中国科学技术大学

博士学位论文

基于生成对抗网络的图像合成

作者姓名：　鲍建敏

学科专业：　信息与通信工程

导师姓名：　李厚强 教授　罗杰波 教授

完成时间：　二〇一九年六月三日

University of Science and Technology of China

A dissertation for doctor's degree

Image Synthesis based on Generative Adversarial Networks

Author: Jianmin Bao

Speciality: Information and Communication Engineering

Supervisors: Prof. Houqiang Li, Prof. Jiebo Luo

Finished time: June 3rd, 2019

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____ 签字日期：_____

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

☑ 公开 ☐ 保密（____ 年）

作者签名：_____ 导师签名：_____

签字日期：_____ 签字日期：_____

摘　要

图像合成是计算机视觉、计算机图形学等领域的重要研究方向，具有着广泛的应用：由一段文字生成图像、图像在不同模态间的转换、图像的修复、编辑、去模糊、超分辨率等。尽管经过几十年的研究，在面对复杂的自然图像时，图像合成模型的表现依然不尽如人意。合成图像面临的主要挑战是真实性、多样性和与输入条件一致性。近些年生成对抗网络的出现提升了合成图像的真实性。但是由于生成对抗网络自身存在的训练不稳定、收敛状态无法判断、模式坍缩等问题，图像合成中的挑战依然存在。

论文的核心贡献在于针对图像合成中的挑战和生成对抗网络的问题提出了一些解决方案。

论文提出了特征匹配损失函数，以解决生成对抗网络中训练不稳定的问题。在训练中，对于判别网络，论文使用了和原始生成对抗网络中一样的二元交叉熵损失函数，使其保持判别能力。而对于生成网络，论文使用了特征匹配损失函数，该损失函数要求生成图像和真实图像在判别网络中的特征中心靠近，这样能够解决生成对抗网络原始损失函数中的梯度消失的问题，也就使得生成对抗网络的训练更加稳定。实验结果表明，该损失函数有效解决了生成对抗网络训练中的不稳定问题，并且提升了生成模型合成图片的质量。

论文提出将编码网络加入生成对抗网络的框架，以解决生成对抗网络中的模式坍缩问题。编码网络将图片空间映射到隐空间，再使用生成网络将隐空间映射回图片空间，因为原图片空间的分布中的图片是多样的，所以生成网络生成的图片也是多样的。这样就解决了生成对抗网络中的模式坍缩问题。实验结果表明，加入了编码网络的生成对抗网络框架生成了更加富有多样性的图片，从而证明该框架有效解决了模式坍缩问题。同时该框架可以完成很多应用：细粒度图片合成、图片修复、图片渐变、图片属性检索、数据增强等。

论文提出了身份保持的生成对抗网络框架，以实现指定身份和属性的人脸图片合成。该框架可以解耦人脸图片中的身份特征和属性特征（角度、表情、光照等），然后重组该身份特征和从另外一张人脸图片提取的属性特征得到一张新的人脸图片，该人脸图片满足给定的身份特征，同时也满足给定的属性特征。实验结果表明，该框架实现了开放集中的身份保持的人脸图片合成。同时该框架可以应用在很多任务中：侧脸图片转正脸图片、人脸识别中的对抗样本检测、人脸图片属性转换等。

关键词：图像合成；生成对抗网络；人脸合成；身份保持；特征匹配

ABSTRACT

Image synthesis is an important research direction in the fields of computer vision and computer graphics. It has a wide rang of applications: image generation from a piece of text, image translation between different domains, image restoration, editing, deblurring, super-resolution, etc. Despite decades of research, the performance of image synthesis models is still not satisfactory in the face of complex natural images. The main challenges in synthesizing images are realism, diversity, and consistency with input conditions. The emergence of generative adversarial networks(GANs) in recent years has increased the realism of generated images. However, the challenges in image synthesis still exist due to problems in GANs such as: unstable training, inability to judge the convergence state, and mode collapse.

The core contributions of this thesis are to propose some solutions to the challenges in image synthesis and problems in GANs.

This thesis proposes a feature matching loss function to solve the problem of unstable training in GANs. In the training stage, for the discriminator network, we use the same binary cross entropy loss function as in the original GANs to maintain discriminative ability. For the generator network, we use the loss function of feature matching, which requires the feature center of generated image to be close to the feature center of real image, thus solving the problem of gradient vanishing in the original loss function of GANs. The training of GANs is more stable. The experimental results show that the loss function effectively solves the instability problem in the GANs, and improves the quality of the generated images.

This thesis proposes to add the encoder network to the GAN framework to solve the problem of mode collapse in GANs. The encoder network maps the image space to the latent space, and then uses the generator network to map the hidden space back to the image space. Because the images in real image space are diverse, the images generated by the generator network are also diverse. This solves the problem of model collapse in GANs. The experimental results show that the GAN framework with encoder generates more diverse images, which proves that the framework effectively solves the problem of mode collapse. At the same time, the framework can be applied to many applications: fine-grained image synthesis, image inpainting, image morphing, image attribute retrieval, data enhancement, and so on.

This thesis proposes identity preserving GANs framework to solve the problem of

open-set identity preserving face synthesis. The framework can disentangle the identity features and attribute features (poses, expressions, illuminations, etc.) in the face image, and then recombine the identity features and the attribute features extracted from another face image, and input them into the generator model and get a new face image. The face image maintains the given identity feature while also maintains the given attribute feature. The experimental results show that the framework can perform open-set identity preserving face image synthesis. At the same time, the framework can be applied to many tasks: profile face to frontal face, adversarial examples detection in face recognition system, face attributes translation, and so on.

