐层表征的组成方式方面相当地灵活。类似地，条件 GAN 的目标函数是带有条件概率的二元极小极大值博弈（two-player minimax game）：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x \mid y)] + E_{z\sim p_z(z)}[\log(1 - D(G(z \mid y)))] \quad (2\text{-}3)$$

2.2.3 循环生成对抗网络（CycleGAN）

一般的 GAN 面向一个域的数据，而循环生成对抗网络 CycleGAN[6]实现的是两个域的数据迁移。CycleGAN 是一个 A→B 单向 GAN 加上一个 B→A 单向 GAN。两个 GAN 共享两个生成网络，然后各自带一个判别网络，所以加起来总共有两个判别网络和两个生成网络。CycleGAN 本质上是两个镜像对称的 GAN，构成了一个环形网络。

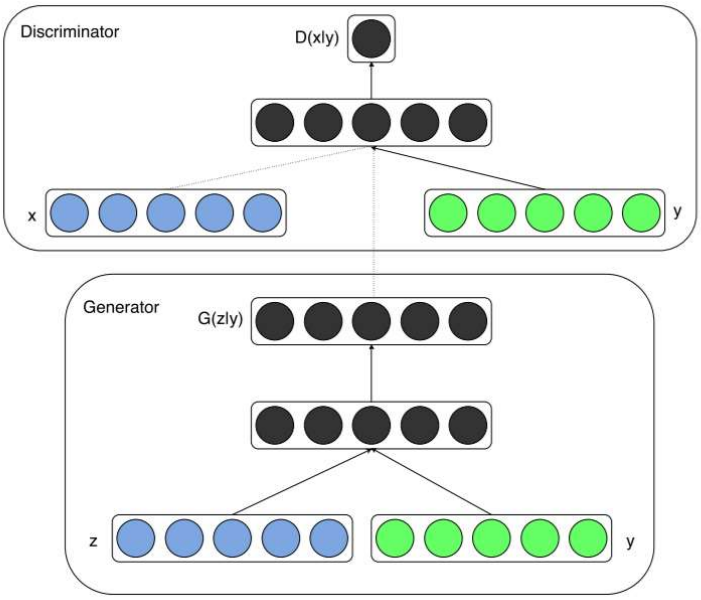


图 2.4 CGAN 网络结构图

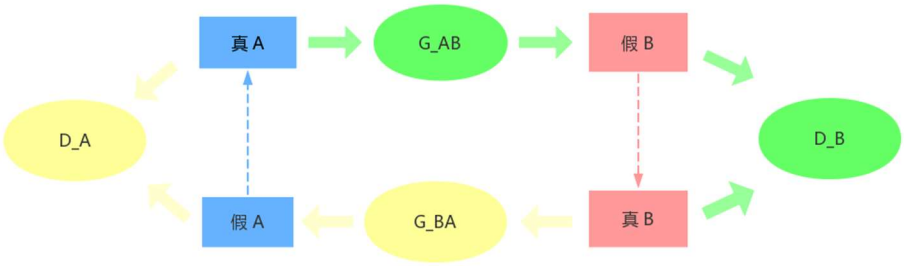


图 2.5 CycleGAN 网络结构图

如图 2.5 所示，真图像 A 经过生成网络G_{AB}表示为假图像 B，把假图像 B 视作为真图像 B；同理可得，真图像 B 经过生成网络G_{BA}表示为假图像 A，把假图像 A 视作为真图像 A。

一个单向 GAN 有两个损失函数，而 CycleGAN 加起来总共有四个损失函数。对于判别网络 A：

$$L_{D_A} = E_{x \in P_A} \log D_A(x) + E_{x \in P_{B2A}} \log(1 - D_A(x)) \qquad (2\text{-}4)$$

对于判别网络 B：

$$L_{D_B} = E_{x \in P_B} \log D_B(x) + E_{x \in P_{A2B}} \log(1 - D_B(x)) \quad (2\text{-}5)$$

对于生成网络 BA：

$$L_{G_{BA}} = E_{x \in P_{B2A}} \log D_A(x) + \lambda E_{x \in P_A} \left\| x - G_{BA}(G_{AB}(x)) \right\| \quad (2\text{-}6)$$

对于生成网络 AB：

$$L_{G_{AB}} = E_{x \in P_{A2B}} \log D_B(x) + \lambda E_{x \in P_B} \left\| x - G_{AB}(G_{BA}(x)) \right\| \quad (2\text{-}7)$$

对于生成网络添加重构误差项，如同对偶学习，能够引导两个生成网络更好地完成编码和译码的任务，而两个判别网络则起到纠正编码结果符合某个域的风格的作用。

第三章 SinGAN

SinGAN：Learning a Generative Model from a Single Natural Image[7]，即从单张自然图像中学习的生成模型。此模型通过使用一种专门的多尺度对抗训练方案，对多个尺度上学习子图像块数据。然后，它可以用来生成新的逼真的图像样本，在创建新的对象配置和结构时，保持原始的子图像块的分布。本章将主要介绍 SinGAN 模型的基本原理，模型细节及该模型的优缺点和部分实验结果。

3.1 SinGAN 相关基础

3.1.1 单项深度模型

现有的几项研究提出将深度模型过度拟合到单个训练实例中，然而，这些方法是为特定的任务而设计的(如：超分辨率，纹理扩展等)。Shocher 等人[8]首先为单个自然图像引入了基于内部GAN的模型，并在重新定向的背景下进行了说明。然而，它们的生成取决于输入图像，即将图像映射到图像，而不是用来绘制随机样本。相比之下，SinGAN 的框架是纯生成的，即将噪声映射到图像样本，因此适合许多不同的图像处理任务。

如图 3.1 所示，无条件的单图像 GANs 仅在纹理生成的环境中被探索过。这些模型在对非纹理图像进行训练时，并不能生成有意义的样本。而 SinGAN 的方法并不局限于纹理，还可以处理一般的自然图像。实际上，用于纹理生成的单一图像模型并不适用于处理自然图像，但是本文提出的可以生成包含复杂纹理和非重复全局结构的真实图像样本。

3.1.2 图像处理的生成模型

在许多不同的图像处理任务，基于 GAN 模型的研究已经证明了对抗性学习的能力，例如：交互式的图像编辑和其他图与图之间的翻译任务。然而，已有的方法大部分都是在具体的数据集上训练，将生成条件设置为另一个输入信号，SinGAN 也是如此。SinGAN 并不着重于提取一般的同类图像特征，而是通过不同来源的训练数据——单幅自然图像的多尺度的全部重叠图像子块。SinGAN 展示了一个强大

图 3.1 SinGAN 与单个图像纹理生成

的生成模型是可以从上述训练数据中学习，并用于多种图像处理任务，下面将详细介绍 SinGAN 模型的基本原理及其应用。

3.2 SinGAN 模型的基本原理

3.2.1 概述

SinGAN 模型的主要目的是学习一个无条件生成模型，它可以捕获单个训练图像的内部统计信息。这个任务在概念上与传统的 GAN 设置类似，只是这里的训练样本是单个图像的子块，而不是来自数据库的整个图像样本。

SinGAN 选择超越纹理生成，并处理更综合的自然图像。这需要捕捉在很多不同尺度下的复杂图像结构分布。例如，SinGAN 想要捕获全局属性，比如图像中大型物体的排列和形状(顶部的天空，底部的地面)，以及图像细节和纹理信息。为了实现这一目标，SinGAN 的生成式结构, 如下图 3.2 所示，包含一个多层次的子块-GANs (马尔可夫链的判别网络)[9],其中每个负责捕捉不同规模 x 的子块分布。GANs 的感受野比较小，而且容量有限, 这些特点阻止它们记忆单一的图像. 同时，相似的多尺度的架构一直在探索传统 GAN 设置，SinGAN 是第一个从单一图像内部学习探索它的网络模型。

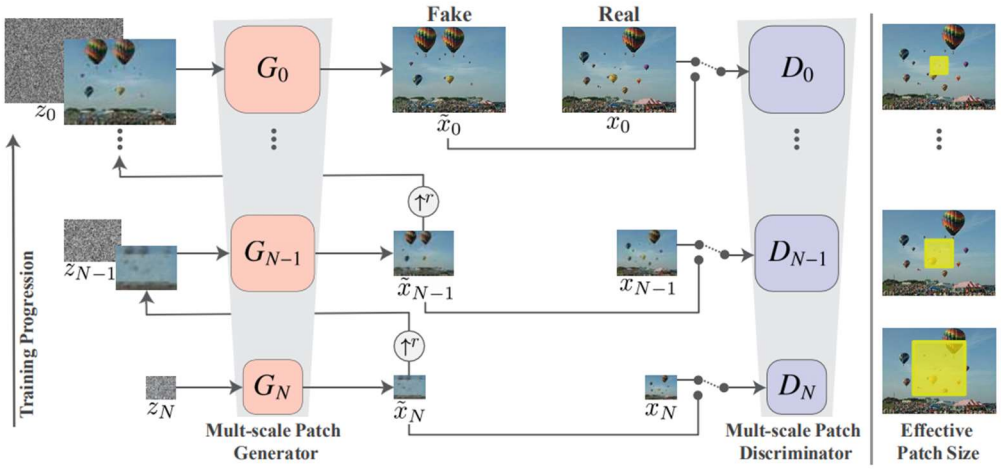


图 3.2 SinGAN 的多尺度传递途径

SinGAN 模型是由一个 GANs 的金字塔结构组成，其中训练和传递都是尺度由大到小的方式完成。在每个尺度，生成网络 G_n 学习生成图像样本，判别网络 D_n 无法将生成图像的所有重叠图像块与降采样训练图像中的图像块 x_n 区分开来。当沿着金字塔向上移动时，有效的图像块尺寸不断减小(在原始图像中用黄色标记以供说明)。输入到生成网络 G_n 是随机噪声图像 z_n。对之前尺寸 $\tilde{x}_n$ 生成的图像进行上采样至当前的分辨率（除了纯生成的最大尺度）。当前尺度的生成过程涉及到所有生成网络 $\{G_N,\cdots,G_n\}$ 和噪声图谱 $\{z_N,\cdots,z_n\}$ 的参与。

