

中国科学技术大学

学士学位论文



基于手绘草图的人脸照片生成和编辑

作者姓名： 程志华

学科专业： 自动化

导师姓名： 陈雪锦 副教授

完成时间： 二〇二〇年五月十二日

University of Science and Technology of China

A dissertation for bachelor's degree



Photorealistic face image synthesis and editing from hand-drawn sketch

Author: Cheng Zhihua

Speciality: Automation

Supervisor: Prof. Chen Xuejin

Finished time: May 12, 2020

致 谢

经过这几个月来的学习和调研、思考和实验，我终于完成了自己的本科毕业设计，此刻我思绪万千，想说的话有很多，但最想表达的是自己满怀的感激。

首先，我要感谢的人就是我的导师陈雪锦副教授。每当我遇到困惑不解的问题时，她总能给我提供宝贵的指导意见，使我拨云见日、柳暗花明。每当我放松懈怠时，她又会及时地点醒我，鞭策我不断向前。在她的引导下，我学会更加深入地思考而不只流于问题的表面。她还亲自动手帮我修改论文，叮嘱我写作论文的注意事项，令我感动不已。老师对我的教诲、包容与鼓励，给了我克服困难、直面挫折的信心和勇气，使我能顺利完成自己的毕业设计。老师严谨的治学态度和孜孜以求的探索精神，给我以后的科研道路树立了榜样，也让我获益匪浅。

其次，我要感谢科技西楼 1205 多媒体计算与通信实验室的所有同学们，特别是组内的李宇航、杨彬鑫和陈梓涵三位师兄，在我完成毕设的过程中给予我很多指导与帮助。正是他们的倾囊相授与无私帮助，才能让我在基础薄弱、经验缺乏的情况下一步步完成本文的工作。

我还要感谢我的班主任唐建老师，在大学四年里给予我很多关怀与帮助；感谢辅导员袁国富老师，多少次的促膝谈心、谆谆教诲，在工作和学习中给我支持与鼓励；感谢我的三位室友以及班里所有的同学们，陪伴我走过大学四年的时光，让我的青春焕发光彩。

同时，我还要感谢我的父母和爷爷奶奶，他们用无私的爱点亮了我的生活，尽他们所能给我营造最好的环境，但凡我取得一点点进步都离不开他们的奉献与付出。他们是我求学路上不断进取的动力，是我的信心之源。

2020 年注定是个多事之秋，我们国家刚刚经历过大疫，我们每个人也都被迫在家不能返校。但是我多想在毕业之前再回去见一面朝夕相处了四年的同学，说不定自兹之后各自天涯，我多想把毕业这一刻的喜悦跟母校分享，将毕业这一刻的美好定格于母校的每个角落。

所以最后，我要感谢我的母校，中国科大，在这里我汲取知识，丰富自我，结交了许多良师益友，度过了青春最珍贵的四年时光，这对我的一生都意义深远。我会继续努力，不断求索，不辜负母校对我的培养，也希望母校能越办越好，创寰宇学府纳天下英才，早日成为世界一流大学。

目 录

中文内容摘要	3
英文内容摘要	4
第一章 绪论	5
第一节 研究背景	5
第二节 本文工作	6
第二章 相关模型介绍	7
第一节 pix2pixHD 的模型结构	7
一、简介	7
二、网络结构	7
三、目标函数	8
第二节 实例标准化	9
第三节 模型引入	11
第四节 数据集生成	12
一、CelebA-HQ 数据集	12
二、轮廓数据集	12
三、手绘草图的生成	12
第三章 特征可视化	14
第一节 t-SNE 原理	14
一、SNE	14
二、t-SNE	15
第二节 草图特征分析	16
第三节 PCA 可视化分析	18
第四节 t-SNE 可视化分析	20
第四章 实验结果及分析	22
第一节 数据增广	22
一、输入数据的全局对齐问题	22
二、训练数据的增广	22

第二节 我们模型的生成效果及对比实验 ······	23
一、针对实例标准化的消融实验 ······	23
二、我们的模型与 pix2pixHD 的对比实验 ······	24
第五章 总结和展望 ······	26
参考文献 ······	28

中文内容摘要

`pix2pixHD` 是目前比较先进的一种图像翻译神经网络，适用于多种不同的图像翻译任务，其中包括从草图到人脸照片的生成。但是 `pix2pixHD` 对手绘草图输入的生成效果并不是很理想，生成质量不高，草图上某一位置的变化会引起生成图像全局的改变。

本文用可视化的方法分析了 `pix2pixHD` 生成器中间层的特征，发现了其网络结构存在的缺陷，并改进了其模型。改进后的模型生成质量更高，纹理细节更加丰富，而且还实现了图像编辑的功能。

本文还对训练数据进行了增广，使得训练和测试的过程更加鲁棒。

关键词： 图像翻译；手绘草图；图像编辑；可视化

Abstract

Pix2pixHD is a relatively advanced image translation neural network which is suitable for a variety of different image translation tasks, including face photo generation from hand-drawn sketches. However, the generation effect of pix2pixHD for hand-drawn sketch input is not really precise, and the generated images are often limited to the lack of some textures. The change of a certain position on the sketch will cause the global change of the generated image.

In this paper, the features of the middle layer of pix2pixHD generator are analyzed by visualization method, the defects of its network structure are found, and the model is improved. The improved model produces higher quality, richer texture details, and realizes the function of image editing.

The training data are also augmented to make the training and testing process more robust.

Key Words: image translation; hand-drawn sketch; image editing; visualization

第一章 绪论

第一节 研究背景

人脸图像生成一直是计算机图形和视觉领域的热门方向，包括人脸姿态仿真^[1]、遮挡人脸恢复^[2]、年龄与表情仿真^[3-4]、面部属性编辑^[5]、人脸艺术生成^[6]、人脸美颜与风格化^[7]等等。而基于草图的人脸图像生成是这一方向的重要分支。

当我们想要从零开始描绘一个物体时，手绘草图无疑是最直观、最有效的方式之一，它比文字描述更清晰具象，可以将抽象的概念转化为具体的视觉表达。随着触屏设备的广泛普及，随手绘制草图变得越来越简单容易。所以如何由手绘草图生成真实的图像一直吸引着领域内研究人员的关注。由手绘草图生成真实的人脸有着广泛的应用。比如在刑侦案件中，根据目击者的手绘草图重塑嫌疑人的脸图像；还可以根据我们自己的兴趣，随心所欲地将手绘的人脸草图转化成真实的人脸照片等等。而且，草图有着很强的通用性，是一种跨越国家和文化的沟通交流方式，不同国家和民族的人们都可以通过手绘草图描绘自己的所见所闻，表达自己的所思所想。草图简单明了，但却可以捕捉对象的结构特征和纹理细节，包含了足够的信息，可以作为有效的输入来生成高质量的图像。

经过近二十年的发展，由草图生成真实图像的方案主要有两种。一种是传统的方法，图像检索。该方法需要基于一个庞大的数据库，将手绘草图分割成独立的部分，利用搜索算法分别检索出与输入的草图相匹配的图像，然后将所有部分进行融合，生成一张完整的图像。该方法缺点很明显，那就是需要大量的数据，保证数据库中有与输入草图相匹配的内容，并且后期的融合会有很明显的痕迹。这种方法显然不适合人脸图像的生成。

另一种是基于深度学习的方法，图像翻译^[8]。图像翻译的目标是给定输入-输出图像对作为训练数据，将输入图像从一个域 \mathcal{A} （源域）转换到另一个域 \mathcal{B} （目标域）。如果用 $\alpha(x, A) \in \mathcal{A}$ 和 $\beta(x, B) \in \mathcal{B}$ 来表示域 \mathcal{A} 和域 \mathcal{B} 的图像，图像翻译的任务可以描述为，寻找一个合适的变换 $f : \mathcal{A} \rightarrow \mathcal{B}$ ，使得

$$f(\alpha(x, A)) = \beta(x, B) \quad (1.1)$$

该方法基于生成对抗网络 GAN^[9]，并且是以图像为输入的条件生成对抗网络^[10]。在训练阶段，需要大量成对的数据作有监督的学习。条件生成对抗网络的网络结构分为两部分，即生成器和判别器。训练时将草图输入生成器，生成一张

图像，然后将生成的图像与真实的图像分别输入判别器，以判别器的输出代入损失函数，分别计算生成器与判别器的损失，通过后向传播算法更新网络参数，使生成器产生更接近真实图像的输出。图像翻译的概念最早由朱俊彦在 pix2pix^[8]这篇论文中提出，而后又在此基础上发表了另一篇文章 pix2pixHD^[11]，可以生成质量更高的图像。本文的主要工作就是围绕 pix2pixHD 展开。