Key Words: Image Synthesis; Generative Adversarial Networks; Face Synthesis; Identity Preserving; Feature Matching

目　录

插 图 清 单

表 格 清 单

算 法 清 单

第 1 章　绪　　论

1.1　图像合成简介

每天，人们会接触到海量的视觉内容，例如在视频网站观看视频、在游戏网站打电子游戏、在社交媒体中分享照片等等。这些海量的图像大大加速了现代计算机对于图像内容的理解。特别是深度学习的出现，计算机对于图像内容的理解取得了突破性的进展，例如图片的分类、检测和分割等任务。但是，图像理解也有与其相反的方向：图像合成。图像合成是将对图像内容的理解转换回图像的过程。例如将一个噪声向量、一句话、一个标签或者一张语义标注图片转换回图片。如下图1.1所示为图像合成的示例，计算机需要根据左侧人提供的信息来合成满足条件的图片。

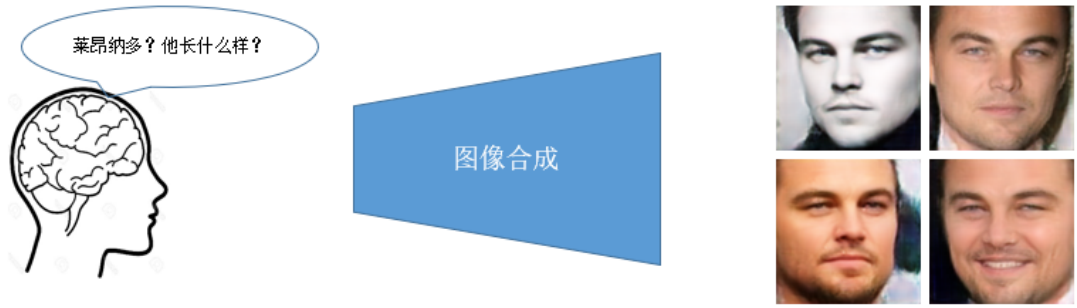


图 1.1　图像合成的示意图。

一般来说，图像合成分为两种形式：一种是无条件输入的生成，即从一个特定的隐空间分布出发合成图片。例如从一个 100 维空间的高斯分布经过模型合成人脸图片。另外一种是有条件输入的生成，给定一个或者多个条件，让合成的图片满足给定的条件，例如给定一张侧脸的图片经过模型合成正脸的图片。比较两种不同的方式可以知道，无条件输入的生成需要合成模型具有很高的数据捕捉能力，能够直接利用数据学习到模型的参数并能合成数据分布中的图片。而有条件输入的则需要模型能够理解并使用条件信息合成满足条件的图片。这两种图像合成的方式现在都在被广泛研究。

图像合成是计算机视觉和计算机图形学中的一个基本问题。它是现在计算机中广泛应用的图片编辑，平面设计，计算机艺术，电影特效等技术的基础。

1.2　研究意义

图像合成作为计算机辅助人类处理视觉和图形学问题的方法。人类利用图像合成可以完成很多的应用，例如，对于给定一个标签生成相关种类的图片；对

图 1.2　图像合成的应用：素描图片到真实图片的转换。[①]

图 1.3　图像合成的应用：文字到图片的生成。[②]

于给定一段话生成相关的图像；完成图片风格的转换，将一张夏天拍摄的图片转换为冬天拍的；完成图片的修复，将一张破损图片修复好，等等。而图像合成需要计算机理解图像，知道图像的具体含义，这样就可以对输入的信息进行整合和处理然后完成图像合成这个任务。

从研究的角度来看，图像合成是当前计算机视觉和计算机图形学中的研究热点。首先图像合成可以帮助减少人们获取图像的代价，例如现在电影工业当中很多的抠图，换景，视频剪辑的使用帮助电影工作者门的节约时间与花费。其次，在图像合成中有很多难题需要解决：首要要解决的问题是合成图像的真实性的问题，只有解决了合成图像的真实性问题，合成的图像被应用在各种应用中才会成为可能。另外研究者也关注条件合成模型中如何使得合成图像满足给定的条件。例如满足在合成人脸图片过程中满足给定的身份信息。这些方面的研究在图像合成中具有重要意义。

从应用角度来看，图像合成可以帮助人类和计算机之间进行更好的交互，帮助人类自动的完成很多应用。以下给出一些相关例子：如果要将一张素描图像转

[①]本图片引用自论文 [1]。

[②]本图片引用自论文 [2]。

图 1.4　图像合成的应用：人脸属性的编辑。

换为真实的彩色图像，一个画家可能要花上几个小时才能做好，但是如图1.2所示，合成模型可以批量的完成从素描图片到真实图片的转换。图像通常比文字具有更好的表达能力，为了更好的表达，如何从一段文字中生成图像呢？合成模型在图1.3中展示了从一段文字到图片的生成。自拍现在是大家生活中常见的场景，如果想编辑自拍人脸图片中的一些属性应该怎么办呢？图1.4展示了合成模型可以做人脸图片属性的编辑。可以看出图像合成模型在计算机视觉和计算机图形学中有着重要的应用。

综上所述，图像合成有着非常多的重要的应用，深深影响着计算机视觉和计算机图形学的发展。随着图像合成的技术的发展，创新的应用越来越多，新的技术也将值得期待。

1.3　主要难点

图像合成面临的最大的一个挑战来自于合成图像的真实性不足。由于在现实生活中，人类被真实图像环绕，所以当一张合成图片出现在眼前的时候，人类往往感觉这张图片不真实。在实际中，人类会对几种常见的人工合成痕迹敏感。如图1.5所示，人们对有着显著的边界的图像 [3]，模糊不清的图像 [4]，缺失结构信息的图像，缺失细节的图像 [5] 等等敏感。如果合成图片中有了这些特征，人们常常会认为这些都不是真实图片，所以图片合成的首要任务就是如何保证模型合成图片的真实性。