3.2.2　多尺度结构

SinGAN 的模型由一个金字塔状生成网络 $\{G_0,\cdots,G_N\}$ 组成，对 x 的图像金字塔 $\{x_0,\cdots,x_N\}$ 进行训练，其中当 $r>1$，x_n 是一个因子 r_n 的 x 的下采样版本。每个生成网络负责生成真实的图像样本，即关于对应图像中的子块分布。通过对抗训练，实现 G_n 学习，欺骗相关的判别网络 D_n，判别网络 D_n 试图将生成样本中的子块与 x_n 中的子块区分开来。

通常，图像样本的生成从最大的尺度开始，依次通过所有生成网络，直到最小的尺度，并在每个尺度都输入噪声。所有的生成网络和判别网络都有相同的感受野，因此在生成过程中捕获的结构尺寸都在减小。在最大尺度上，生成结果是纯生成的，即 G_N 将空间高斯白噪声 z_n 映射到图像样本 $\tilde{x}_N$，即

$$\tilde{x}_N = G_N(z_N) \tag{3-1}$$

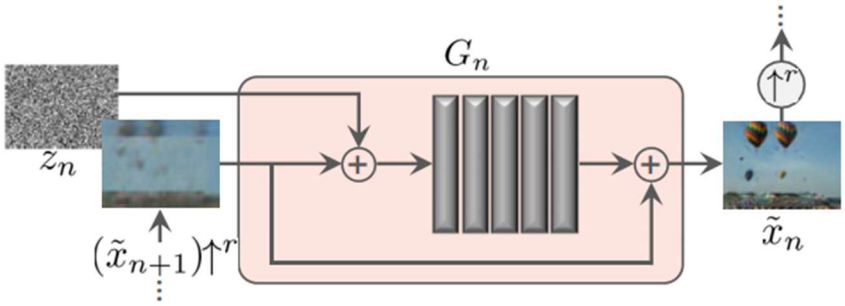


图 3.3 单规模迭代

一般而言，这一层的有效感受野一般是图像高度的 1/2，因此G_N可以生成图像的总体布局和对象的全局结构。每个生成网络在更小的尺度（$n<N$）上添加之前尺度没有生成的细节。因此，除了空间噪声z_n外，每个生成网络G_n还接受较大尺度图像的上采样版本，即

$$\tilde{x}_n = G_n(z_n, (\tilde{x}_{n+1})\uparrow^r)，\ n<N \tag{3-2}$$

实际上，所有的生成网络都具有类似的架构，如图 3.3 所示。在被输入到一系列卷积层之前，噪声z_n要被加在图像$(\tilde{x}_{n+1})\uparrow^r$。这确保了 GAN 不会忽略噪声，如同随机条件方案中经常发生的情况。卷积层的作用是生成缺失的细节$(\tilde{x}_{n+1})\uparrow^r$（残差学习）[10]，即$G_n$执行如下操作：

$$\tilde{x}_n = (\tilde{x}_{n+1})\uparrow^r + \psi_n(z_n + (\tilde{x}_{n+1})\uparrow^r) \tag{3-3}$$

其中，ψ_n是一个有 5 个卷积层的完全卷积网络。SinGAN 在最大的尺度上从每个块的 32 个内核开始，然后内核数量每 4 个尺度增加 2 倍。因为生成网络是全卷积网络，所以 SinGAN 可以在测试时生成任意大小和宽高比的图像(通过改变噪声图像的规模)。

在每个尺度n上，对之前尺度的图像$\tilde{x}_{n+1}$向上采样并输入噪声图谱z_n中。其结果输入至 5 个卷积层，输出是一个补充到$(\tilde{x}_{n+1})\uparrow^r$的残差图像，即生成网络$G_n$的输出$\tilde{x}_n$。

3.3　SinGAN 模型的应用

SinGAN 在许多图像处理任务中都有应用，主要应用为：超分辨率、图画到图像的画风迁移、图像调和、图像编辑和单一图像生成动画。应用基于 SinGAN 原始模型，因为 SinGAN 只能生成与训练图像具有相同子块分布的图像，所以可以通过在 $n < N$ 的某个尺度将图像(可能是下采样的版本)注入到生成网络金字塔中，并通过生成网络将其前馈，使其子块分布与训练图像的子块分布匹配，从而进行操作。不同的输入规模导致不同的效果。

图 3.4 SinGAN 模型的应用展示

3.3.1　超分辨率

SinGAN 将输入图像的分辨率提高了 s。SinGAN 在低分辨率(LR)图像上训练模型,得出重构损失权重为 $\alpha = 100$ 和生成网络金字塔的比例因子 $r = \sqrt[k]{s}$，$k \in N$。在自然场景不同尺度中，小型结构往往反复出现，因此在测试时，SinGAN 通过一个 r 因子对 LR 图像进行上采样，并将其连同噪声输入最后一个生成器 G_0。SinGAN 重复 k 次以获得最终的高分辨率输出，示例结果如图 3.5 所示。从对比结果可以看出，SinGAN 重建的视觉质量超过了目前最先进的内部生成方法，也超过了以最大信噪比为目标的外部生成方法。SinGAN 尽管只需要一张图像，但结果可以与外部训练的 SRGAN[11]方法相媲美。在 BSD100 数据集[12]上，基于失真程度(RMSE)和感知质量(NIQE[13])两个指标比较了 5 种方法的性能，结果展示在表 3.1 中，注：这

表 3.1 超分辨率对比

方法 / 性能指标	外部训练方法		内部训练方法		
	SRGAN	EDSR	DIP	ZSSR	SinGAN
RMSE	16.34	12.29	13.82	13.08	16.22
NIQE	3.41	6.50	6.35	7.13	3.71

图 3.5 超分辨率效果对比。

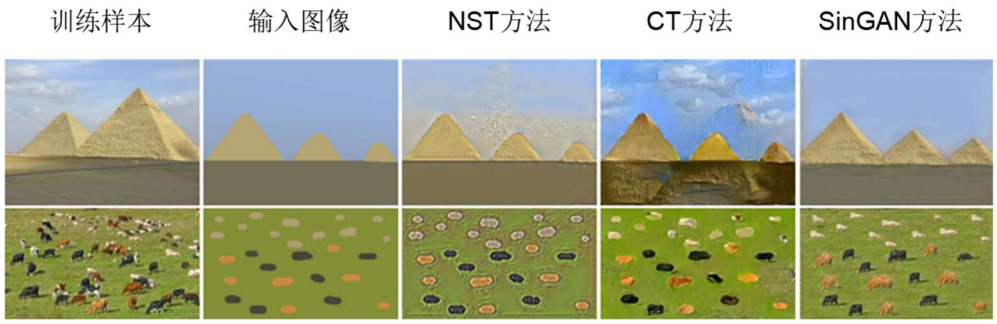


图 3.6 画风迁移效果对比

两个指标本质上是相互冲突的。从表 3.1 中所示的结果可以看出：SinGAN 擅长感知，其 NIQE 数值仅略低于 SRGAN，但 RMSE 数值要高于 SRGAN。

另外，当 SinGAN 被训练在一个低分辨率的图像上时，可以进行超分辨率操作。这是通过不断迭代对图像进行采样，并将其输入到 SinGAN 的最小尺度的生成网络来实现的。可见，SinGAN 的视觉质量优于 SOTA 的内部训练方法 ZSSR 和 DIP，也与在大规模集合上进行外部训练的 SRGAN 方法的训练结果相近。括号中显示了相应的 PSNR 和 NIQE 的数值。

3.3.2 图画到图像的画风迁移

图画到图像的画风迁移即将剪贴画转换成逼真的图像。这是通过对剪贴画图像向下采样，并将其输入至一个较大尺度(例如 $N-1$ 或 $N-2$)的生成网络来实现的。从图 3.6 可以看出，SinGAN 保留了画面的整体结构，真实地生成了与原图匹配的

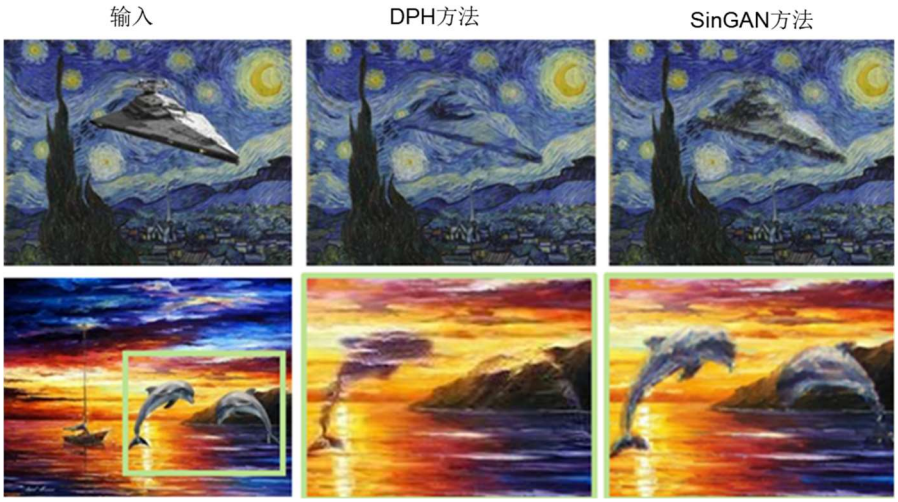


图 3.7 图像调和效果对比

纹理和高频信息。SinGAN 的画风迁移结果在视觉质量上要优于风格迁移（Syle Transfer）方法。在目标图像上训练 SinGAN，并在测试时将下采样的图画输入到较大生成网络中，生成图像保留了剪贴画的布局和一般结构，同时生成与训练图像匹配的真实纹理和精密细节。