第二节 本文工作

本文旨在探究利用 pix2pixHD 实现从草图到人脸照片生成的过程，发现对于手绘草图的输入，其输出结果的质量不够令人满意，特别是某些细节比较模糊，而且草图上某一位置的改变常引起生成结果全局的变化。针对这一缺陷，我们用可视化的手段分析了草图特征，对其网络结构进行了深度剖析，在此基础上改造了其生成器的网络结构，我们改进后的模型对手绘草图输入的生成图像质量更高，并且能更好地实现图像编辑的功能。

首先，本文用人脸轮廓-照片数据对训练 pix2pixHD 的原始模型，然后用手绘的草图进行测试，发现其生成的人脸照片不尽人意，具体表现为纹理特征不够清晰，丧失了对细节的控制，如果在草图上改动一处可能影响生成的整体结果。

而后，利用 pca、tsne 等可视化工具对生成器网络提取到的草图特征进行可视化分析，结合对网络结构的研究，本文发现生成器编码图片所用的实例标准化层是造成这一现象的根本原因。接下来我们针对生成器中的实例标准化层进行了消融实验，做对比研究，发现去掉前两层实例标准化操作的模型生成质量更高，且能更好地对输入草图进行编辑。

针对模型对于输入草图空间位置变化的鲁棒性差的问题，我们用增广的数据重新训练了模型，成功解决了这一问题。

第二章 相关模型介绍

第一节 pix2pixHD 的模型结构

一、简介

pix2pixHD 是一种基于条件生成对抗网络的图像翻译模型，能够从输入的语义标签图出生成高质量、高分辨率的图像。

在它之前的方法存在两个问题：1. 很难生成高分辨率的图像，如 pix2pix^[8]；2. 生成的图像缺少细节特征和真实的纹理，如 CRN^[12]。

所以，为了解决以上两个问题，该模型提出了一个新的鲁棒性更强的对抗学习损失函数，并采用了新的多尺度生成器和判别器结构。运用该模型可以生成分辨率达到 2048×1024 的真实感更强的图像，效果超越了以前的方法。

此外，该模型还可以对生成图像进行交互的图像编辑。首先，可以在输入中加入对象的实例分割信息，实现对物体的编辑，比如在生成图像中增删物体或者改变物体的类别；其次，给定相同的输入，可以编辑生成图像中某一物体的外观。

二、网络结构

1. 多层级的生成器

生成器网络分为两个子网络： G_1 和 G_2 ，如图 2.1 所示。 G_1 称为全局生成器网络 (global generator network)， G_2 称为局部增强器网络 (local enhancer network)。 G_1 用来生成基础图像，而 G_2 用来提高图像的分辨率。为了进一步提高分辨率，甚至可以继续叠加 G_2 。

G_1 由 3 部分组成，分别是一个卷积前端 $G_1^{(F)}$ ，一系列的残差模块 $G_1^{(R)}$ 和一个逆卷积后端 $G_1^{(B)}$ 。 $G_1^{(F)}$ 由 5 个卷积层组成， $G_1^{(R)}$ 由 9 个残差模块组成， $G_1^{(B)}$ 由 4 个逆卷积层和最后的 1 个卷积层组成。这种结构被证明用在神经风格迁移任务中很成功。网络中每一个卷积和逆卷积操作后都用实例标准化进行处理，使训练过程更加鲁棒。

G_2 也是由 3 部分组成：一个卷积前端 $G_2^{(F)}$ ，一系列的残差模块 $G_2^{(R)}$ 和一个逆卷积后端 $G_2^{(B)}$ 。其中， $G_2^{(F)}$ 由 2 个卷积层组成， $G_2^{(R)}$ 由 3 个残差模块组成， $G_2^{(B)}$ 由 1 个逆卷积层和 1 个卷积层组成。 $G_2^{(F)}$ 输出的特征图与 $G_1^{(B)}$ 第 4 个逆卷积层的特征图进行每像素相加，相加的结果作为 $G_2^{(R)}$ 的输入。

在训练阶段，先分别训练 G_1 和 G_2 ，然后将两部分合并训练，对网络参数进行微调。这种由粗到精的网络结构可以有效地融合局部和全局的信息，最终生成高分辨率、高真实感的图像。

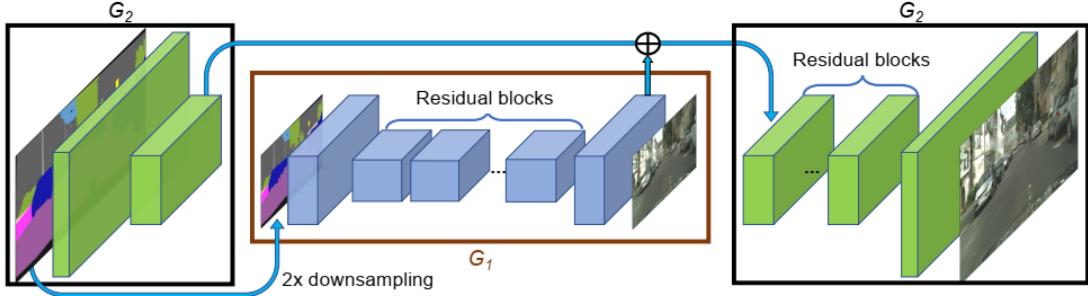


图 2.1 pix2pixHD 生成器网络结构

2. 多尺度的判别器

为了判别高分辨率的真实图像与生成图像，判别器需要有大的感知野。如果通过增加网络深度或者提高卷积核尺寸来解决，将占用大量的显存，并且容易导致过拟合的问题。

所以该模型采用多尺度的判别器，由粗到精地分辨生成图像和真实图像。具体来说，使用了 3 个网络结构完全相同的判别器，分别称为 D_1 , D_2 和 D_3 。将用于训练的真实图像和生成图像进行 $1/2$ 和 $1/4$ 降采样，连同原始尺寸图像，分别作为 D_1 , D_2 和 D_3 的输入。对于以最小尺寸图像为输入的 D_2 来说，其拥有最大的感知野，能更好的对全局信息进行判别，因而可以使生成图像在全局视野下更真实；对于以原始尺寸图像为输入的 D_3 来说，其拥有最精细的粒度，对局部信息更敏感，可以促进细节的生成。

判别器的网络结构比较简单，由 5 个降采样的卷积层堆叠而成，其卷积核大小为 4。

三、目标函数

该模型采用了 3 个目标函数，第一个是常规的对抗生成损失 $L_{GAN}(G, D_k)$ ，第二个是特征匹配损失 $L_{FM}(G, D_k)$ ，第三个是 VGG 损失 $L_{VGG}(G)$ ，下面分别对这 3 个目标函数进行介绍。

1. 对抗生成损失

$L_{GAN}(G, D_k)$ 是普通的对抗生成损失，没有需要特别说明之处。其数学表达如下所示：

$$L_{GAN}(G, D_k) = \mathbb{E}_{(s, x)}[\log D_k(s, x)] + \mathbb{E}_s[\log(1 - D(s, G(s)))] \quad (2.1)$$

其中， s 表示语义标签图的域， x 表示对应的真实图像的域， \mathbb{E}_s 表示 $\mathbb{E}_{s \sim p_{data}(s)}$ ， $\mathbb{E}_{(s, x)}$ 表示 $\mathbb{E}_{(s, x) \sim p_{data}(s, x)}$ 。

2. 特征匹配损失

首先用多尺度判别器提取生成图像和真实图像的特征，然后对两类特征求 L_1 损失，使特征尽可能相似。这种匹配中间过程表达的做法可以帮助稳定训练过程。其数学形式为：

$$L_{FM}(G, D_k) = \mathbb{E}_{(s, x)} \sum_{i=1}^T \frac{1}{N_{ki}} [\|D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s))\|_1] \quad (2.2)$$

$D_k^{(i)}$ 表示判别器 D_k 的第 i 层特征提取器， T 表示判别器中特征提取的层数， N_{ki} 表示第 k 个判别器第 i 层特征图的元素总数。

3. VGG 损失

首先用预训练好的的 VGG19 模型作为特征提取器，将生成图像和真实图像分别输入 $VGG19$ ，得到中间层的特征，然后对两类特征求 L_1 损失。其本质也是一种感知损失，按如下方式计算：

$$L_{VGG}(G) = \mathbb{E}_{(s, x)} \sum_{i=1}^L \frac{1}{M_i} [W_i \cdot \|V^{(i)}(x) - V^{(i)}(G(s))\|_2] \quad (2.3)$$