在有条件图像合成中，如何让合成图像满足给定的条件输入也是一个非常大的挑战，如图1.6所示侧脸图片转换为正脸图片的任务中。左侧为输入侧脸图片，中间为正确的输出正脸图片，右侧为错误的输出正脸图片。正确输出的正脸

图 1.5　合成图片的真实性面临的挑战。[①]

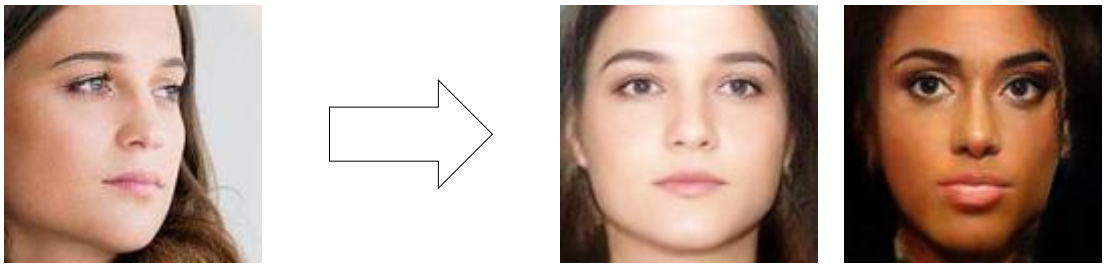


图 1.6　侧脸转正脸的过程中应该保持身份等一致。

的图片应保持身份特征、表情、背景等信息和输入侧脸图片一致。所以如何设计框架，以使得合成模型充分使用输入的条件信息也是图像合成任务的研究方向之一。

图像合成的另外一个挑战来源于合成的图片的多样性有限，如图1.7所示用轮廓图片作为输入合成手提包图片的任务中，一个轮廓图片可以对应着合成多种颜色的手提包。所以当在使用计算机完成这个任务的时候，人们希望计算机可以生成多样的结果。所以怎样将多样性引入到合成模型中也是图像合成工作中的一个研究热点。

图像合成质量的评价标准是另外一个挑战，现在需要衡量的标准一般有：真实性，多样性，与输入条件一致性。这三个标准的定量衡量方法都难以定义，已有的衡量方法一般有两种，一种是人为定义的数值衡量标准，一种是基于用户调研的方式进行质量评价。在人为定义的数值衡量标准中，有峰值信噪比（peak signal-to-noise ratio, PSNR）、结构相似性指标（structural similarity index, SSIM）、

[①]左上图片和右下图片分别引用自论文 [3] 和论文 [5]。

图 1.7　合成图片过程中一个输入往往对应着多个可能的输出。

Inception Score 和 Fréchet Inception Distance(FID) 等衡量标准，这些标准可以在一定程度上反应合成图片的质量，但是评价的方面往往比较单一。另一种使用用户调研的方式进行质量评价，虽然比人为定义的数值衡量标准更加的全面，但是往往需要花费大量的时间。

1.4　研究现状

图像合成的研究从 20 世纪 60 年代开始，经历了从简单线段、规则形状的合成 [6]，到规则图像的合成，如纹理图像、人脸图像的合成 [7-8]，再到复杂自然图像的合成，如数据集 ImageNet[9] 中的图片合成 [10-12]。随着图像数据量逐渐变大，现代计算机的计算能力的提升，自然图像合成的质量不断提升。

图像合成中最关键的是合成模型，下面将从模型的角度分析，图像合成模型经历的三个阶段。

第一阶段是基于图像像素的图像合成模型。其中最具代表性的是 Smith 等人 [13] 和在图像合成中使用像素颜色信息进行前景，背景的线性组合方法。它采集未知像素点周围的前景和背景，然后利用像素的统计信息算出未知像素的具体值。当前景与背景的颜色，纹理或者风格相近时，这种方法可以工作的很好，但是当差异很大时，这种方法常常得不到想要的结果。后来，泊松克隆 [3] 的提出在图像的像素梯度上解决这个问题，泊松过程保证了在边界处前景和背景图是无缝的，从而使得图像合成的效果更好。

第二阶段是基于特征表达的图像合成模型。特征表达是将图像表达为特征的过程，如果将此过程逆过来，则得到了图像合成的方法，其中代表的是主成分分析 (Principal Component Analysis, PCA)[8] 方法，主成分分析可以将人脸表达为特征向量与特征矩阵，反过来，可以利用特征向量便可以合成各种各样的人脸图

片。同样的道理，稀疏表达 [14] 也是图片表达的重要手段，和主成分分析一样的道理，人们可以从稀疏表示中合成图片。

第三阶段是深度卷积神经网络的图像合成模型。得益于深度卷积神经网络的优秀表达能力，图像合成在深度学习时代得到快速的发展。其中最典型的模型如生成对抗网络（Generative Adversarial Network, GAN）[15], 生成对抗网络由一个生成网络 G（Generative Network）和一个判别网络 D（Discriminative Network）构成。生成网络 G 和判别网络 D 在训练的时候用对抗的方式进行学习，生成网络 G 尝试合成图片使得判别网络无法分辨真假，判别网络 D 尝试判别合成的图像是假的，在这样的对抗中生成网络 G 就可以学到怎么合成真实的图像。还有典型的合成模型变分自编码器 (Variationl Auto-Encoder, VAE)[4], 变分自编码器由一个编码器（Encoder）和一个解码器（Decoder）构成，利用变分推断的方法可以学到从一个隐空间（latent space）到真实图片空间的解码器。同时也有基于密度分布建模的 PixelRNN 和 PixelCNN[16] 等方法。