3.3.3　图像调和

图像调和为将粘贴对象与背景图像真实地混合在一起。在背景图像上训练 SinGAN，并在测试时输入原始粘贴合成的下采样样本。SinGAN 将生成图像与原始背景相结合。从图 3.7 可以看出，SinGAN 模型对粘贴对象的纹理进行了裁剪以匹配背景，并且与其他图像调和方法相比，更好地保留了对象的结构。在 2、3、4 尺度下，粘贴对象的结构和转移背景纹理之间可以取得很好的平衡。SinGAN 模型能够保持粘贴对象的结构，同时调整其外观和纹理，而其他的协调方法过度混合对象与背景。

3.3.4　图像编辑

图像编辑为将图像区域复制并粘贴到其他位置，进行无缝衔接合成。将合成的下采样样本输入到较大尺度生成网络中。然后，将编辑区域的 SinGAN 的输出与原始图像结合起来，如下图所示，SinGAN 重新生成了精细的纹理，并无缝衔接了粘贴部分，产生了比 Photoshop 的 Content-Aware-Move 方法更好的效果。

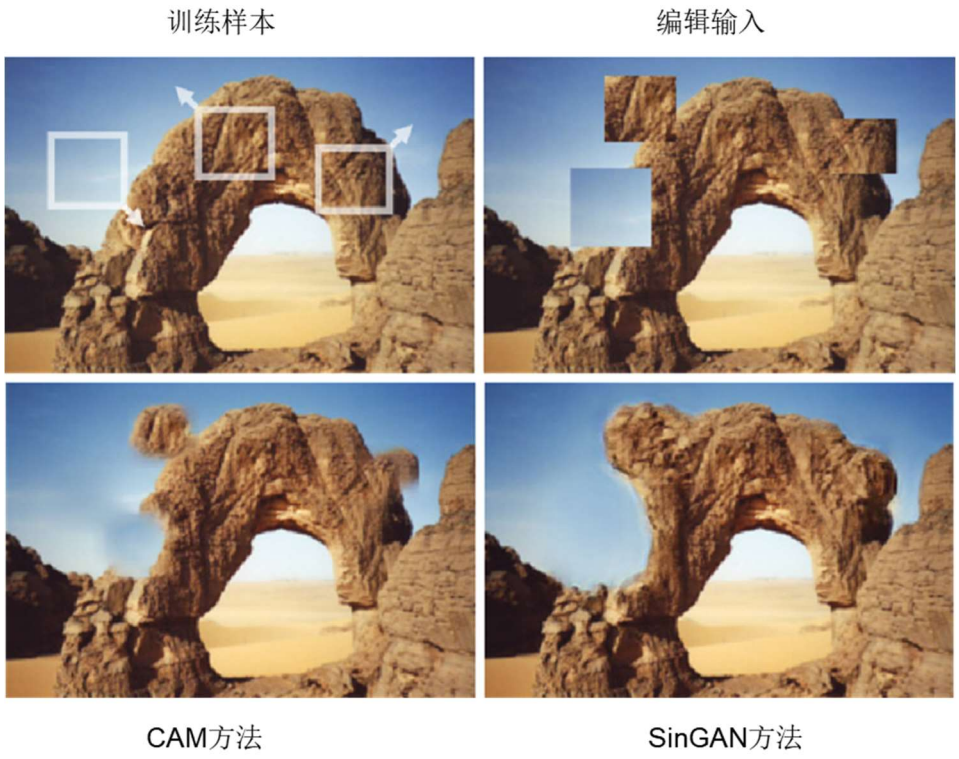


图 3.8 图像编辑效果对比

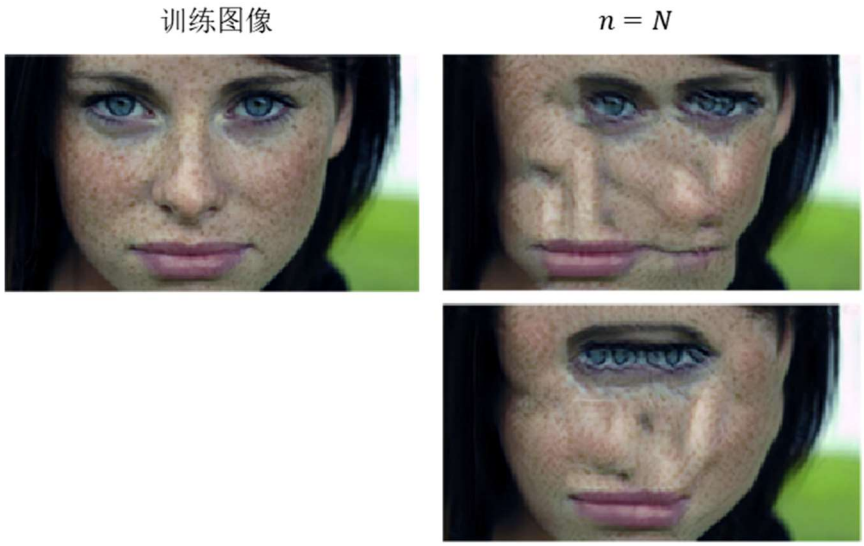


图 3.9 人脸图像训练 SinGAN 模型的生成效果

3.3.5 单一图像生成动画

单一图像生成动画为输入单一图像，生成真实物体运动的短视频。自然图像往往包含重复的部分，这显示不同的同一动态对象的“快照”(例如，一群鸟的图像显示了一只鸟的所有飞行姿势)。使用 SinGAN，可以沿着图像中物体的所有形象流

形前进，从而将单一图像合成运动视频。对于许多类型的图像，真实效果是通过 z 空间中的随机漫步实现的，即在所有生成尺度中，第一帧画面由 z^{rec} 开始生成。

3.4　本章小结

SinGAN 模型是首次使用单张自然图像训练、非条件的生成式模型。SinGAN 模型生成的效果目前已经可以做到以假乱真，它可以生成新的具有真实感的图像样本，在保留了原始的图像块分布的基础上，创造了新的物体外形和结构。SinGAN 模型具有超越纹理和生成自然复杂图像的各种真实样本的能力，为广泛的图像处理任务提供非常强大的工具。

然而，SinGAN 模型也存在一定的局限性，这可能源于该模型是“单张图像训练”的设定，具体表现为：第一，当图像块差异较大时，容易产生不真实的现象，无法学到很好的分布。如图 3.9 所示，如果直接使用人脸图像训练 SinGAN，生成的图像失真严重。这个问题也是本文针对面部图像生成对 SinGAN 进行改进的出发点，目的是生成无失真更真实的人脸表情。第二，与外部训练生成方法相比，SinGAN 经过内部学习生成图像的内容语义多样性受到了限制，例如：如果训练图像是一只猫，模型不会生成不同猫品种的样本。

第四章 GANimation

鉴于 SinGAN 生成图像语义单一的局限性，以及本论文的目的是为了探索仅凭单张人脸图像便可生成多种语义图像的相关研究，因此在考虑 SinGAN 的同时也考虑到其他一些 GAN 模型，经过研究对比之后，以基于人脸动作单元调节表情且具有生成表情连续自然、较为清晰等特点的 GANimation 模型作为本文算法的另一种参考模型。因此，本章从 GANimation 相关基础，模型架构和方法以及优缺点分析等三个部分进行论述，为下一章本文提出算法的介绍奠定理论基础。

4.1 GANimation 模型的相关基础

4.1.1 非匹配的图像转换

在 GANimation 框架中，一些工作解决了使用非匹配训练数据的问题。在图像个别领域的边缘分布中，首次尝试应用依赖马尔科夫随机场先验的贝叶斯生成模型。其他模型则探索了利用变分自动编码器策略来增强 GANs。后来，一些模型应用了驱动系统生成变换样式映射的思想，而且没有改变原始输入图像内容。GANimation 方法更接近于那些利用循环一致性来保存输入和映射图像之间的关键特征的模型，比如 CycleGAN[6]、DiscoGAN[14]和 StarGAN[15]。

4.1.2 面部图像处理

人脸生成与编辑是计算机视觉和生成模型研究的热点。大多数的工作都是处理属性编辑的任务，试图修改诸如添加眼镜、改变头发颜色、性别交换和老化等属性类别。这些工作与 GANimation 最相关的是面部表情的合成。早期的方法是使用质量-弹簧模型来模拟皮肤和肌肉运动[16]。这种方法的问题是很难产生自然的面部表情，因为有许多细微的皮肤运动是很难用简单的弹簧模型渲染的。另一种思路是依赖于 2D 和 3D 的形态[17]，但在区域边界周围产生了强大的伪影，无法模拟光照变化。

最近的研究训练了能够处理自然环境下图像的高度复杂卷积网络。然而，这些方法都是基于离散的情感类别(例如，快乐、中性情绪和悲伤)。相反，GANimation

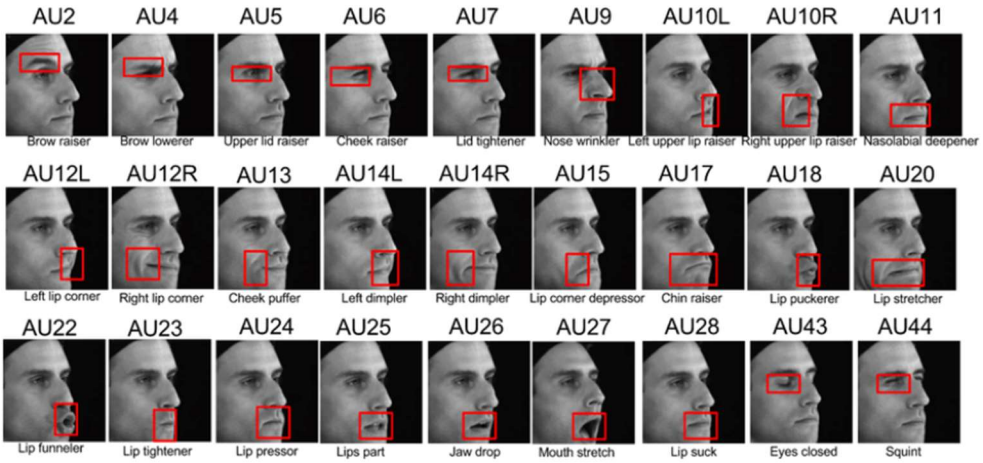


图 4.1 常见 AU 类别表示

模型恢复了皮肤和肌肉建模的想法，但将其整合到现代深度学习机制中。更具体地说，GANimation 学习了一个基于肌肉运动的连续嵌入 GAN 模型，允许在视频序列中生成大量基于人脸结构的面部表情以及平滑的面部运动转换。