其中， $V^{(i)}$ 表示 $VGG19$ 中的第 i 层特征提取器， W_i 为每层的权重参数， M_i 代表每层特征图的元素总和， L 为特征提取器的层数。

pix2pixHD 总的目标函数如下所示：

$$\min_G (\max_{D_1, D_2, D_3} \sum_{k=1,2,3} L_{GAN}(G, D_k)) + \lambda (\frac{1}{3} \sum_{k=1,2,3} L_{FM}(G, D_k) + L_{VGG}(G(s))) \quad (2.4)$$

第二节 实例标准化

实例标准化 (IN)^[13] 是众多标准化方法中的一种，由 Ulyanov 等人在 2017 年提出，在风格迁移、图像翻译等任务上有很好的作用效果。实例标准化通过计算

单样本单通道的均值和方差对输入进行标准化，首先对输入的单张特征图的某一通道的像素值求均值和方差，像素值减去均值后再除以方差进行标准化，之后再进行放缩和平移。其实现过程的数学表达形式如公式(2.5)~(2.8)所示：

$$\mu_{ni} = \frac{1}{HW} \sum_{l=1}^W \sum_{m=1}^H x_{nilm} \quad (2.5)$$

$$\sigma_{ni}^2 = \frac{1}{HW} \sum_{l=1}^W \sum_{m=1}^H (x_{nilm} - \mu_{ni})^2 \quad (2.6)$$

$$\hat{x}_{nijk} = \frac{x_{nijk} - \mu_{ni}}{\sqrt{\sigma_{ni}^2 + \epsilon}} \quad (2.7)$$

$$y_{nijk} = \gamma \hat{x}_{nijk} + \beta \quad (2.8)$$

用 $\mathbf{x} \in \mathbb{R}^{N \times C \times W \times H}$ 表示一个批次的图片或特征， $x_{nijk} \in \mathbf{x}$ 表示其中的一个元素，其中， n 表示在批次中的索引， i 表示特征通道， j, k 表示空间位置。 γ 和 β 是可学习的参数，分别对标准化后的值进行放缩和平移。 ϵ 是为了防止方差为 0 而加入的微小的正数。

实例标准化有以下几个作用：

首先，对特征进行标准化，使其服从均值为 0、方差为 1 的正态分布，缩小了同一通道内不同元素间的差异，从而避免了梯度消失和梯度爆炸的问题，使训练过程更加稳定。

其次，通过减少梯度对初始值尺度的依赖，从而可以用较大的学习率训练网络，从而加速网络的收敛。

其实，实例标准化与批标准化非常相似，唯一的不同是批标准化对一个批次内所有的特征图求均值和方差，而实例标准化只对单个样本进行计算。对于图像翻译这类任务，重点关注每张图像的内容，可以把每个样本都看成一个单独的域，因此批归一化便不再适用。因为批归一化考虑批次内所有样本的信息，从而造成每个样本细节信息的丢失。而实例归一化只考虑单样本单通道的信息，更适合对每像素有更高要求的任务。

实例归一化沿用了批归一化的 γ 和 β 两个参数，可以自适应地调节输出的范围，使其不致受限于标准高斯分布，提高了网络的表达能力。

第三节 模型引入

我们对 pix2pixHD 的模型做了一些改进, 在其全局生成器 G_1 的基础上, 去掉了前两层的实例标准化操作, 作为我们自己的生成器。生成器的结构如图 2.2 所示:

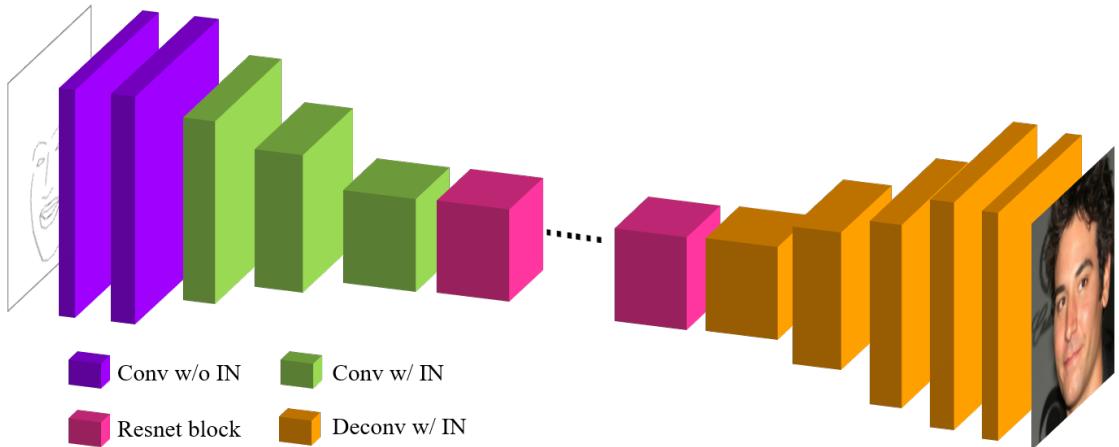


图 2.2 我们的模型的生成器网络结构

判别器沿用了多级判别器的设计, 并利用了 PatchGAN^[8] 的思路, 最终判别器输出的是一个矩阵而非一个标量。若用 $D_k, k = 1, 2, 3$ 表示我们的判别器, 其中 D_k 的网络结构如图 2.3 所示:

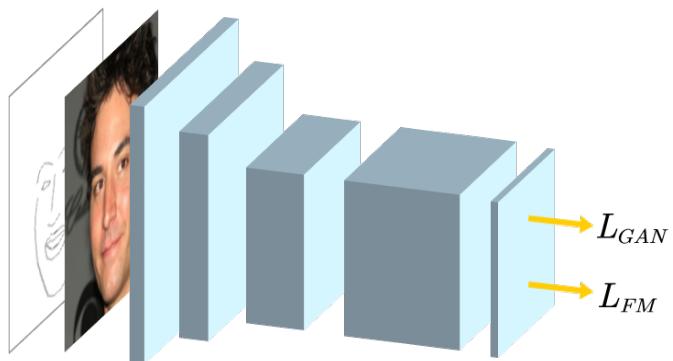


图 2.3 我们模型的判别器 D_k 的网络结构

我们把草图和人脸照片在通道方向连接起来作为判别器的输入, 判别器的输出用来计算生成对抗损失 L_{GAN} 和特征匹配损失 L_{FM} 。我们的目标函数跟 pix2pixHD 中保持了一致, 也是由 L_{GAN} 、 L_{FM} 和 L_{VGG} 3 部分组成。

第四节 数据集生成

一、CelebA-HQ 数据集

CelebA 数据集是香港中文大学的开放数据库，搜集了 10177 位名人的 202599 张人脸照片，并且对每张图片标注了人脸特征点 (landmark)。而 CelebA-HQ 数据集是在 CelebA 的基础上，利用 PGGAN^[14] 生成的一个新的数据集。

CelebA-HQ 包含 30000 张分辨率为 1024×1024 人脸照片，因为裁剪图片时以人脸特征点为参照，所以每张图片都是全局对齐的，即有一个标准化的位置和角度。

二、轮廓数据集

我们从 CelebA-HQ 数据集的人脸照片中提取了 68 个特征点，然后用像素值为 2 的线将这些点依次连接，就形成了轮廓图。

为了实验需要，我们将图片进行了缩放，每张图的分辨率为 512×512 。经过挑选，训练集中包含 14973 对轮廓和人脸照片，测试集中包含 4992 对轮廓和人脸照片。

以轮廓作为草图有很多好处，其中重要的一点是它比边缘图更加简洁，更符合我们手绘的习惯。因此用轮廓拟合手绘草图的数据分布更加合理。

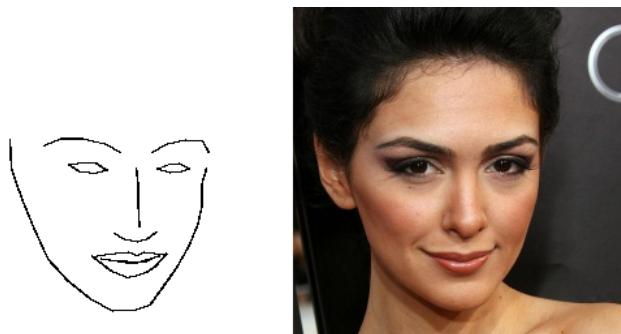


图 2.4 轮廓图数据集示例

三、手绘草图的生成

为了测试模型对手绘草图的真实效果，我们开发了一个用于绘制和收集草图并能实时生成、展示结果的交互界面。我们在绘制窗口设置训练草图的平均脸为底图，作为手绘时的参照。设置笔触的大小为 2 像素值，从而保证与训练数据线条宽度相同。我们可以开放交互界面让不同的人来画，绘制的草图会自动存