通过这三个阶段的对比，可以看到图像合成模型取得了非常大的发展，尤其是随着深度卷积神经网络的出现，图像合成到了一个新的阶段。但是也应该看到，图像合成中如何保证合成图片的真实性、多样性、与输入条件一致性等问题在现在的深度模型中依然存在。虽然生成对抗网络的出现极大的解决了合成图片真实性的问题，但是由于生成对抗网络网络自身存在的训练不稳定 [17]、收敛状态无法判断 [18]、模式坍缩（mode collapse）[17] 等问题，关于如何改进生成对抗网络的研究工作层出不穷。

为了解决生成对抗网络中出现的训练不稳定的问题。论文在第三章中提出了改进生成对抗网络损失函数的方法，论文提出了特征匹配（Feature Matching）的损失函数。在训练中，对与判别网络 D，论文使用了和原始生成对抗网络中一样的二元交叉熵损失函数，使其保持判别能力。而对于生成网络 G，论文使用了特征匹配（Feature Matching）的损失函数，该损失函数要求生成图像和真实图像在判别网络中的特征中心靠近，这样解决了生成对抗网络原始损失函数中的梯度消失的问题，也就使得生成对抗网络的训练更加稳定。同时由于特征优秀的表达能力，这会使得生成对抗网络收敛的更快。

对于一个已经训练好的的生成对抗网络，如果使用不同的隐变量进行测试，会发现生成模型在不同的隐变量上合成的图片出现内容相同的现象。该现象被称为声称对抗网络中的模式坍缩现象。第四章中，论文提出将编码网络加入生成对抗网络的训练中以解决模式坍缩的问题。论文利用编码网络将图片空间隐射到隐空间，再使用生成网络将隐空间隐射回图片空间，因为原图片空间的分布中的图片是多样的，所以生成网络生成的图片也是多样的。这样解决了生成对抗网络中的模式坍缩问题。

现有的基于身份标签的人脸图片合成模型常常能处理的只是已有身份标签的人脸图片合成。在第五章中，为了解决这个限制，论文提出了一个用于解决开放人脸数据集的人脸图片合成框架。该框架可以从一张人脸图片中解耦身份特征和属性特征，然后再结合身份特征和属性特征合成新的人脸图片。这样对于一个不在训练数据中的身份的人脸图片，可以使用论文提出的框架提取其身份特征，然后任意选择图片作为属性特征进行开放数据集的人脸图片合成。

1.5　研究趋势

基于以上分析，图像合成现在正在逐渐被用在越来越多的应用当中，比如灵活的人脸编辑，有了这样的人脸编辑模型，人们可以指定给脸部做怎样的编辑，修改嘴巴的大小，鼻子的形状，肤色等等；游戏画面低清转为高清，模型将低清游戏画面转化为高清游戏画面，让游戏者有更好的游戏体验；给黑白照片上色，让黑白图片重新焕发光彩。

同时图像合成正在逐渐向如何合成高分辨率高质量的图片发展。高分辨率图像合成可以帮助研究者们得到高质量的图片，更大的提升图像的视觉效果。有了高分辨率图像的合成，使得合成图像应用在现在流行的电影，电视，海报中成为可能。

同时在各种移动设备上，如何将图像合成技术用在流行的自拍、短视频、图片编辑中也是一个研究热点。另一方面，提升图像合成的效率也是将其用在移动端的一个关键点，如何设计精确高效的图像合成算法也是一个被学术和工业界所关注的方向。

1.6　主要创新点

基于章节1.3的讨论，图像合成有着很多的挑战，本文的工作重要在于设计图像合成网络可以完成真实的，满足给定条件的图像合成。主要创新点如下：

1. 提出了特征匹配损失函数，该损失函数改善了生成对抗网络的训练稳定性。在训练中，对与判别网络，论文使用了和原始生成对抗网络中一样的二元交叉熵损失函数，使其保持判别能力。而对于生成网络，论文使用了特征匹配（Feature Matching）的损失函数，该损失函数要求生成图像和真实图像在判别网络中的特征中心靠近，这样解决了生成对抗网络原始损失函数中的梯度消失的问题，也就使得生成对抗网络的训练更加稳定。同时，该损失函数可以用在条件生成框架中，帮助合成模型在条件生成中合成更加符合条件的图片。实验结果表明，该损失函数使生成对抗网络的训练更加

稳定，提升了生成模型的合成图片的质量，帮助生成模型合成更加符合条件的图片。

2. 提出了一个新的条件变分生成对抗网络（CVAE-GAN）的训练框架，该框架将编码网络加入生成对抗网络的训练中以解决模式坍缩的问题。论文利用编码网络将图片空间隐射到隐空间，再使用生成网络将隐空间隐射回图片空间，因为原图片空间的分布中的图片是多样的，所以生成网络生成的图片也是多样的。这样解决了生成对抗网络中的模式坍缩问题。实验结果表明，该框架改善了生成对抗网络的模式坍缩问题，同时可以完成很多应用：图片的修复、图片的渐变、相同属性的图片的检索、数据增强等。
3. 提出了面向在开放数据集中的人脸合成的身份保持的生成对抗网络框架，该框架可以解耦人脸图片中的身份特征和属性特征（角度，表情，光照等等），然后重组该身份特征和从另外一张人脸图片提取的属性信息，将其输入进生成模型得到一张新的人脸图片。该人脸图片满足给定的身份特征，同时满足给定的属性特征。实验证明，该框架可以完成开放数据集中的人脸图片合成，同时可以完成很多应用：人脸属性转换、侧脸图片转换为正脸图片、人脸识别中对抗样本的检测等。