4.1.3 人脸动作单元（AU）

人脸动作单元（AU）源于面部动作编码系统（FACS），是一种基于面部表情对人类面部运动进行分类的系统，而 AU 是人脸单个肌肉或一组肌肉的基本动作。人脸做出表情时，面部区域会有不同程度的变化，即多种 AU 会有一定的强度变化。常见的 AU 包括内侧眉头上扬，上眼睑上扬，嘴唇提起等，具体 AU 类别表示如图 4.1 所示。GANimation 数据集的标签便是基于 AU，使用表情向量来表示面部各区域不同程度的变化。通过调节表情向量使得 GANimation 模型输出不同程度的表情。表情向量如下所示：

$$y_r = \left(y_1, y_2, \ldots, y_N\right)^T \tag{4-1}$$

其中，N 为向量长度，y 表示 AU 运动强度，即 $y_i \in [0,1], i \in [1,N]$。

4.2 GANimation 模型架构和方法

4.2.1 待解决的问题

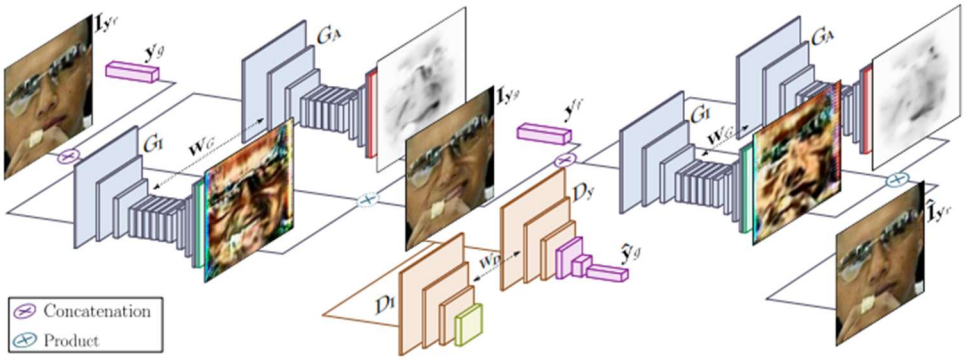


图 4.2 GANimation 模型的结构

定义一个在任意面部表情下捕获的输入 RGB 图像为$\mathbf{I}_{\mathbf{y}_r} \in \mathbb{R}^{H\times W\times 3}$。每个表情表达式都由 N 个动作单元$\mathbf{y}_r = (y_1,\ldots,y_N)^\top$决定，其中每个 y_n 表示第 n 个动作单元大小的归一化值，其值范围为从 0 到 1。由于这种连续的表现，自然插值可以在不同的表情之间，渲染的范围更加真实，面部表情更加光滑。

GANimation 的目标是学习一个映射 $\mathcal{M}$，将$\mathbf{I}_{\mathbf{y}_r}$转换成输出图像$\mathbf{I}_{\mathbf{y}_g}$条件下的动作单元目标$\mathbf{y}_g$，例如：映射为：$\mathcal{M}:\left(\mathbf{I}_{\mathbf{y}_r},\mathbf{y}_g\right)\rightarrow\mathbf{I}_{\mathbf{y}_g}$。为此，GANimation 对 $\mathcal{M}$ 进行无监督训练，并借 $\mathcal{M}$ 训练三元向量组$\left\{\mathbf{I}_{\mathbf{y}_r}^m,\mathbf{y}_r^m,\mathbf{y}_g^m\right\}_{m=1}^M$，其中目标向量$\mathbf{y}_g^m$随机生成。GANimation 既不需要同一个人在不同表情下的成对图像，也不需要期望的目标图像$\mathbf{I}_{\mathbf{y}_g}$。

如图 4.2 所示，该网络结构由两个主要部分组成:一个用于回归注意力的生成网络 G 和颜色掩膜;判别网络 D 要对生成图像的真实性 D_I 和表情条件完成度$\hat{\mathbf{y}}_g$进行评估。需要说明的是，GANimation 是无监督的，即同一个人不同表情的图像对和目标图像$\mathbf{I}_{\mathbf{y}_g}$都假设是未知的。

4.2.2　网络结构

设 G 为生成网络块，因为它是双向应用的(例如，将任一输入图像映射到所需表情，反之亦然)，在下文中，将使用下标 o 和 f 来表示起点和终点。给定图像$\mathbf{I}_{\mathbf{y}_o}\in\mathbb{R}^{H\times W\times 3}$和编码所需的表达式 N 维向量$\mathbf{y}_f$，将生成器的输入作为一组串联$\left(\mathbf{I}_{\mathbf{y}_o},\mathbf{y}_o\right)\in\mathbb{R}^{H\times W\times(N+3)}$，其中$\mathbf{y}_o$表示为大小为$H\times W$的 N 个数组。

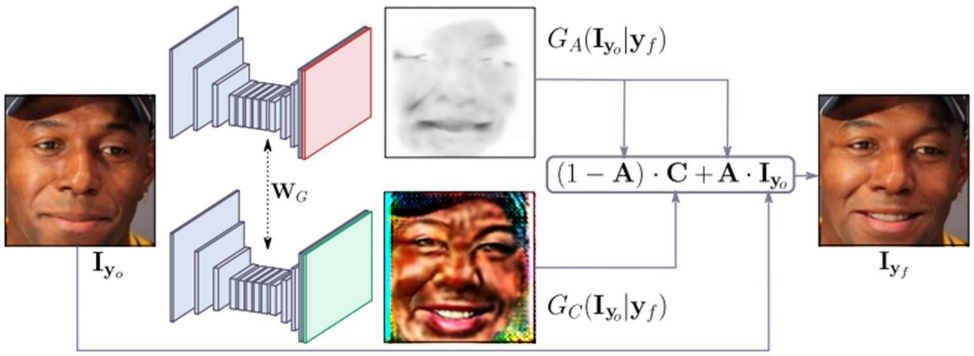


图 4.3 基于注意力机制的生成网络

GANimation 系统的一个关键组成部分是使 G 只专注于图像中那些负责合成新表情的区域，并且保持图像中其他元素，如头发、眼镜、帽子或珠宝等非表情元素不受影响。为此，GANimation 在生成网络中嵌入了注意力机制。具体来说，GANimation 生成网络输出两个掩膜，一个颜色掩膜 C 和一个注意力掩膜 A，而不是回归一个完整的图像。最终图像的获得方式如下：

$$\mathbf{I}_{\mathbf{y}_f}=(1-\mathbf{A})\cdot\mathbf{C}+\mathbf{A}\cdot\mathbf{I}_{\mathbf{y}_0} \tag{4-2}$$

其中，$\mathbf{A}=G_A\left(\mathbf{I}_{\mathbf{y}_0}\mid\mathbf{y}_f\right)\in\{0,\ldots,1\}^{H\times W}$ 和 $\mathbf{C}=G_C\left(\mathbf{I}_{y_o}\mid\mathbf{y}_f\right)\in\mathbb{R}^{H\times W\times 3}$。掩膜 A 表示扩展 C 的每个像素并对输出图像 $\mathbf{I}_{\mathbf{y}_f}$ 有贡献。通过这种方式，生成网络不需要渲染静态元素，只关注定义面部运动的像素，从而生成更清晰、更真实的合成图像。此过程如图 4.3 所示。在整个图像上，给定了输入图像，目标表情，生成网络回归表达式和注意力掩膜 A 和 RGB 颜色转换掩膜 C。注意力掩膜定义了每个像素的强度，确定了将原始图像每个像素的扩展程度，并将在最终呈现的图像中起作用。

条件化判别网络，即以生成图像真实性和期望表情完成度作为评价标准的判别网络。$D(\mathbf{I})$ 的结构类似于 PatchGan 网络[18]，从输入图像 I 映射到一个矩阵 $\mathbf{Y}_{\mathrm{I}}\in\mathbb{R}^{H/2^6\times W/2^6}$，其中 $\mathbf{Y}_{\mathrm{I}}[i,j]$ 表示重叠图像块 ij 为真实的概率。此外，为了评估其条件作用，在网络顶端添加副回归项首部，来估计在图像中 AUs 的激活函数 $\hat{\mathbf{y}}=\left(\hat{y}_1,\ldots,\hat{y}_N\right)^{\top}$。

4.2.3 模型学习

GANimation 定义的损失函数包含四个内容，即：由 Gulrajani 等人修改后的图像对抗损失函数 WGAN-GP[19]，将生成图像的分布拓展到训练图像的分布；使注意力面罩光滑并防止其饱和的注意力损失函数；将生成图像的表情设置为与期望图像相似的条件化表情损失函数；有利于保持人面部纹理一致性的一致性损失函数。下面将给出上述损失函数的详细信息：

① 图像对抗损失函数

为了了解生成网络 G 的参数，GANimation 使用了 WGAN-GP 提出的标准 GAN 算法的修正版本。具体来说，原始的 GAN 公式是基于 Jensen-Shannon (JS)散度损失函数，其目的是最大化真实图像的分类正确概率和当生成网络欺骗判别网络时，对图像进行渲染。这种损失可能不是连续的生成网络参数，而且局部饱和会导致判别网络中的梯度消失。通过 WGAN[20]替换连续地球移动距离的 JS 函数，可以解决此类问题。为了保持 Lipschitz 约束，WGAN-GP 为判别网络添加一个梯度惩罚作为判别网络输入的梯度范数。