入我们的手绘草图数据库，从而可以不断丰富扩大我们的数据集。交互界面如图 2.5 所示：

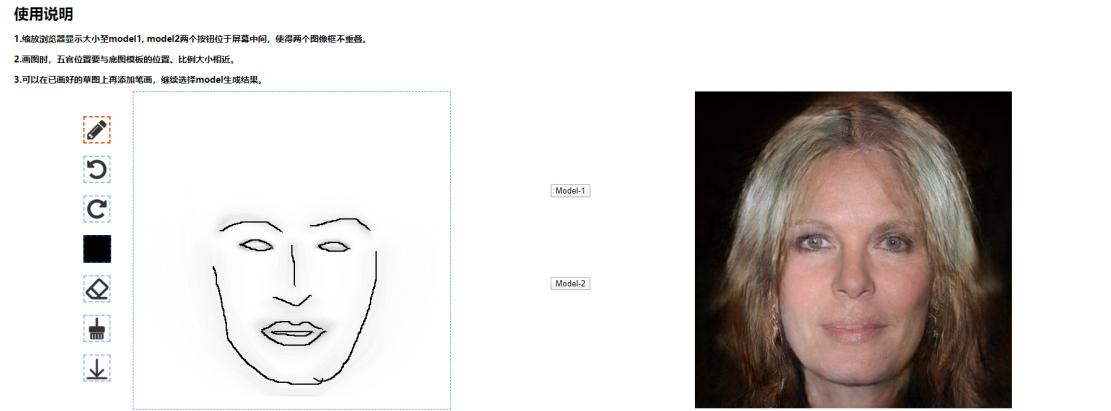


图 2.5 手绘草图交互界面示意图

第三章 特征可视化

第一节 t-SNE 原理

高维数据的可视化在许多不同的领域都一直是一个非常重要的问题，近几十年来提出了许多可视化的方法，而数据降维是其中非常重要的一种。降维，顾名思义就是将高维空间中的数据降低维度。通常变成二维或三维的形式，方便以散点图的形式展现，帮助人们更好地分析数据。降维后的数据能否保留原始数据之间的关系以及降维后的可视化效果成为衡量一种降维方法优劣的两大标准。

降维方法一般分为线性降维和非线性降维，线性降维包括主成分分析(PCA)、多维尺度变换(MDS)等，非线性降维包括等度量映射(Isomap)、局部线性嵌入(LLE)和随机近邻嵌入(SNE)等。线性降维通常会将高维空间中不同的数据点在低维表示中远远分开，对于高维空间中相同或相近的数据点其表示能力则存在不足。非线性降维可以很好地弥补这一缺点。t-SNE便是一种非线性降维方法，它在SNE的基础上做了几点改进。

一、SNE

随机近邻嵌入首先将高维空间数据点之间的欧氏距离转换为表示相似性的条件概率分布。考虑高维空间中两个数据点 \mathbf{x}_i 和 \mathbf{x}_j ，以 $p_{j|i}$ 表示 \mathbf{x}_i 选择 \mathbf{x}_j 作为其近邻点的条件概率。 \mathbf{x}_j 与 \mathbf{x}_i 的欧氏距离越小，则 $p_{j|i}$ 越大。如果用高斯分布描述这种条件概率，则其数学表达形式为：

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)} \quad (3.1)$$

其中 σ_i 表示以 \mathbf{x}_i 为中心的高斯分布的方差。设 \mathbf{x}_i 和 \mathbf{x}_j 映射到低维空间的点分别为 \mathbf{y}_i 和 \mathbf{y}_j ，同样地，用条件概率 $q_{j|i}$ 表示低维空间点的相似性，其数学表达形式为：

$$q_{j|i} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)} \quad (3.2)$$

为了使 \mathbf{y}_i 和 \mathbf{y}_j 可以真实地反映 \mathbf{x}_i 和 \mathbf{x}_j 之间的关系，理论上应该让条件概率 $p_{j|i}$ 与 $q_{j|i}$ 完全相等。用 \mathbf{P}_i 表示给定 \mathbf{x}_i 与其余所有点之间的条件概率分布， \mathbf{Q}_i 表示 \mathbf{y}_i 与其余各点之间的条件概率分布，应该使 \mathbf{Q}_i 与 \mathbf{P}_i 完全相等，所以用 \mathbf{P}_i 与

Q_i 的 KL 距离作为代价函数:

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (3.3)$$

由于 KL 散度是非对称的, 因此在低维空间中, 不同类型的误差在两两距离上的权重并不相等。特别地, 如果高维空间中相距很近的点被映射到低维空间后距离很远, 将得到很大的代价; 而如果在高维空间中相距很远的点被映射成距离很近的点, 得到的代价不会很大。这说明, SNE 的一大缺陷就是更关注局部结构而忽视了全局。

代价函数 C 对 \mathbf{y}_i 求梯度后有一个简单的形式:

$$\frac{\delta C}{\delta \mathbf{y}_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(\mathbf{y}_i - \mathbf{y}_j) \quad (3.4)$$

常常利用带动量的随机梯度下降算法优化低维空间点的分布:

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}) \quad (3.5)$$

二、t-SNE

t-SNE 在 SNE 的基础上做了两点改进:

1. 将原始的 SNE 转变成对称 SNE

在原来的 SNE 中, $p_{j|i} \neq p_{i|j}$, 且 $q_{j|i} \neq q_{i|j}$, 是非对称的, 则在高维空间和低维空间中分别构造出对称的联合概率分布 P 和 Q , 使得对任意的 i, j , 均有 $p_{ij} = p_{ji}, q_{ij} = q_{ji}$ 。构造的 p_{ij} 有如下形式:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \quad (3.6)$$

其中, n 为数据点的数量, 这种定义既满足了对称性, 也保证了对于离群点 \mathbf{x}_i 代价惩罚不致太小。代价函数重写为:

$$C = KL(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (3.7)$$

梯度函数变为:

$$\frac{\delta C}{\delta \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j) \quad (3.8)$$

可见比之前的形式更为简洁, 计算效率更高。

2. 将低维空间中的数据分布表示为 t 分布

t 分布是长尾分布，相比于正态分布可以更好地将高维数据映射到低维空间。举例来说，在高维空间中距离相近的点，为了满足 $p_{ij} = q_{ij}$ ，映射到低维空间后距离会更近；而在高维空间中距离较远的点，映射到低维空间后距离会拉大。这使得同类别的点或者说相似的点在低维空间分布得更为紧密，而不同类别的点或者说相似性较低的点分布得更为稀疏，从而使可视化的效果更好。

用 t 分布重新定义低维空间的联合概率分布：

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \quad (3.9)$$

代价函数与之前相同，梯度的形式变为：

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \quad (3.10)$$

之后用带动量的随机梯度下降算法优化低维空间的数据表示 \mathcal{Y} 即可。

第二节 草图特征分析

为了验证现有图像翻译神经网络对精确度不高、带有几何变形的手绘草图是否能够提取到与作者意图相符的人脸形状特征，我们手绘了 198 张草图，分辨率为 512×512 。这些草图可以划分成 11 类，分别是 G1：添加头发；G2：添加新属性，例如胡须、皱纹、耳朵；G3：改变脸型；G4：改变眉毛；G5：改变眼睛形状；G6 改变眼睛大小；G7：涂鸦型手绘；G8：改变嘴巴；G9：改变鼻子；G10：改变嘴巴（眼睛与 G9 相同）；G11：改变鼻子（眼睛与 G8 相同）。具体的，G7 的涂鸦型手绘草图各图之间没有关联。而其余 10 类中草图都只改变一个特定位置或属性而其他部分保持不变。除 G8 和 G11、G9 与 G10 之间眼睛部位相同以外，其余各类草图的眼睛线条各不相同。图 3.1 展示了这 11 组手绘草图示例。

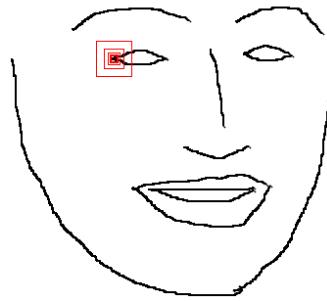
如第二章所述，我们选用了目前较为先进的图像翻译神经网络 pix2pixHD^[11] 作为我们从手绘草图到人脸图像的网络模型。

我们以 G_1 指代全局生成器， G_2 指代局部增强器。由于外层的 G_2 只是单纯为了提高生成图像的分辨率，为了可以更直接地探究问题的本质，我们截取了 G_1 作为模型的生成器。由前文 2.2.1 可知， G_1 由 3 部分组成，分别是：一个卷积前端 $G_1^{(F)}$ ，9 个残差模块 $G_1^{(R)}$ 和一个逆卷积后端 $G_1^{(B)}$ ，而 $G_1^{(F)}$ 由 5 个下采样的卷积层组成。我们分别称 $G_1^{(F)}$ 的前 5 个卷积层为 $L0 \dots L4$ 。由 2.3 可知，我们的模型在 pix2pixHD 的 G_1 的基础上去掉了 $L0$ 和 $L1$ 的实例标准化操作，作为我们自己的生成器模型。