1.7　章节安排

本文主要是关注于图像合成框架设计，以合成真实的，多样的，满足给定条件的图像。具体章节安排如下：

第 2 章介绍相关工作，从图像合成的基本框架、生成对抗网络的改进工作、图像合成的应用与评价标准三个方面介绍图像合成历史上的著名方法以及本文的相关工作。

第 3 章提出特征匹配条件生成对抗网络（FM-CGAN）图像合成框架，在该框架中，论文提出特征匹配损失函数以改进生成对抗网络的训练稳定性。该损失函数改善了原始损失函数中的梯度消失的问题，使得生成对抗网络的训练更加稳定。同时该损失函数也被用在条件合成模型中以改善合成图片与输入条件的一致性。

第 4 章提出条件变分生成对抗网络（CVAE-GAN）图像合成框架。该框架将编码网络加入生成对抗网络的训练中以解决生成对抗网络中的模式坍缩的问题。有了编码网络的存在，合成模型可以合成更加真实与多样的图片。同时论文展示了条件变分生成对抗网络有着广泛应用，如图像修复、图像渐变、数据增强等等。

第 5 章提出了面向在开放数据集中的身份保持的生成对抗网络（IP-GAN）人

脸图片合成框架，该框架设计了一个从人脸图片中解耦身份特征和属性特征的方法。然后针对开放数据集中的人脸图片，框架先提取其身份特征，然后和其他属性特征组合合成新的人脸图片。同时论文在实验中展示了身份保持的生成对抗网络可以被应用在很多实际任务中，如：人脸属性转换、侧脸转正脸、人脸识别系统中的对抗样本检测等等。

第 6 章是回顾本文内容，总结内容与创新点，并且发现其中的不足点与未来工作。

第 2 章　相 关 工 作

如第 1 章所介绍，图像合成技术最核心的的部分是图像合成模型的选择。不论是无条件输入的图像合成还是有条件输入的图像合成。一个好的合成模型的使用可以使图像合成的结果更加真实，多样和满足输入条件。目前为止，图像合成技术已经有了 50 多年的历史。图像合成技术的发展本质上就是图像合成模型的发展。在图像合成的早期，人们只能进行简单线段、规则形状的合成。后来随着特征表达技术的发展，如主成分分析（PCA）[8]、独立成分分析（Indepedent Component Analysis, ICA）[19] 的出现使得图像合成可以完成规则纹理、结构简单的图像合成。最近深度卷积神经网络的发展催生了很多深度图像合成模型如变分自编码器（VAE）[4]、生成对抗网络（GAN）[15]、自回归模型（autoregression）[20] 等等。这些深度合成模型使得图像合成可以完成复杂自然图像的合成。因此在2.1章节中论文具体介绍图像合成模型的发展。

在深度图像合成模型中，生成对抗网络模型的出现大大促进了图像合成的发展。生成对抗网络通过对抗学习的方式可以直接学到图像的复杂分布。其使用的对抗损失函数使得生成的图像更加真实，但是其自出现开始就一直存在着训练不稳定 [17]、收敛状态无法判断 [18]、模式坍缩（mode collapse）[17] 等问题，所以关于其改进的工作层出不穷。因为本文的工作也都基于生成对抗网络，所以在2.2章节中论文将介绍基于生成对抗网路的改进工作。

如第 1 章中所介绍，图像合成发展另外一个重要的原因是图像合成具有着广泛的应用。所以在2.3章节中本文将介绍图像合成在一系列问题中的应用，注重介绍其在文字到图片转换，图片到图片转换，图片的修复、编辑、去模糊、超分辨率等中的应用。同时，论文在2.3章节中也将介绍与图像合成评价标准相关的一些工作。

2.1　图像合成模型

建立一个高效的生成自然图像的模型一直是计算机视觉和计算机图形学中的一个重要问题。它旨在从真实的图片分布中学习到模型的参数。因此，一个好的生成模型能够表达出真实的数据分布。在图像合成研究的早期，传统图像合成模型利用很多典型的特征表达方法进行图像合成。但是受限于特征表达能力的限制，模型只能进行简单图像的合成。深度卷积神经网络的出现大大改进特征的表达能力，使得真实图像的复杂分布的特征表达成为可能，所以许多深度图像合成模型应运而生 [4, 15, 17-18, 20-25]。下面将从传统图像合成模型开始介

绍，然后注重介绍几种重要的深度图像合成模型：变分自编码器（VAE）[4]、生成对抗网络（GAN）[15]、PixelCNN 和 PixelRNN[16]、和条件合成模型条件自编码器（Conditional Variational Auto-Encoder, CVAE）[26] 或者条件生成对抗网络（Conditional Generative Adversarial Network, CGAN）[27]）等等，下面将主要介绍这些生成框架。