令 $\mathbf{I}_{\mathbf{y}_o}$ 作为初始条件 $\mathbf{y}_o$ 的输入图像，$\mathbf{y}_f$ 为期望的最终条件，$\mathbb{P}_o$ 为输入图像的数据分布，$\mathbb{P}_{\tilde{I}}$ 为随机插值分布。然后，判别损失 $\mathcal{L}_{\mathrm{I}}\left(G, D_{\mathrm{I}}, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_f\right)$ 为：

$$\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[D_{\mathrm{I}}\left(G\left(\mathbf{I}_{\mathbf{y}_o}\mid\mathbf{y}_f\right)\right)\right]-\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[D_{\mathrm{I}}\left(\mathbf{I}_{\mathbf{y}_o}\right)\right]+\lambda_{\mathrm{gp}}\mathbb{E}_{\tilde{I}\sim\mathbb{P}_{\tilde{I}}}\left[\left(\left\|\nabla_{\tilde{I}}D_{\mathrm{I}}(\tilde{I})\right\|_2-1\right)^2\right] \quad (4\text{-}3)$$

其中，λ_{gp} 为惩罚系数。

② 注意力损失函数

在训练模型时，与颜色掩膜 C 类似，没有对注意力掩膜 A 进行 ground-truth 注释，而从判别模块的结果梯度和其他损失函数中学习的。然而，注意力掩膜很容易饱和到 1，这使得 $\mathbf{I}_{\mathbf{y}_o}=G\left(\mathbf{I}_{\mathbf{y}_o}\mid\mathbf{y}_f\right)$，也就是说，生成网络没有起效。为了防止这种情况，GANimation 用一个 l_2 权重惩罚系数来调整掩膜。同时，为了在将输入图像像素与颜色变换 C 相结合时，进行平滑的空间颜色变换，GANimation 对 A 进行全变差正则化。因此，注意力损失 $\mathcal{L}_{\mathrm{A}}\left(G, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_f\right)$ 可以定义为：

$$\lambda_{\mathrm{TV}}\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\sum_{i,j}^{H,W}\left[\left(\mathbf{A}_{i+1,j}-\mathbf{A}_{i,j}\right)^2+\left(\mathbf{A}_{i,j+1}-\mathbf{A}_{i,j}\right)^2\right]\right]+\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|\mathbf{A}\|_2\right] \tag{4-4}$$

其中，$\mathbf{A}=G_A\left(\mathbf{I}_{\mathbf{y}_o}\,|\,\mathbf{y}_f\right)$，$\mathbf{A}_{i,j}$是$A$的第$i,j$个入口。$\lambda_{\mathrm{TV}}$是惩罚系数。

③ 条件化表情损失函数

在减少图像对抗损失的同时，生成网络还必须减少D上 AUs 回归产生的误差。这样，G不仅学会了渲染真实的样本，还学会了满足$\mathbf{y}_f$编码的目标面部表情。这个损失由两个部分定义:一个是用于优化G的伪图像的 AUs 回归损失，另一个是用于学习D上回归的真图像的 AUs 回归损失。这个损失$\mathcal{L}_{\mathrm{y}}\left(G,D_{\mathrm{y}},\mathbf{I}_{\mathbf{y}_o},\mathbf{y}_o,\mathbf{y}_f\right)$如下所示：

$$\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|D_{\mathbf{y}}\left(G\left(\mathbf{I}_{\mathbf{y}_o}\,|\,\mathbf{y}_f\right)\right)\right]-\mathbf{y}_f\|_2^2\Big]+\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\left\|D_{\mathbf{y}}\left(\mathbf{I}_{\mathbf{y}_o}\right)-\mathbf{y}_o\right\|_2^2\right] \tag{4-5}$$

④ 一致性损失函数

由上文所述的损失函数，生成网络进行生成逼真的面部转换。但是，如果没有 ground-truth 监督，就无法保证输入和输出图像中的人脸源于同一个人。通过使用循环一致性损失函数[21]，惩罚原始图像$\mathbf{I}_{\mathbf{y}_o}$和其重建之间的差异使得生成网络保持每个个体的一致性。具体公式如下：

$$\mathcal{L}_{\mathrm{idt}}\left(G,\mathbf{I}_{\mathbf{y}_o},\mathbf{y}_o,\mathbf{y}_f\right)=\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\left\|G\left(G\left(\mathbf{I}_{\mathbf{y}_o}\,|\,\mathbf{y}_f\right)|\,\mathbf{y}_o\right)-\mathbf{I}_{\mathbf{y}_o}\right\|_1\right] \tag{4-6}$$

为了生成逼真的图像，对低频信号和高频信号都进行建模。GANimation 的 PatchGan 基于判别网络D_{I}，通过限制对局部图像块结构的注意力来强化高频信号的准确性。为了捕获低频信号，使用l_1范数便已足够。在初步实验中，尽管没有性能的提升，还是尝试用更复杂的感知损失函数[22]来代替l_1范数。

⑤ 全损失函数

为了生成目标图像$\mathbf{I}_{\mathbf{y}_g}$，通过线性组合上文所述的部分损失函数，来建立全损失函数$\mathcal{L}$：

$$\begin{aligned}\mathcal{L}=&\mathcal{L}_{\mathrm{I}}\left(G,D_{\mathrm{I}},\mathbf{I}_{\mathbf{y}_r},\mathbf{y}_g\right)+\lambda_{\mathrm{y}}\mathcal{L}_{\mathrm{y}}\left(G,D_{\mathrm{y}},\mathbf{I}_{\mathbf{y}_r},\mathbf{y}_r,\mathbf{y}_g\right)\\&+\lambda_{\mathrm{A}}\left(\mathcal{L}_{\mathrm{A}}\left(G,\mathbf{I}_{\mathbf{y}_g},\mathbf{y}_r\right)+\mathcal{L}_{\mathrm{A}}\left(G,\mathbf{I}_{\mathbf{y}_r},\mathbf{y}_g\right)\right)+\lambda_{\mathrm{idt}}\mathcal{L}_{\mathrm{idt}}\left(G,\mathbf{I}_{\mathbf{y}_r},\mathbf{y}_r,\mathbf{y}_g\right)\end{aligned}\tag{4-7}$$

其中，$\lambda_{\mathrm{A}},\lambda_{\mathrm{y}}$ 和 λ_{idt} 控制每个部分损失函数相对重要性的超参数。最后，定义极大极小问题，如下所示：

$$G^{\star}=\arg\min_{G}\max_{D\in\mathcal{D}}\mathcal{L}\tag{4-8}$$

其中，$G^{\star}$ 从数据分布中抽取样本。另外，将判别网络 D 约束在 $\mathcal{D}$ 中，$\mathcal{D}$ 表示 1-Lipschitz 函数的集合。

4.3　本章小结

图 4. 4 GANimation 模型生成的失败结果

GANimation 是一种基于 AU 标注的 GAN 条件化方法，该方法在连续的流行中描述了定义人类表情的面部解剖运动。GANimation 模型采用完全无监督策略训练，只需要激活 AU 标注图像，并利用注意力机制，便可对不断变化的背景和光照条件具有鲁棒性。相比于 StarGAN[15]只能由数据集决定生成离散的表情，其生成

的图像连续自然，较为清晰。相比于其他条件生成模型，GANimation 在合成多类表情和处理自然图像的能力上均有超越。

GANimation 在某些情况下会出现失败结果，如下图所示。这可能是因为输入图像仅为一张，训练数据不足引起的。当输入极端表情时，颜色掩膜没有及时调整权重，会导致局部出现透明化。如果输入图像的对象是非人类，模型的效果也会很差。GANimation 生成图像作为人脸表情数据集的扩充，数量方面还远远不够，尤其不足以满足深度学习表情识别海量训练数据的需要，此方面待以改进。

第五章　SinGANimation 表情生成算法实验

5.1　SinGANimation 模型架构

本文意图构建基于单幅图像的面部表情生成模型，即输入单幅人脸表情图像，经过模型训练，可以输出多种人脸表情的多幅图像。SinGAN 模型生成人脸表情的种类单一，而且容易出现失真的现象。于是，本文对 SinGAN 模型进行了改进，解决原有模型人脸生成失真的问题，并创新引入 GANimation 模型，构建出一种新的完全无监督表情生成算法 SinGANimation，使生成人脸表情的类别大幅增加。具体架构如图 5.3 所示。

该算法的基本原理为：输入一种表情类别为 C_0 的单幅图像 I_{C_0}，首先通过 GANimation，进行单个 AU 变换、多个 AU 连续变换，多个 AU 离散变换等操作，对图像的表情种类扩充至 N 个，值得注意的是此时每种表情类别还是单幅图像。然后，将多种表情的单幅图像输入 SinGAN 中，进行再生成操作，对每种表情图像增加至 M 个。因为 SinGAN 再生成的图像与训练图像的差别较小，但又与完全复制不同，所以 SinGAN 再生成只改变每种图像的数量，并不会改变图像种类的多少，即最终结果为 $M(\mathrm{I}_{C_1}+\mathrm{I}_{C_2}+\cdots+\mathrm{I}_{C_N})$。这也是本文最大的亮点所在。

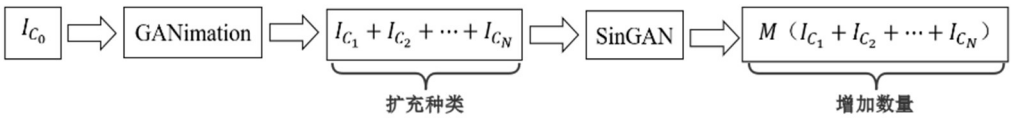


图 5.1 SinGANimation 模型架构

5.2　SinGANimation 模型训练

5.2.1　GANimation 模型训练

GANimation 生成网络建立在 Johnson 等人[23]提出的网络变化基础上，其被证明在图像之间匹配中取得非常好的结果。对它进行了轻微的修改，将最后一个卷积层替换为两个并行卷积层，其中一个是对颜色掩膜 C 的回归计算，另一个是针对注意掩膜 A。通过将生成网络中的批量归一化替换为实例归一化，可以提高训练的稳定性。对于判别网络，采用 PatchGAN 架构，但是去掉特征归一化。否则，在计算