类别	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11
样本数	26	15	12	17	19	9	12	26	19	26	19
手绘草图示例											

图 3.1 分类手绘草图示意图

把手绘草图输入我们模型的生成器，得到 L_0 到 L_4 的 5 张特征图，然后对应输入草图的左眼眼角位置，坐标为 $(170, 250)$ ，分别在 5 张特征图上抽取一个点的全通道特征向量，我们称之为 $\mathbf{v}_k, k = 0, \dots, 4$ 。5 张特征图上对应点的坐标分别为 $(170, 250)、(85, 125)、(43, 63)、(22, 32)$ 和 $(11, 16)$ ，它们在输入草图上的感知野大小分别为 7、9、13、21 和 37。感知野的示意图如图 3.2 所示：

图 3.2 $L_0 \sim L_4$ 特征图的左眼角点在输入草图上的感知野

同样地，我们也将手绘草图输入 pix2pixHD 的 G_1 ，按上述方法得到的 5 个特征向量分别称为 $\mathbf{v}'_0 \dots \mathbf{v}'_4$ 。这 5 个特征向量的维数分别为 48、96、192、384 和 768。

将 198 张输入草图的 \mathbf{v}_0 和 \mathbf{v}_1 分别放在一起比较，发现没有改变眼睛的组，比如改变嘴巴的 G8 和改变鼻子的 G9，这两个向量在类内是一致的；而改变了眼睛的组，如改变眼睛形状的 G5、改变眼睛大小的 G6、涂鸦型手绘 G7 三组，这两个向量在它们各自的类内是不相等的。把不同草图的 \mathbf{v}'_0 和 \mathbf{v}'_4 分别放在一起比

较，发现在各自的类内这两个向量都是不相等的。针对发现的这个问题，我们对提取的特征向量进行了可视化的分析。

第三节 PCA 可视化分析

主成分分析 (PCA) 是一种常用的线性降维方法，可以较好地保留数据在高维空间中的特征。接下来我们用 PCA 对 $v_0 \sim v_4$ 进行降维后的可视化分析。

图 3.3(a) 和图 3.3(b) 分别是对 v_0 和 v_1 用 PCA 降维后可视化的结果，图 3.3(c) 是图例 (该图例对以下第三章的内容都适用)。从图中可以看出 G1、G2、G3、G4 类的数据各自分散地分布于同一点，G8 和 G11 类数据分布于同一点，G9 和 G10 类数据分布于同一点，而 G5、G6、G7 类的数据分布的非常分散。这说明去掉前两层的实例标准化操作之后，左眼角角点位置的特征只受感知野内的草图内容的影响，感知野内草图的改变会使提取到的对应点的特征发生变化，而改变草图其他部分不会对该点的特征产生作用。

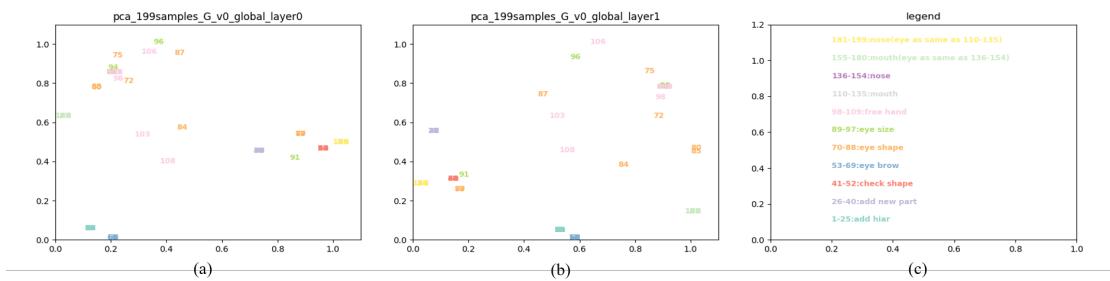


图 3.3 v_0 和 v_1 的 PCA 可视化结果

$v_2 \sim v_4$ 的可视化结果分别如图 3.4(a)(b)(c) 所示。观察 v_2 的 PCA 可视化结果可以发现，没有改动眼睛的类别，类内的特征分布开始不同，说明实例标准化操作将草图上其他位置的变化传达到了眼睛位置，使眼睛位置的特征发生改变，而不单单只影响改动位置的特征。

对比 v_2 和 v_3 的 PCA 可视化结果，可以发现，没有改动眼睛的草图类别，如 G1、G2、G3、G4、G8、G9、G10、G11 类，类内距离逐渐增大，即同一个类别内特征分布得更为分散。这说明随着逐层增加实例标准化操作，在输入草图上改动其他部位对眼睛位置的特征影响越来越明显，导致最终生成图像的眼睛也发生改变。

对比从 v_2 到 v_4 的 PCA 可视化结果，还可以发现，随着卷积层数的加深，G1、G2、G3、G4、G8、G9、G10、G11 类特征的类间距离逐渐减小，类间界限逐渐模

糊，即不同类别的特征分布得更为集中。这说明随着逐层增加实例标准化操作，由于实例标准化的平均化作用，即使在输入草图上改变眼睛，对特征图上眼睛位置的影响作用也越来越弱。这种现象最终导致生成图像的纹理细节变得模糊，不够清晰。

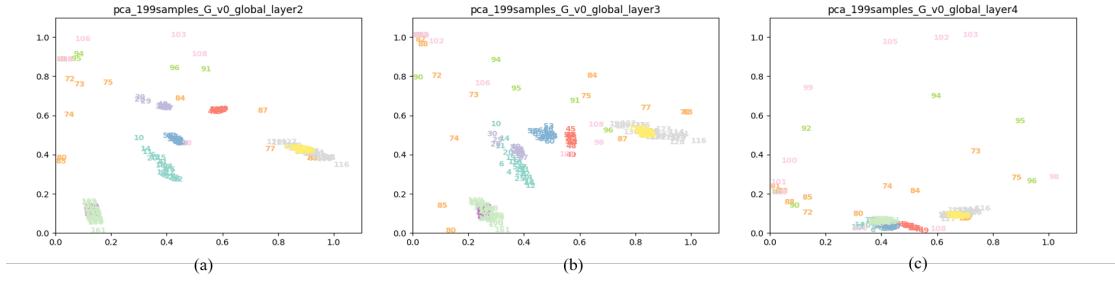


图 3.4 $\mathbf{v}_2 \sim \mathbf{v}_4$ 的 PCA 可视化结果

接下来做的是模型的对比实验，即，把 $\mathbf{v}_0 \sim \mathbf{v}_4$ 的可视化结果与 $\mathbf{v}'_0 \sim \mathbf{v}'_4$ 进行横向对比。

将 $\mathbf{v}'_0, \mathbf{v}'_1$ 的可视化结果分别与 $\mathbf{v}_0, \mathbf{v}_1$ 的进行对比，图 3.5(a)(c) 分别为 \mathbf{v}_0 和 \mathbf{v}_1 的可视化结果，图 3.5(b)(d) 分别为 \mathbf{v}'_0 和 \mathbf{v}'_1 的可视化结果。因为前两个卷积层有实例标准化的缘故，所以 G1、G2、G3、G4、G8、G9、G10、G11 类的特征向量 \mathbf{v}'_0 和 \mathbf{v}'_1 各自分散地分布，而没有像 \mathbf{v}_0 与 \mathbf{v}_1 一样重合于一点。

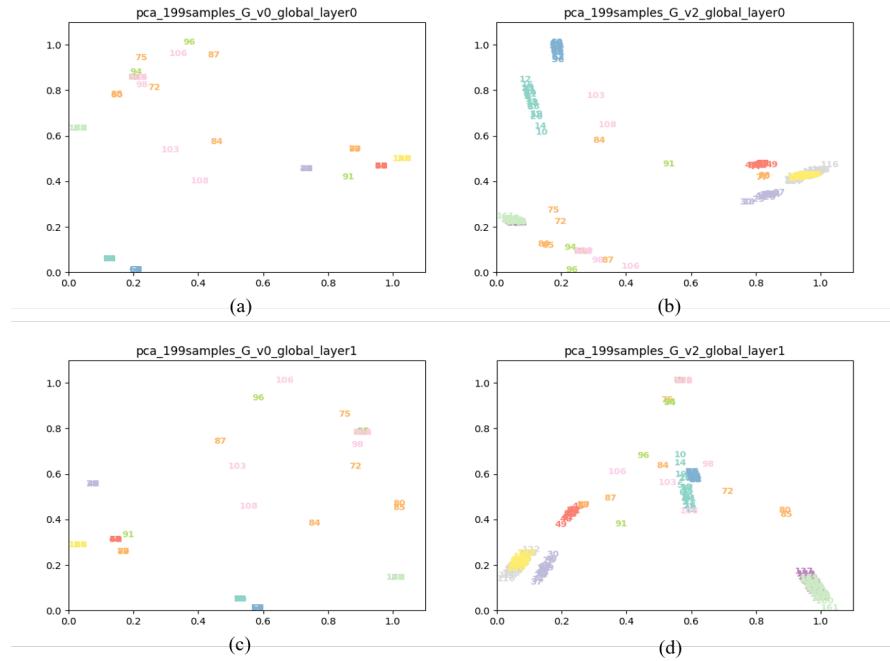


图 3.5 $\mathbf{v}_0, \mathbf{v}_1$ 与 $\mathbf{v}'_0, \mathbf{v}'_1$ 的 PCA 可视化结果对比