2.1.1　传统图像合成模型

图像合成技术出现的早期，研究者们只能使用计算机完成简单线段、规则形状的合成。后来有一些的工作 [13] 从图像像素上的一些关系出发约束图像的修复或者融合过程。再后来研究者们发现可以利用图像梯度上的连续性进行图像的融合，所以如泊松克隆 [3] 等经典方法出现来解决图像融合的问题。后来随着特征表达技术的发展如：主成分分析（PCA）[8]、独立成分分析（ICA）[19]、和高斯混合模型（GMM）[28-30] 等等。在这些模型中，图片的分布都是被假设成一个非常简单的分布，所以这些模型只能处理简单纹理，规则结构图片的合成，因而不能处理复杂数据分布的合成。后面的出现的模型，如因马尔科夫模型（HMM）[31]，马尔科夫随机场（MRF）[32]、限制玻尔兹曼机（RBMs）[33-34] 和通过判别模型训练生成网络的方法 [35] 在简单图片分布的合成上取得更多的进步。但是由于它们特征表达能力的限制，所以它们只能做一些简单的图像合成，数字的图像，规则人脸的图像合成。

2.1.2　变分自编码器

深度卷积神经网络的出现，大大提升了图像的特征表达能力，所以其在很大程度上提升了图像合成的质量，在深度合成模型中，其中最经典的工作之一便是变分自编码器（VAE）[4]。变分自编码器是由 Diederik P. Kingma 和 Max Welling 在 2014 年首次提出，一经提出就引起了学术界的广泛关注。变分自编码器由编码网络（Encoder）和解码网络（Decoder）构成，其中编码器负责将图片转换为隐空间（Latent Space）的表达，而解码网络则负责将隐空间的表达再转换到图片空间中。

其基本的框架如图2.1所示。输入图片 X 进入编码网络得到其在隐空间的均值表达 $\mu(X)$ 和方差表达 $\Sigma(X)$，如果直接在该估计的分布 $N(\mu(X), \Sigma(X))$ 上采样，则无法进行梯度的回传，所以采用了重新参数化的技巧（reparameterization trick），用该技巧采样一个 $z = \mu(X) + \Sigma(X) \odot N(0, I)$，$\odot$ 表示每个元素位置上的相乘。然后再将 z 输入进解码网络得到图片 X 的重构 $f(z)$。其所使用的损失函数为：

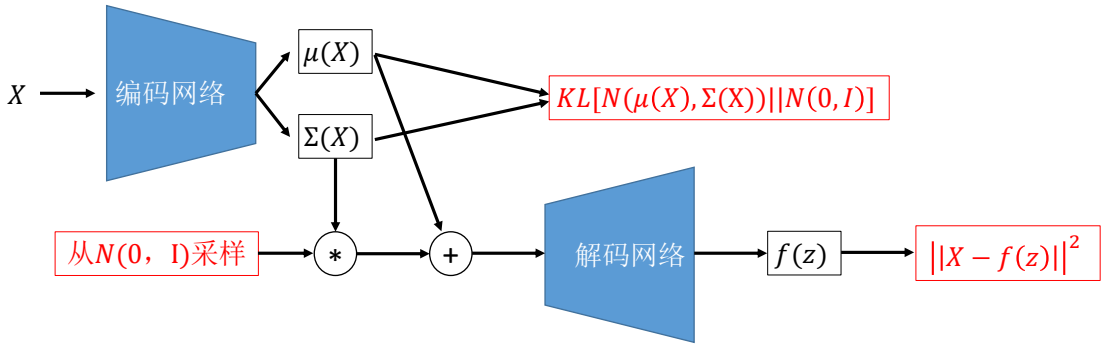


图 2.1　变分自编码器（VAE）的示意图。

$$\mathcal{L}_{VAE} = \mathcal{KL}[N(\mu(X), \Sigma(X))||N(0, I)] + ||f(z) - X||_2^2, \tag{2.1}$$

其中 $z = \mu(X) + \Sigma(X) \odot N(0, I)$，$f(z)$ 为 z 经过解码网络的重构。在整体损失函数中 $\mathcal{KL}[N(\mu(X), \Sigma(X))||N(0, I)]$ 用于将图片经过编码网络的隐空间分布向 $N(0, I)$ 靠近，而 $||f(z) - X||_2^2$ 则让重构的图片尽可能向真实图片靠近。

变分自编码器在测试阶段只需要解码器，直接将一个从 $N(0, I)$ 分布中采样的点 z 输入进解码器便可以进行相关图片的合成。后续的工作 DRAW[36] 基于变分自编码器提出使用注意力的机制将图像的合成过程转换为一个部分接着一个部分生成的序列化操作。

2.1.3　生成对抗网络

在深度合成模型中，其中另外一个经典的工作便是生成对抗网络（GAN）[15]。Ian Goodfellow 在 2014 年首次提出生成对抗网络，一经提出引起了学术界的广泛关注，生成对抗网络的基本原理其实非常简单，这里以生成图片为例进行说明。如图2.2所示，假设有两个网络，生成网络 G（Generator）和判别网络 D（Discriminator）。

生成对抗网络是由两部分组成：生成模型 G（generative model）和判别模型 D（discriminative model）。生成模型是用来学习到真实数据的分布。判别模型是一个二分类器，用来判别输入是真实数据还是生成数据。x 是真实数据，符合 $P_r(x)$ 分布。z 是隐空间变量，符合 $P_z(z)$ 分布，比如高斯分布或者均匀分布。然后从假设隐空间 z 进行抽样，通过生成模型 G 之后生成数据 $x' = G(z)$。然后真实数据与生成数据一起送入判别模型 D，输出判定类别。在原始的生成对抗网络框架中，判别模型需要进行一个二分类的判别，所以此时最基本的想法就是使用二元的交叉熵损失函数的方法。对于真实的图片，其给定的标签为 1，对于生成的图片其给定的标签为 0。而生成模型 G 尝试合成图片使判别模型 D 判别为真，