梯度惩罚时，判别网络的梯度范数将对整个批进行计算，而不是对每个单独的输入进行计算。

在参数优化方面，使用 Adam 优化算法[24]，其参数为学习率 0.0001，β_1 0.5，β_2 0.999，批量大小为 25。训练 30 个周期，学习率在最后的 10 个周期内线性衰减到 0。每 5 次判别网络的优化，对应执行一次生成网络的单一优化。损失函数的权重系数设为 $\lambda_{gp}=10$，$\lambda_A=0.1$，$\lambda_{TV}=0.0001$，$\lambda_y=4000$，$\lambda_{idt}=10$。为了提高稳定性，尝试在不同的生成网络更新中，使用带有生成图像的缓冲区来更新判别网络，但是没有明显性能的改进。该模型需要在 GTX 1080Ti GPU 上训练两天。

5.2.2 SinGAN 模型训练

根据顺序训练多尺度体系结构，从最大的尺度到最小的尺度。一旦每个 GAN 被训练，它就会被固定下来。对第 n 个 GAN 的训练损失包括对抗阶段和重建阶段，即

$$\min_{G_n}\max_{D_n} L_{adv}(G_n,D_n)+\alpha L_{rec}(G_n) \tag{5-1}$$

对抗损失 L_{adv} 是为 x_n 的子块距离分布和生成样本 $\tilde{x}_n$ 的子块距离分布构造的惩罚函数。重建损失 L_{rec} 确保可以产生 x_n 的特定噪声图谱。

对抗损失每个生成网络 G_n 都与一个马尔可夫链的判别网络 D_n 相结合，该 D_n 将其输入的每个重叠的子块分类为真或假。本文使用 WGAN-GP 损失函数来增加训练的稳定性，其中最终的判别得分是图像块判别得分的平均值。相对于纹理的单一图像 GANs，本文定义了整个图像的损失，而不是随机的切割图像。这允许网络学习边界条件，这是 SinGAN 设置的一个重要特性。D_n 的架构与 G_n 中的 ψ_n 网络一样,所以它的块大小(网络的感受野)是 11×11。

为了确保特定的输入噪声图谱，可以生成原始图像 x。本文特别选择了 $\{z_N^{rec},z_{N-1}^{rec}\cdots,z_0^{rec}\}=\{z^*,0,\cdots,0\}$，其中 z^* 是一些固定的噪音图谱(只绘制一次，在训练时保持固定)。在使用图像的噪声图谱时，由 $\tilde{x}_n^{rec}$ 表示第 n 个尺度的生成图像。则对于 $n<N$，即

$$L_{\text{rec}} = \left\| G_n(0,(\tilde{x}_{n+1}^{rec})\uparrow^r) - x_n \right\|^2 \quad (5\text{-}2)$$

对于 $n=N$，$L_{\text{rec}} = \left\| G_N(z^*) - x_n \right\|^2$。重建的图像 $\tilde{x}_n^{rec}$ 在训练中还负责确定每个尺度中噪声 z_n 的标准差 σ_n。具体来说，把 σ_n 与 $(\tilde{x}_{n+1}^{rec})\uparrow^r$ 和 xn 之间的均方误差(RMSE) 成正比，这提供说明了此尺度内需要添加的细节数量。

5.3　实验数据集

为了验证本文提出算法的性能，两个面部表情数据集被采用进行实验分析：CelebA 和 RAF-DB 数据集，其中 CelebA 数据集的图像光照均匀，而 RAF-DB 数据集的图像因在一般环境下拍摄，光照分布相对不规则。下面将分别对这两个数据集的细节以及数据处理进行介绍。

5.3.1　CelebA 数据集

CelebA 数据集[25]（CelebFaces Attribute）数据集包含 10177 个名人身份的 202599 张人脸图片，每张图片都有特征标记，包含 40 个二进制属性标注，5 个人脸特征点坐标等。图 5.2 展示了一些 CelebA 数据集的面部图像。

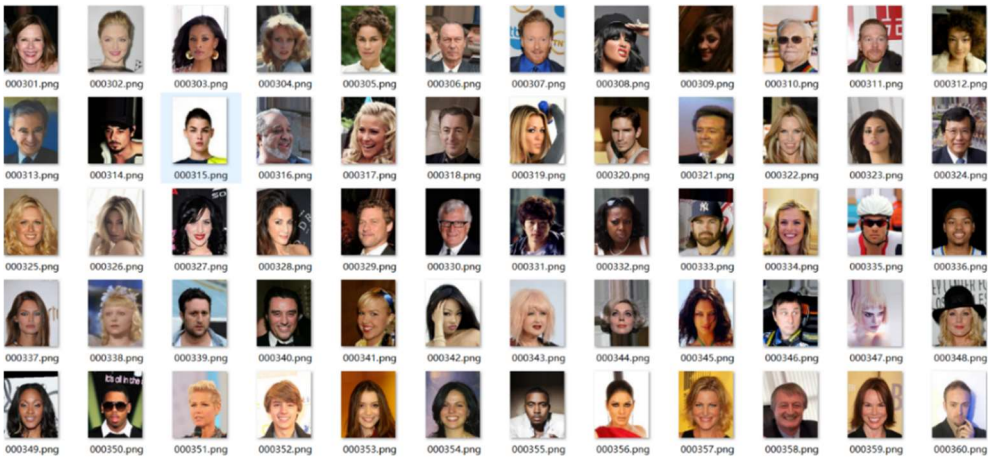


图 5.2　CelebA 数据集

5.3.2　RAF-DB 数据集

图 5.3 RAF-DB 数据集

RAF-DB 数据集[26]即真实世界的情感面孔数据库，是一个大规模的面部表情数据库，包括从网上下载的大约 3 万张多样的面部图像。该数据库中的图像在受试者的年龄，性别和种族，头部姿势，光照条件等方面变化很大。图 5.3 展示了一些 RAF-DB 数据集的面部表情图像。

5.3.3 数据预处理

在实验中，随机选取数据集的 80%作为训练集，余下的 20%作为测试集。为了加快训练时间，实验中使用 OpenCV，将 CelebA 数据集的图像尺寸从178×218调整为128×128。另外，由于 RAF-DB 数据集原生图像格式不一，因此对其进行统一裁剪提取图像的人脸区域，本文采用 OpenFace 提取每个图像动作单元，并将每个输出存储在与图像同名的 csv 文件中，以供后续训练模型使用。对 GANimation 模型生成的初步结果进行下采样，便于 SinGAN 模型生成结构更完整、图像更清晰的结果。

5.4 实验结果及其定性分析

5.4.1 单个 AU 变换结果

首先，对模型在不同强度激活 AUs 的能力进行评估。该部分实验使用 CelebA 数据集进行测试，在该数据集上进行单个 AU 变换，9 个 AU 子集分别转换为 4 个强度级别(0、0.33、0.66、1)，实验结果如图 5.4 所示。从图 5.4 所示的结果可发现：当强度为 0 时，不改变相应的 AU；当强度非 0 时，可以观察到每个 AU 是如何逐

步变化。在不同强度下，本文提出的算法模型都能很好地生成与输入图像相对应的结果。为了避免引入不需要的面部运动，恒等变换是至关重要的。

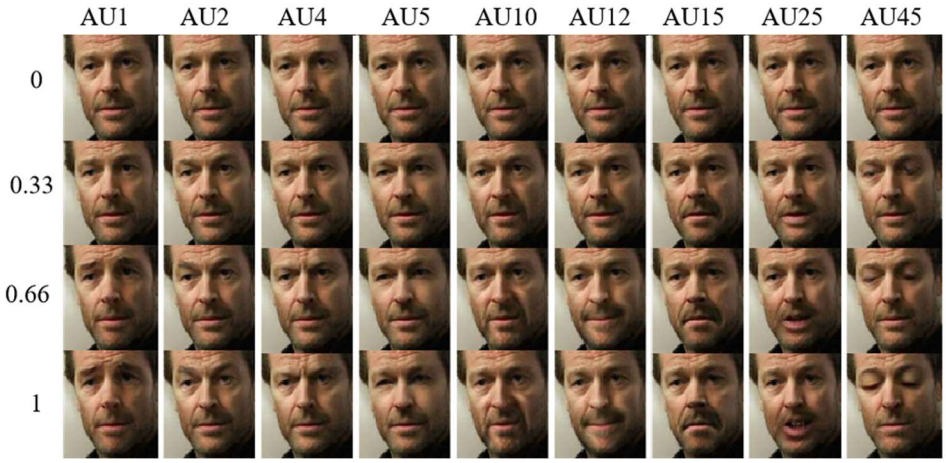


图 5.4 单个 AU 变换在 CelebA 数据集生成的结果

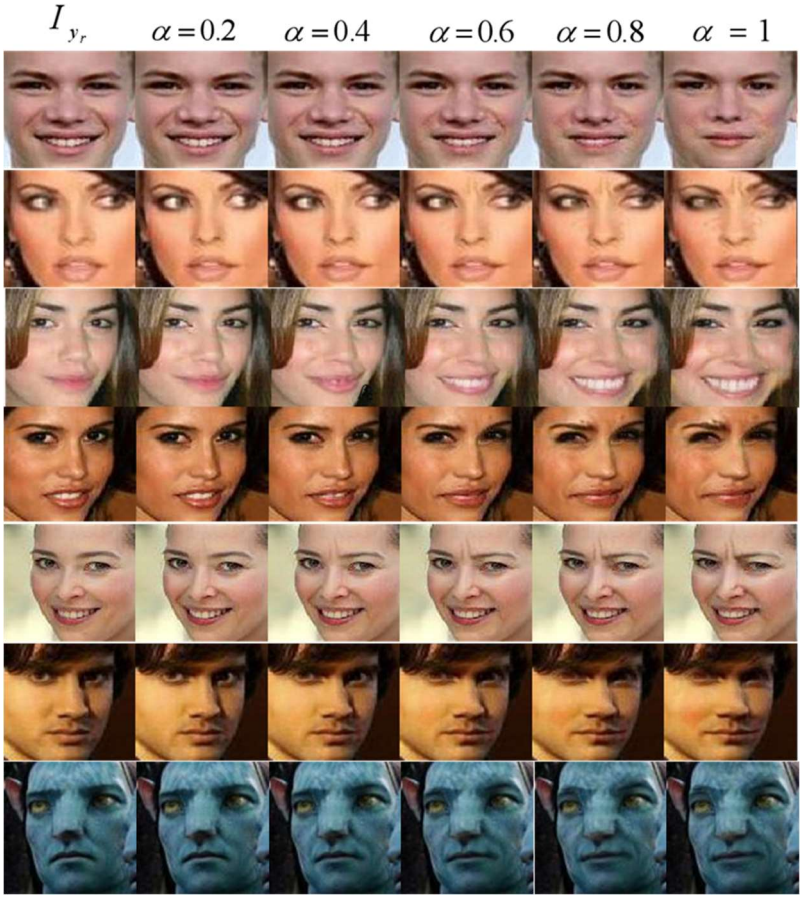


图 5.5 多个 AU 连续变换在 CelebA 数据集生成的结果

此外，从实验结果中也可看出，模型非常成功地渲染复杂的面部运动，生成图像与真实图像难辨真假。生成网络对面部肌肉集群学习训练具有独立性，无混杂交叠的情况，比如：对应眼部和人脸上半部分的 AU（AU1，AU2，AU4，AU5，AU45）