分别将 $\mathbf{v}'_2 \sim \mathbf{v}'_4$ 与 $\mathbf{v}_2 \sim \mathbf{v}_4$ 的可视化结果进行横向对比，如图 3.6 所示，(a)(b)(c)

分别为 $\mathbf{v}_2 \sim \mathbf{v}_4$ 的 PCA 可视化结果, (d)(e)(f) 分别为 $\mathbf{v}'_2 \sim \mathbf{v}'_4$ 的可视化结果。我们发现 \mathbf{v}'_2 , \mathbf{v}'_3 , \mathbf{v}'_4 中没有改变眼睛的类别, 如 G1、G2、G3、G4、G8、G9、G10、G11 类, 其类内距离更大, 类别内部分布得更为分散, 而 \mathbf{v}_2 , \mathbf{v}_3 和 \mathbf{v}_4 则有效减少了类内距离, 使类别内部分布得更为集中。这恰恰说明了我们将 pix2pixHD 的 G_1 的前两个卷积层的实例标准化操作去掉, 可以更好地提取到输入草图的低层特征, 从而可以有效降低在草图上改变某一部位而对生成图像其他部位产生的影响, 同时也丰富了生成图像的纹理细节。

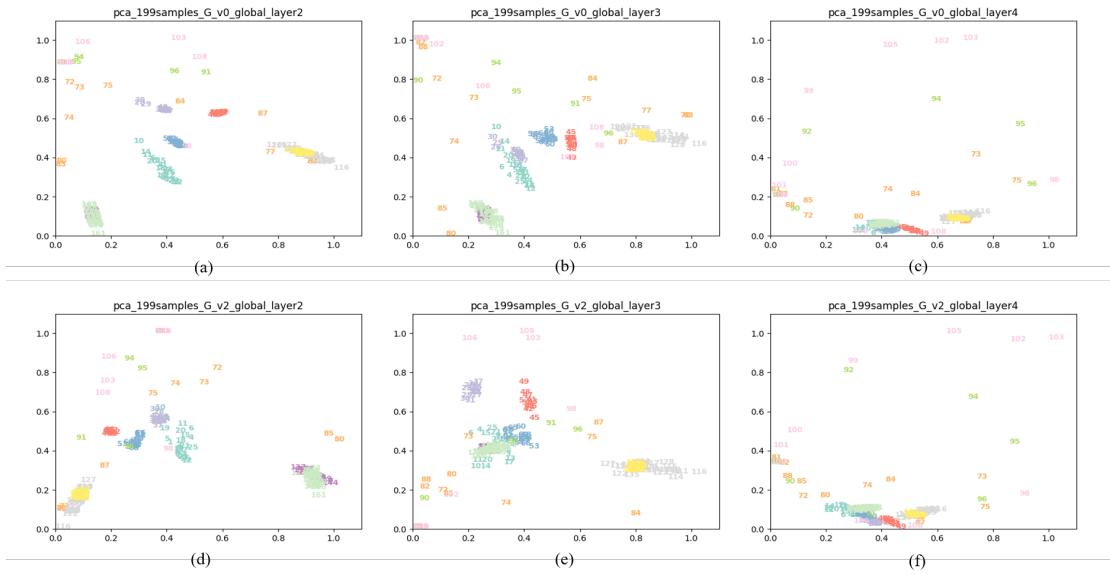


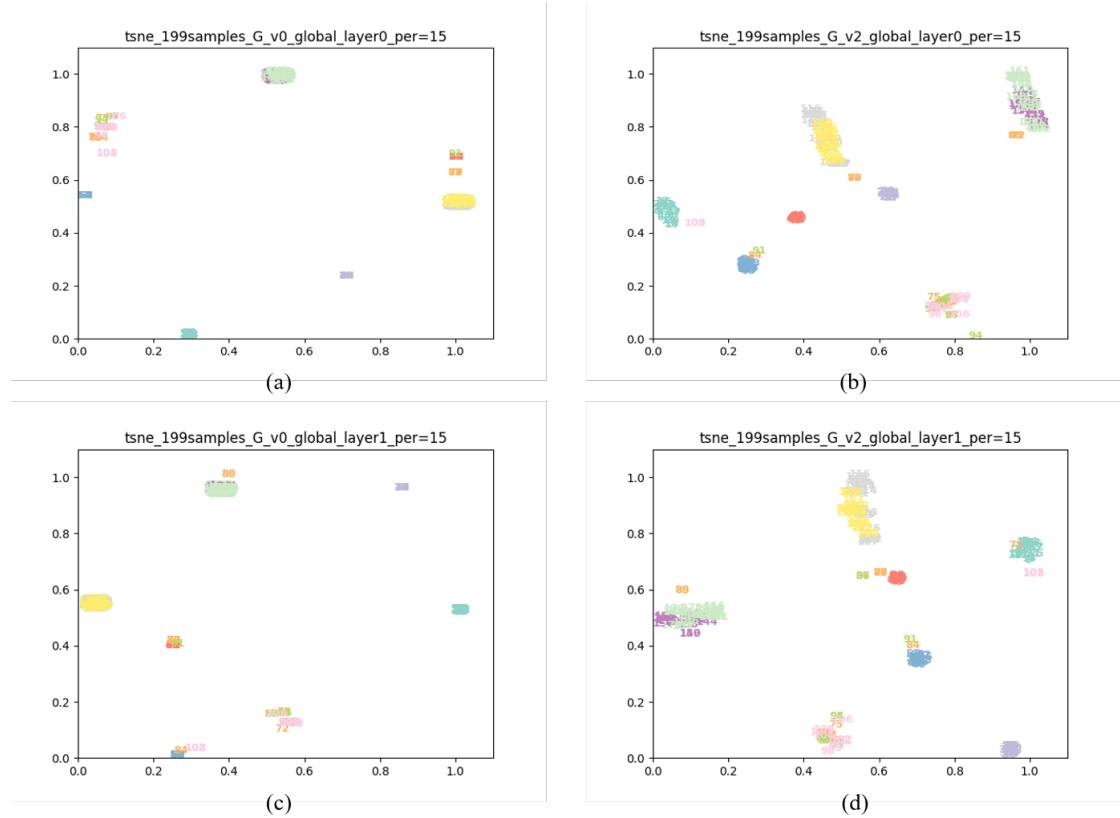
图 3.6 $\mathbf{v}_2 \sim \mathbf{v}_4$ 与 $\mathbf{v}'_2 \sim \mathbf{v}'_4$ 的 PCA 可视化结果对比

第四节 t-SNE 可视化分析

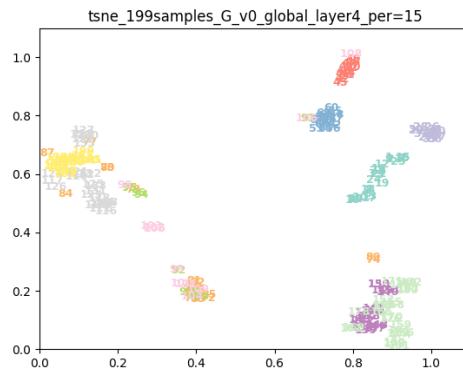
t-SNE 的困惑度参数 (perplexity) 代表了在求高维空间数据分布时考虑的近邻点的个数, 由于我们的草图数据每类的数量平均为 15 左右, 所以取困惑度的值为 15。

图 3.7(a)(b) 分别展示了 \mathbf{v}_0 与 \mathbf{v}'_0 的 t-SNE 可视化结果, 图 3.7(c)(d) 分别展示了 \mathbf{v}_1 与 \mathbf{v}'_1 的 t-SNE 可视化结果。从图中可以发现, G1、G2、G3、G4、G8、G9、G10、G11 类的 \mathbf{v}_0 、 \mathbf{v}_1 各自聚类在一起, 各类集中地分布, 而 \mathbf{v}'_0 和 \mathbf{v}'_1 则在类内分布得比较分散。这也说明在 G_1 的 L_0 , L_1 去掉实例标准化操作后, 在草图上改变其他部位对前两层特征图的眼睛位置没有影响。