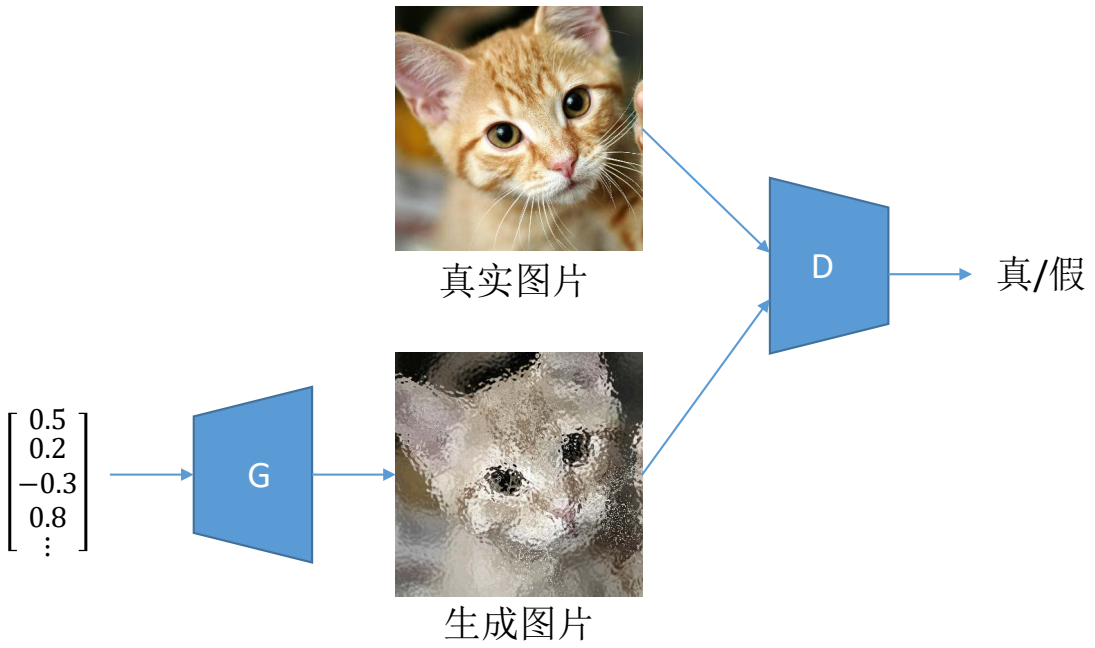


图 2.2　生成对抗网络（GAN）的原理示意图。

这样可以得到其损失函数为：

$$\min_D \max_G V(D,G) = -\mathbb{E}_{x\sim P_r}[\log D(x)] - \mathbb{E}_{z\sim P_z}[\log(1 - D(G(z))], \tag{2.2}$$

其中 P_r 是真实数据分布，P_z 是假设的隐空间的分布。

如果用一个简单模型比喻的话，其训练过程会如图2.3中所示的过程：

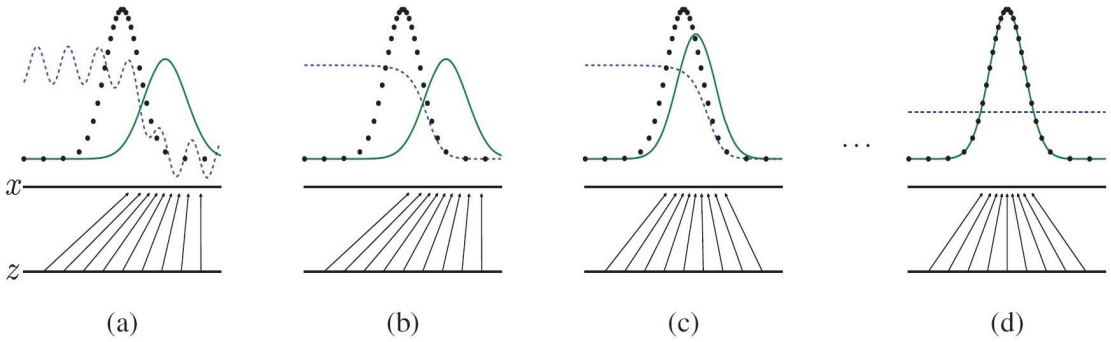


图 2.3　生成对抗网络的迭代优化过程。[①]

在图2.3中，z 表示隐变量空间，x 数据空间，从 z 到 x 的箭头表示 GAN 学到的生成网络映射 $x = G(z)$。黑色的虚线表示真实数据的分布，绿色的实线表示通过生成网络 G 生成的数据分布，蓝色的虚线表示判别网络 D 的判别函数，对于图 (a)，判别网络 D 是一个部分精确的分类器，只能部分区分真实数据和生成数据。当训练过程到图 (b) 中所示，判别网络 D 得到了一些训练，对生成数据和真实数据有比较明显的分类能力，这个时候判别网络 D 会促进生成网络 G 生成

[①]本图片引用自论文 [15]。

自己无法分类的数据。再然后对于图 (c)：绿色实线一步步向黑色虚线偏移，也就是说生成数据分布在想真实数据分布靠近。最后训练的过程收敛示意图如 (d) 中所示，蓝色的虚线为一条水平的线，这时判别网络没有了判别能力。生成网络生成的数据分布和真实数据分布相同，即 $p_g(x) = p_r(x)$。到这里，G 网络和 D 网络就处于纳什均衡状态，也是整个系统达到收敛的状态了。算法2.1展示了 GAN 的训练过程。在训练过程中，先使用真实的数据和生成的数据分布训练判别网络 k 次，然后再训练生成网络 G 一次。在这样的迭代过程中，G 达到收敛的状态。