不影响嘴部的 AU。同样，嘴部 AU 的变化(AU10，AU12，AU15，AU25)也不影响眼部和眉毛的 AU。

5.4.2 多个 AU 连续变换结果

为了充分展现本文提出算法的性能，该部分实验尝试在 CelebA 数据集上进行多个 AU 连续变换，评估其插入多种表情的能力。实验结果如图 5.5 所示，其中，第一列为表情 $\mathbf{y}_r$ 的原始图像，最后一列是以 $\mathbf{y}_g$ 为目标表情的综合生成图像，其余几列的结果根据生成网络的条件生成，其线性插值的初始和目标表达式为 $\alpha\mathbf{y}_g+(1-\alpha)\mathbf{y}_r$。由图 5.5 所示的实验结果可知，本文提出的算法其跨帧转换的平滑一致性非常显著。特别地，在该实验中特意选择了具有挑战性的样本，来验证提出算法对光照条件的鲁棒性，甚至是对非现实世界数据分布的鲁棒性，而这些现有的模型中是看不到的。实际上，对于多个 AU 连续变化的实验结果对于之后该模型扩展到视频生成领域具有一定的指导意义。

5.4.3 多个 AU 离散变换结果

接下来，本文将 GANimation 与 DIAT[27]、CycleGAN[6]、IcGAN[28]和 StarGAN[15]模型进行比较。为了公平比较，本文采用了这些模型的结果，即是由 StarGAN 在 RAF-DB 数据集中生成的离散情绪结果(例如，快乐、悲伤和恐惧)。因为 DIAT 和 CycleGAN 是非条件生成，所以对于每一对可能的原始和目标情绪，它们被独立来训练。下面将首先对几种对比算法模型进行简单介绍：

- DIAT：给定输入图像 $x\in X$ 和参考图像 $y\in Y$，DIAT 学习 GAN 模型在图像 x 上渲染参考域 Y 的属性，同时保持人物的不变性。通过经典的对抗性损失和循环损失 $\left\|x-G_{Y\to X}\left(G_{X\to Y}(x)\right)\right\|_1$ 进行训练，以保持人物的一致性。
- CycleGAN：如第二章所述，与 DIAT 类似，CycleGAN 也学习两个域之间的映射 $X\to Y$ 和 $Y\to X$。为了训练域之间的变换，使用正则项来表示两个周期的周期一致性损失：$\left\|x-G_{Y\to X}\left(G_{X\to Y}(x)\right)\right\|_1$ 和 $\left\|y-G_{X\to Y}\left(G_{Y\to X}(y)\right)\right\|_1$。
- IcGAN：对于给定的输入图像，IcGAN 使用预训练的编码-解码器将图像编码为潜在表示，并与表情向量 y 连接，然后重构原始图像。在通过解码器之前，用目标表情替换 y 来修正表情。

- StarGAN：它使用一个掩码向量来忽略未指定的标签，并且只对已知的真值标签进行优化。当同时使用多个数据集进行训练时，它会产生更实际的结果。

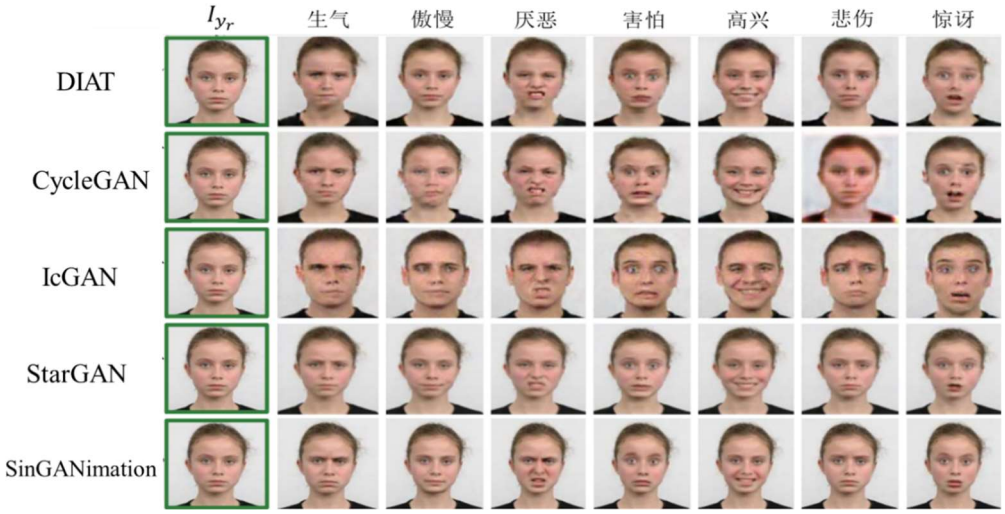


图 5.6 多个 AU 离散变换在 RAF-DB 数据集生成的结果

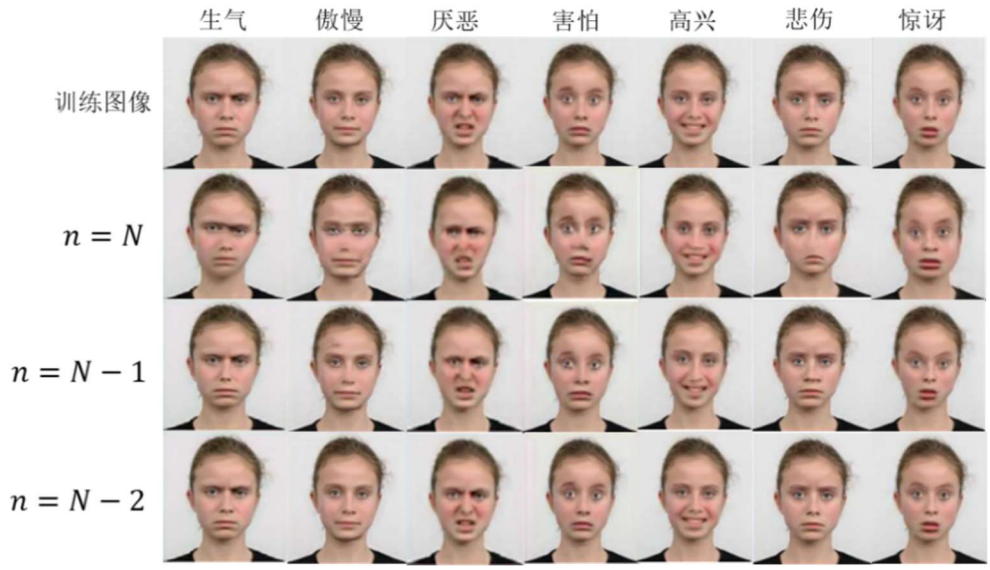


图 5.7 在 RAF-DB 数据集的再生成结果

相比以上对比算法模型，本文提出的算法与它们主要在两个方面不同。首先，本文提出的算法不以离散的情感类别为模型设定条件，而是学习了一种在解剖学上可行的变形理论，它允许产生连续的表情。其次，在使用掩膜时，只允许在裁剪后的面部进行变形，并将其放回到原始图像上，而不会产生任何人为修改痕迹。图 5.6 展示了本文算法和 4 种对比算法在多个 AU 离散变换下对 RAF-DB 数据集的面

部表情生成结果。从图 5.6 所示的结果可知，本文提出的算法模型比其他模型生成的图像更真实清晰、灵活生动。

5.4.4 SinGANimation 再生成结果

如在第三章的优缺点分析所述，当图像的整体结构非常重要时，SinGAN 可能会产生不真实的结果。为此，本文对 SinGAN 原始模型进行了改进，通过在更小的范围内启动生成过程来避免人脸图像失真的情况。图 5.7 展示了本文算法在 RAF-DB 数据集上的生成结果。从图 5.7 所示的结果可知，在原始模型下，以最大尺度（$n=N$）开始生成会导致图像严重失真。然而，通过对图像进行下采样，在更小的尺度上输入 $N-1$ 层和 $N-2$ 层，可以保持面部的整体结构，同时只改变更微小的细节，如眼睛、鼻子和嘴唇的形状、皮肤和眉毛的纹理等。

5.5 SinGANimation 实验结果定量分析

上节展示了对 SinGANimation 模型的定性分析，本节将量化其生成图像的真实性，以及发现模型如何捕捉训练图像的内部统计数据。实验使用 AMT 真伪用户测试和单幅图像 FID[29]测量来进行定量分析。

5.5.1 AMT 真伪用户测试

AMT 真伪用户测试是基于亚马逊人工智能平台的“图片真伪判别任务”，来评估模型生成图像的真实性。该实验在两种情况下进行测试实验：

① 匹配(真伪对比)

受试者面前有 50 个实验序列，每个序列中仅有一张伪图像(SinGANimation 生成)，其余均为真实图像。在进行 1 秒钟的对比后，受试者被要求挑选出伪图像。