另外, 我们观察 \mathbf{v}_4 的可视化结果, 由图 3.8 可以发现, G8 与 G11 类、G9 与 G10 类的 \mathbf{v}_4 还是分布在一起, 并没有互相分离。这说明, 对于比较深层的眼睛

图 3.7 v_0, v_1 与 v'_0, v'_1 的 t-SNE 可视化结果对比

部位的特征来说，在草图上修改眼睛依然是最重要的改变，在草图眼睛相同的情况下，无论改变嘴巴还是鼻子都不能对眼睛特征产生足以区分类别的显著影响。输入数据的局部变动对提取到的特征的影响也是局部的，而非全局性的。

图 3.8 v_4 的 t-SNE 可视化结果

根据以上观察可以说明，我们的模型能更好的保留草图最底层的特征和最原始的语义信息。

第四章 实验结果及分析

第一节 数据增广

一、输入数据的全局对齐问题

当我们用手绘草图测试训练好的模型时，发现如下问题：当手绘草图的人脸特征点与参考底图基本重合时，生成效果非常不错；但是当我们手绘草图的位置偏离底图时，生成效果则非常糟糕。

由于 CalebA-HQ 数据集的人脸照片是以面部特征点为参照进行裁剪的，我们的训练数据——轮廓图也是基于面部特征点获取的，所以我们直觉地认为所有训练数据的特征点都具有空间一致性。

为证实以上假设，我们计算了所有训练数据的平均脸，如图 4.1 所示。我们发现其五官和面部轮廓基本处于相同位置，换句话说训练数据是全局对齐的。



图 4.1 轮廓图平均脸

这就导致了测试结果对输入数据空间分布变化包容性差的问题。使用者便只能被限制在固定的区域，不能随心所欲地在界面上不同位置绘制草图了。

二、训练数据的增广

我们通过对训练集的草图做一定程度的平移和旋转来模拟用户手绘草图空间位置的变化。平移在水平和竖直两个方向上进行，平移的距离在 ± 25 像素之间随机选取，旋转的范围在 $-7^\circ \sim +7^\circ$ 之间。但是人脸照片并不做平移和旋转处理，因为我们期望无论输入草图的空间位置如何，生成结果始终保持全局对齐。我们用增广后的数据重新训练模型。

用手绘草图对模型进行测试，发现对于不同方向产生位移的草图，使用增广数据训练后的模型的生成效果要明显优于原始模型，如图 4.2。

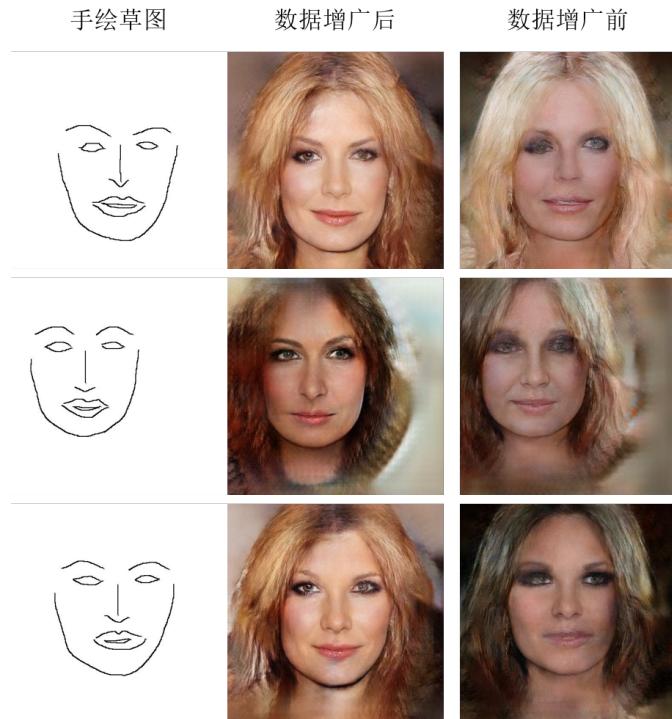


图 4.2 数据增广前后生成结果对比

第二节 我们模型的生成效果及对比实验

一、针对实例标准化的消融实验

我们称原始的 pix2pixHD 的模型为 M_1 ，在 pix2pixHD 的 G_1 基础上去掉前 5 层实例标准化的模型为 M_2 。用手绘的草图对我们的模型及以上 2 个模型分别进行测试，得到的结果如图 4.3 所示：

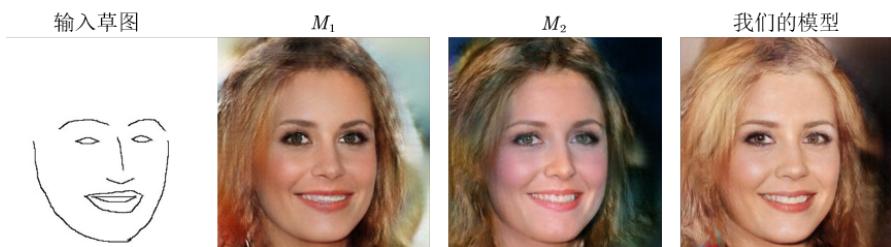


图 4.3 针对实例标准化层数量的对比实验

可以看到，我们模型的生成效果最好，图像的对比度更高，光照也更加自然，而 M_2 的生成结果会出现蓝绿等不正常色块，降低了人脸的真实感。这是因为

M_2 把生成器特征提取阶段的实例标准化操作都去掉之后，使最终提取出来的特征很多地方值为 0，造成反向传播过程中梯度消失，训练过程很难收敛。

下面，只比较我们的模型与原模型 pix2pixHD 之间生成结果的差异。

二、我们的模型与 pix2pixHD 的对比实验

首先我们的细节更加逼真，纹理更加丰富。比如牙齿位置，pix2pixHD 的结果产生模糊，而我们的结果则纹理清晰，细节丰富。再比如额头位置，pix2pixHD 的生成结果经常出现一些乱发状的噪声，而我们的结果则没有这个问题，如图 4.4 所示。

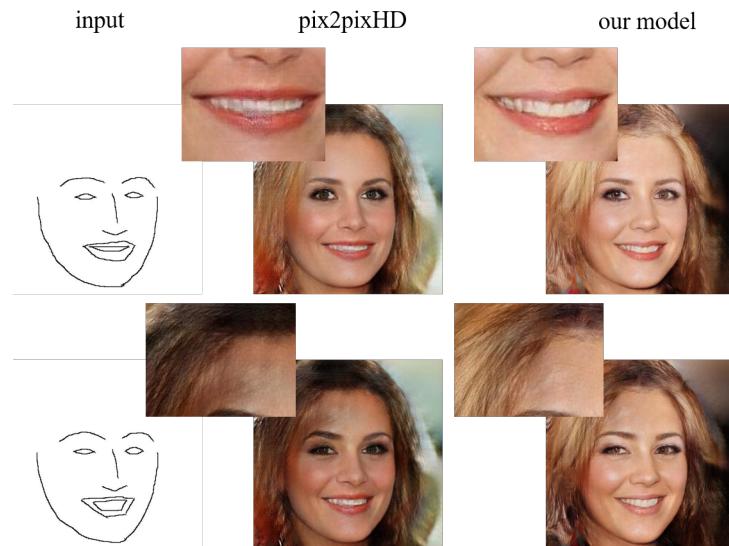


图 4.4 我们的模型与 pix2pixHD 的生成结果细节对比

其次，我们的模型可以实现图像编辑的功能，而 pix2pixHD 则不能完全做到。比如我们在输入草图中只改变嘴巴的形状，我们模型的生成结果很明显地改变了嘴部且与草图形状相吻合，而 pix2pixHD 的生成结果在嘴部的改变并不明显，并且嘴部生成质量较差，如图 4.5 第 1, 2 行所示。

我们的模型还能保证生成结果仅仅在草图编辑位置上产生改变，其余位置保持不变，而 pix2pixHD 则不然。如图 4.6 所示，当我们在草图上编辑鼻子形状的时候，我们的模型生成的结果眼睛保持不变，而 pix2pixHD 生成结果的眼睛朝向发生了明显的变化。如图 4.5 第 3 行所示，当在草图上编辑嘴部形状时，我们模型的生成结果除嘴部以外没有发生改变，而 pix2pixHD 的生成结果则全局都变差，特别是眼部的生成质量明显下降。