② 非匹配(不区分真伪)

受试者一秒钟看一张图片，然后被问这是否为虚假图像。总共有 50 张真实图像和 50 张不重复的虚假图像随机呈现给每位受试者。

本次实验对两种情况都采用了两种生成方法：从较大尺度 $N-1$ 开始生成和从 $N-2$ 尺度开始生成，如图 5.7 所示。在这 4 个测试中，有 50 个不同的受试者。在所有测试中，前 10 个测试都是包含反馈的教程，具体的实验结果如表 5.1 所示。

表 5.1 真伪 AMT 测试结果

输入尺度	调查类型	混淆率
$N-1$	匹配	$21.45\%\pm1.5\%$
	非匹配	$42.9\%\pm0.9\%$
$N-2$	匹配	$30.45\%\pm0.9\%$
	非匹配	$47.04\%\pm0.8\%$

表 5.2 SIFID 结果

输入尺度	SIFID	调查类型	SIFID/AMT 混淆率
$N-1$	0.09	匹配	-0.55
		未匹配	-0.22
$N-2$	0.05	匹配	-0.56
		未匹配	-0.34

通过实验得出两个生成过程的混淆率：从最大尺度 $N-1$ 开始(生成具有较大多样性的样本)和从第二个最大尺度 $N-2$ 开始(保存原始图像的全局结构)。在每种情况下，都进行了匹配和非匹配测试，方差通过 Bootstrap[30]计算。从获得的实验结果可知，

实验结果正如预期的那样，在非配对情况下，混淆率始终较大，即使改变了大型结构，这表明通过本文提出算法生成的图像也很难与真实图像区分开来(混淆率50%意味着完全混淆真实图像和虚伪图像)。

5.5.2　单幅图像 FID 测量

接下来量化了 SinGANimation 对输入图像 x 的内部数据的捕捉程度。GAN 评估的一个常用度量是 FID，它测量生成图像的深度特征分布与真实图像的深度特征分布之间的偏差。然而，在本文实验的设置中，仅仅只有一张真实图像，并且对它的内部图像块数据非常感兴趣。因此，本文提出了单幅图像 FID ，简称 SIFID 测量，它没有使用 Inception 网络[31]中最后一个池化层之后的激活向量(每个图像一个向量)，而是使用在第二个池化层之前使用卷积层输出深层特征的内部分布(地图中每个位置一个向量)。本文提出的 SIFID 是真实图像和生成图像中特征统计数据之间的 FID。

实验中将 FID 指标应用于单个图像，并得出完全生成的 50 个图像的平均分。实验结果如表 5.2 所示。表 5.2 可以看出，AMT 结果的相关性表明，SIFID 与人为测试结果一致。由 $N-2$ 尺度生成的 SIFID 平均值低于 $N-1$ 尺度生成的 SIFID 平均值，这与用户测试结果保持一致。该部分实验还说明了 SIFID 值与虚伪图像混淆率之间的相关性，两者之间存在明显的的反相关性，这意味着越小的 SIFID 通常对应较大的混淆率。匹配测试的相关性更强，因为 SIFID 是匹配的度量标准，其作用于匹配图像 x_n，$\tilde{x}_n$。

5.6 本章小结

本文提出的算法 SinGANimation 结合了 GANimation 和 SinGAN，在完全无监督的情况下，对单幅人脸图像进行学习训练，并且生成的表情图像种类多样，规模庞大，质量较高。在 SinGANimation 模型搭建过程中，通过将训练数据下采样输入，解决了 SinGAN 原有模型的人脸生成失真问题，进一步优化 SinGAN 模型，拓宽了 SinGAN 的适用范围。对于 SinGANimation 模型的性能评估使用了 CelebA 和 RAF-DB 两个数据集，前者为特定环境下拍摄的数据集，后者为在自然环境下拍摄的数据集，两者的光照条件，人脸特征等方面差距较大。在两个数据集上生成的图像均达到不错的效果，证明 SinGANimation 可以在不同数据集上应用，具有良好的鲁棒性。由于现有的人脸表情数据库的图像种类和数量有限，而且例如深度学习等人脸表情识别需要大规模的训练数据，所以在人脸数据量较少的情况下，SinGANimation 模型可以对人脸表情图像进行种类和数量的扩充，满足实验的需求。SinGANimation 可生成连续的表情序列，所以在后续的工作中，可以应用于视频序列等工作中。由此可见，SinGANimation 具有非常重要的商用和科研实验价值。另外，SinGANimation 的再生成图像，与对应种类的图像差别较小。这可能是 SinGAN 内部学习训练的原因，在语义多样性方面具有一定的限制。训练 SinGANimation 模型耗时较长，算法时间复杂度较高，还待算法性能进一步优化。

第六章　工作总结

人脸表情生成算法与自然图像生成相比，在保证生成图像的画质清晰以外，还需要确保人脸的高度结构性。原有的 SinGAN 对自然图像的生成效果很好，但对人脸图像的效果差强人意。于是，本文对人脸图像的特性进行研究，发现将图像下采样后再输入，可以保证人脸结构的一致性，解决了 SinGAN 原有模型人脸生成失真的问题。即便这样，人脸表情的数量有所增加，但种类基本保持不变。因此，本文首次引入了 GANimation 到 SinGAN 模型中，构建出一种新的完全无监督表情生成算法 SinGANimation。在 SinGANimation 模型中，通过 GANimation 进行单个 AU 变换、多个 AU 连续变换、多个 AU 离散变换等操作，对图像的表情种类进行扩充，这样有效地实现了生成表情的数量和种类都有一定程度的增加。实验中，通过与其他经典模型对比，发现本文提出的 SinGANimation 模型可以控制 AU 变换，既可以生成连续自然的表情，还可以生成离散情绪的图像。此外，本文进行了 AMT 真伪用户测试和单幅图像 FID 测量，得到的生成图像的深度特征分布与真实图像的深度特征分布之间的偏差分别为 0.09 和 0.05，混效率接近 50%，表明本文提出的 SinGANimation 模型生成的图像与真实图像高度相似。同时，在不同数据集上实验结果均达到不错的效果，也验证了 SinGANimation 算法具有良好的鲁棒性。

致 谢

2020 年是特殊的一年，尤其对于我们这些毕业生意义非凡。从最初的恐慌，中期的顽强抗疫，直到现在的形势大好，希望这场灾难尽早落下帷幕。在无尽的等待中，发现自己快要毕业了。对于现在的我，大学是最重要的一个阶段，它让我收获颇多。这一路走来，感谢的人有很多。谢自己，一路坚持，追寻自己的梦想。谢父母，默默支持，给予我如空气般难以察觉却必需的爱。谢益友，排忧解难，总有你们在我身边。谢良师，传道解惑，以身作则教我做人。谢母校，提供资源平台让我开拓视野，发现山外的山。谢白衣天使们，有你们的前线抗战，才有我们的后方安逸。谢祖国，此生无悔入华夏。

本文是由 xxx 老师悉心教导下完成的。我与 x 老师初见，是在《智能数据挖掘》的课堂。x 老师讲课深入浅出，算法原理分析细致严谨，让我感慨还有老师如此重视授课的不易。此外，我在课外还向 x 老师请教一些职业生涯规划的问题，x 老师都一一解答，让我真的很感动。这次，有幸选到了 x 老师的毕设课题，在我毕设的学习过程中，老师每周一次与我耐心讨论，给予我工作的反馈和指导。甚至在疫情返校不方便的情况下，老师还为我搭建远程服务器，这才让我的毕设赶上进度。在此，向 x 老师致以最诚挚的谢意。

同时，感谢向我伸出援手的外援：舍友陈少宏，光辉学长，东南大学的李阳师兄，常洪丽师姐。谢谢你们为我指点迷津。

最后，希望自己的初心不变，勇往前行。期待着初夏的西电，与我最想见的你们重逢。

参考文献

[1] Mehrabian, Albert,Silent Messages (1st ed.,Belmont, CA: Wadsworth. ISBN 0-534-00910-7.1971.

[2] Ekman, P. & Friesen, W. V The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica, 1, 49–98.1969.

[3] Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua. Generative Adversarial Networks. 2014. arXiv:1406.2661

[4] A Radford, L Metz, S Chintala - arXiv preprint arXiv:1511.06434, 2015 - arxiv.org

[5] Goodfellow Ian, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.

[6] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J]. arXiv preprint arXiv:1703.10593, 2017.

[7] Udwary D W, Zeigler L, Asolkar R N, et al. Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica[J]. Proceedings of the National Academy of Sciences, 2007, 104(25): 10376-10381.

[8] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani.Ingan: Capturing and remapping the "DNA" of a natural image. arXiv preprint arXiv: arXiv:1812.00231, 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. arXiv preprint, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[11] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo realistic single image super-resolution using a generative ad versarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681–4690,2017.

[12] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In null, page 416. IEEE, 2001.

[13] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. IEEE Signal Processing Letters, 20(3):209–212, 2013.

[14] Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML .2017.

[15] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. CVPR.2018.

[16] Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers 22(1), 67–92,1973.

[17] Yu, H., Garrod, O.G., Schyns, P.G.: Perception-driven facial expression synthesis.Computers & Graphics 36(3),2012.

[18] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR ,2017.

[19] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: NIPS ,2017.

[20] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint arXiv:1701.07875 ,2017.

[21] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV ,2017.

[22] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV ,2016.

[23] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV,2016.

[24] Kingma, D., Ba, J.: ADAM: A method for stochastic optimization. In: ICLR (2015)

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[26] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion 24(8), 1377–1388,2010.

[27] Li, M., Zuo, W., Zhang, D.: Deep identity-aware transfer of facial attributes. arXiv preprint arXiv:1610.05586,2016.

[28] Perarnau, G., van de Weijer, J., Raducanu, B., Alvarez, J.M.: Invertible conditional ´GANs for image editing. arXiv preprint arXiv:1611.06355 ,2016.

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner,Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, pages 6626–6637, 2017.

[30] Bradley Efron. Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics, pages 569–593.Springer, 1992

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet,Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015