综上所述，我们改进后的模型较之于原始的 pix2pixHD 模型，针对由手绘草

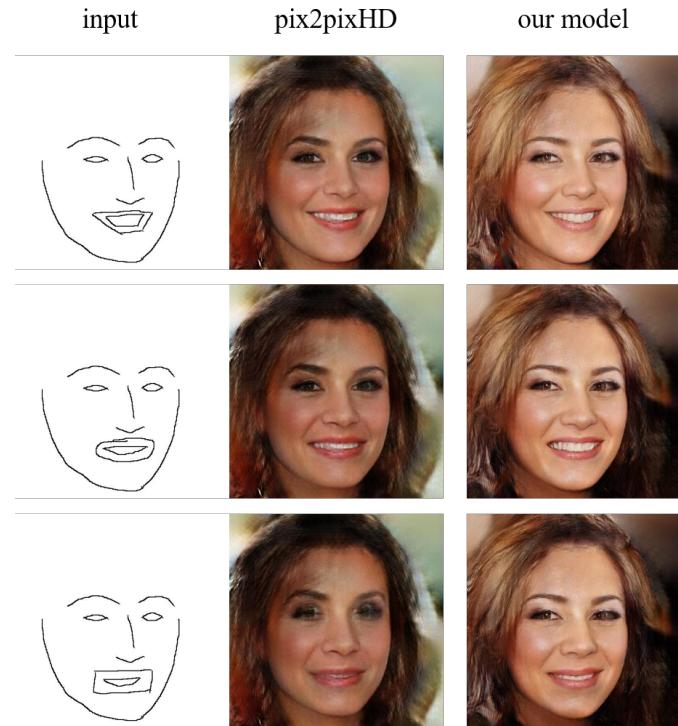


图 4.5 嘴部编辑的生成结果对比

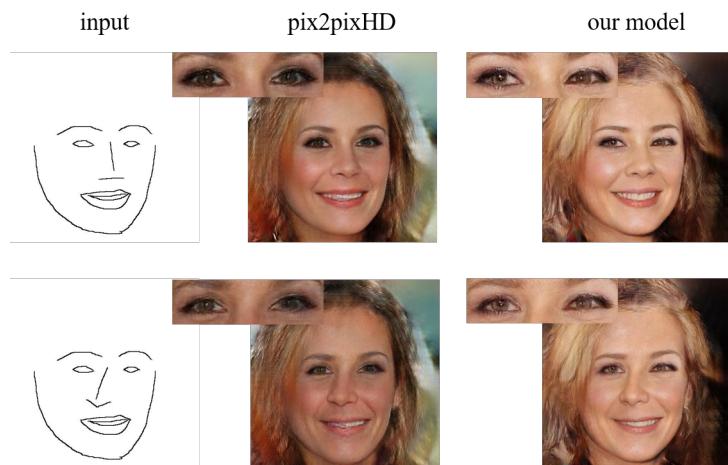


图 4.6 鼻子编辑的生成结果对比

图生成真实人脸的任务，其生成图像质量更高，细节更完备，并且能更好地实现图像编辑的功能。

第五章 总结和展望

本文研究了目前最成功的图像翻译模型 pix2pixHD^[11]，并且在我们的 CelebA-Contour 数据集上训练了该模型。之后用手绘的草图对训练好的模型进行了测试，从测试结果发现 pix2pixHD 存在以下问题：

1. 生成结果细节不够完美，前额、牙齿等部位经常出现模糊现象；
2. 在草图上改变某一部位常造成生成结果全局性的改变，不能实现图像编辑的功能。

所以，我们通过可视化草图的特征，理论分析了造成以上两点问题的原因，发现根源在于每一层的实例标准化操作。我们遂改进了 pix2pixHD 的模型，去掉了网络前两层的实例标准化，重新训练模型，并用手绘草图进行测试，发现生成效果明显好于原模型，同时能很好地实现了预期的图像编辑功能。

我们还探索了去掉前面 n 层（如前 5 层）实例标准化操作的模型，发现如果去掉的太多，则会造成训练过程很难收敛，最终生成结果反而更差。

在测试过程中我们还发现，如果输入草图偏离参考位置会使生成结果变差。在综合分析原因后我们将数据进行了增广，在训练过程中对输入草图加入了一定程度的平移和旋转。最终的模型对输入草图空间位置变化的鲁棒性大大提高。

但是我们的模型也存在很多问题，比如对于精确度不高、带有几何形变的草图，其生成结果会完全忠实于输入从而使真实感降低，如图 5.1 所示。但是这与图像编辑功能是相矛盾的。实现图像编辑意味着模型需要对输入“敏感”，草图上的细微改变也能影响生成的结果；而对精度很差的输入也要生成真实感强的照片则需要模型降低对输入的依赖，变得“迟钝”，才能具有纠偏的能力。

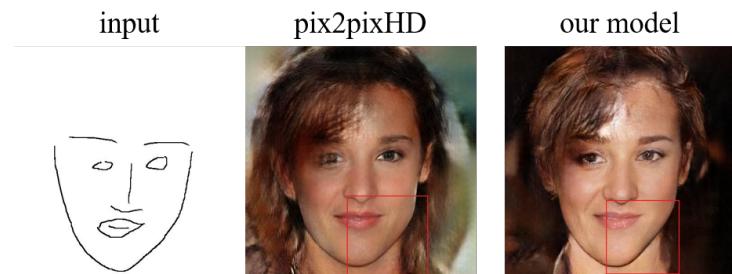


图 5.1 带有变形输入的生成结果对比

所以我们未来要设计一种新的空间自适应的标准化方法，使得以上两点要求能同时满足。对于眼睛、眉毛、鼻子等力求结构精细的部位，我们希望生成结

果尽可能贴近输入草图的形状；而对于脸型属性，由于其本身形状比较规则且左右对称，而人们手绘的草图脸型往往很难画好，所以我们希望模型能减少对脸型轮廓信息的依赖程度，增强生成照片的真实感。

我们的模型对头发的生成做的也不是很好，这是由于训练集轮廓图缺少头发信息造成的。其他类型的草图，比如边缘图或掩膜边界图等，都包含有头发信息，但是它们都比较复杂，缺少绘画基础的普通人很难画出类似的手绘草图。所以我们期望找到一种更好地草图生成方法，在轮廓图的基础上加入头发的样式信息，同时保证可以手绘出风格相近的草图。

希望经过最终的努力，能逐步解决上述问题，不断完善现有的模型，同时完善理论层面的论证，使得模型的生成质量越来越高，鲁棒性越来越强，图像编辑功能越来越强大。

参 考 文 献

- [1] Yin Xi, Yu Xiang, Sohn Kihyuk, et al. Towards large-pose face frontalization in the wild[C]//*The IEEE International Conference on Computer Vision (ICCV)*. 2017.
- [2] Li Yijun, Liu Sifei, Yang Jimei, et al. Generative face completion[C]//*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
- [3] Antipov G, Baccouche M, Dugelay J. Face aging with conditional generative adversarial networks[C]//*2017 IEEE International Conference on Image Processing (ICIP)*. 2017: 2089-2093.
- [4] Zhang F, Zhang T, Mao Q, et al. Geometry guided pose-invariant facial expression recognition[J]. *IEEE Transactions on Image Processing*, 2020, 29:4445-4460.
- [5] He Z, Zuo W, Kan M, et al. Attgan: Facial attribute editing by only changing what you want[J]. *IEEE Transactions on Image Processing*, 2019, 28(11):5464-5478.
- [6] Cao Kaidi, Liao Jing, Yuan Lu. Carigans: Unpaired photo-to-caricature translation [J]. *CoRR*, 2018, abs/1811.00222.
- [7] Tingting Li, Ruihe Qian, Chao Dong, et al. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network[C]//*Multimedia Conference on Multimedia Conference(MM)*. 2018: 645-653.
- [8] Isola Phillip, Zhu Jun-Yan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks[C]//*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
- [9] Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, et al. Generative adversarial nets[M]//*Advances in Neural Information Processing Systems(NIPS)*. 2014: 2672-2680.
- [10] Mirza M, Osindero S. Conditional generative adversarial nets[J]. 2014.
- [11] Wang Ting-Chun, Liu Ming-Yu, Zhu Jun-Yan, et al. High-resolution image synthesis and semantic manipulation with conditional gans[C]//*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [12] Chen Qifeng, Koltun Vladlen. Photographic image synthesis with cascaded refinement networks[C]//*The IEEE International Conference on Computer Vision*

- (ICCV). 2017.
- [13] Ulyanov Dmitry, Vedaldi Andrea, Lempitsky Victor. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis [C]//*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
- [14] Tero Karras, Timo Aila, Samuli Laine, et al. Progressive growing of gans for improved quality, stability, and variation[J]. *CoRR*, 2017, *abs/1710.10196*.