算法 2.1　生成对抗网络的训练算法。

while 生成网络 G 没有收敛 **do**
  for k 步 **do**
    从 $P_z \sim N(0, I)$ 中采样 m 个样本 $\{z_{(1)}, ..., z_{(m)}\}$。
    从 $P_r(x)$ 中采样 m 个真实数据样本 $\{x_{(1)}, ..., x_{(m)}\}$。
    使用梯度上升法更新判别网络：
    $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [log(D(x_{(i)})) + log(1 - D(G(z_{(i)})))]$。
  end
  从 $P_z \sim N(0, I)$ 中采样 m 个样本 $\{z_{(1)}, ..., z_{(m)}\}$。
  使用梯度下降法更新生成网络：
  $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} log(1 - D(G(z_{(i)})))$。
end

基于上面介绍的变分自编码器和生成对抗网络，后续也有工作将变分自编码器中的解码网络和生成对抗网络中的生成网络合并为一个网络，从而形成一个 VAE/GAN[23] 的结构。同时也有工作将对抗损失函数用在变分自编码器的隐空间中得到 AAE（Adversarial Auto-Encoder）[37] 结构。

2.1.4　自回归模型

自回归模型（autoregression）是由 Hugo Larochelle 和 Iain Murray 在 2011 年提出，它将图像合成过程近似于逐个像素点的合成，同时后面生成的像素点将以前面生成的像素点作为参考。相当于将预测一张图像上所有像素点的联合分布转换为对条件分布的预测。于是可以得到下式：

$$p(x) = \prod_{t=1}^{n} p(x_i | x_1, \cdots, x_{i-1}), \tag{2.3}$$

其中 x_i 就是指在 i 处的像素点。

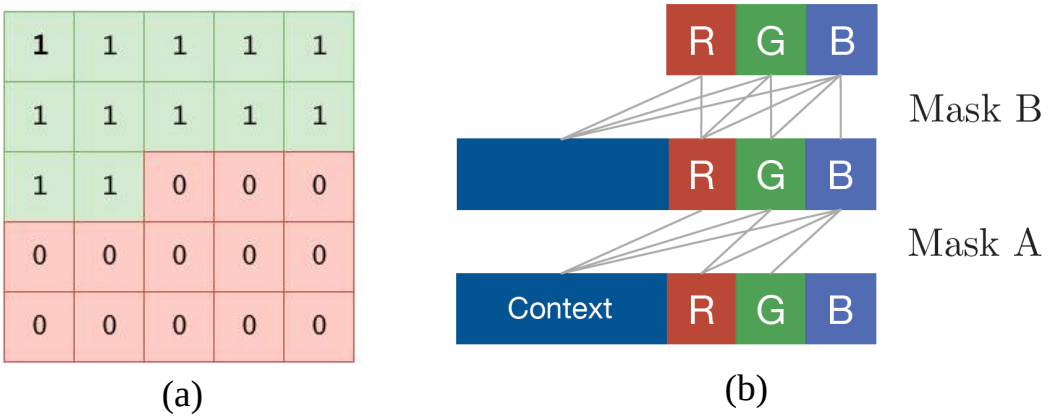


图 2.4　(a) 为一个卷积核为 5×5 的带掩模的卷积层（masked convolution）的示意图。(b) 为卷积层中不同连接示意图，在 Mask A 中，不包含自身到自身的连接，而在 Mask B 中包含自身到自身的连接。[①]

在自回归模型的基础上，Aaron van den Oord 等人在 2016 提出 PixelRNN 和 PixelCNN[16] 图像生成模型。在自回归模型中，由于采用的是有顺序的预测的方法得到像素值的方法，所以 PixelRNN 自然而然就使用递归神经网络（RNN）[38] 的时序思想处理这个预测过程，所以该模型被命名为 PixelRNN。在实现中，他们的网络中使用了多个长短期记忆（LSTM）[39] 层来构建递归神经网络模型。但是由于递归神经网络模型计算复杂度高，需要的训练时间长，所以作者在此基础上又提出了 pixelCNN 模型，pixeCNN 在模型中使用了带掩模的卷积层（masked convolution）来近似递归神经网络的过程。同时在模型中去掉了池化层。如图2.4所示，(a) 为一个卷积核为 5×5 的带掩模的卷积层（masked convolution）的示意图。(b) 为卷积层中不同连接示意图，在 mask A 中，不包含自身到自身的连接，而在 mask B 中包含自身到自身的连接。PixelCNN 得益于在卷积时可以并行的进行计算，所以训练时间较 PixelRNN 有一定的提升。

2.1.5　条件合成模型

为了让生成对抗网络满足有条件的生成，Mehdi Mirza 在 2014 年提出条件生成对抗网络（Conditional GANs）[27]。假设有条件信息 y 作为监督，那么如何将监督信息用在生成网络和判别网络中呢？条件生成对抗网络将条件信息 y 作为输入放入生成网络和判别网络中，其基本示意图如图2.5所示，条件信息和输入信息直接使用连接的方式输入网络。

条件生成对抗网络的损失函数如下式所示：

$$\min_D \max_G V(D,G) = -\mathbb{E}_{x\sim P_r}[\log D(x,y)] - \mathbb{E}_{z\sim P_z}[\log(1 - D(G(z,y),y)], \tag{2.4}$$

其中 z 为隐空间变量，y 为条件信息，$G(z,y)$ 为将条件信息 y 和隐空间变量 z 输

[①]本图片引用自论文 [16]。

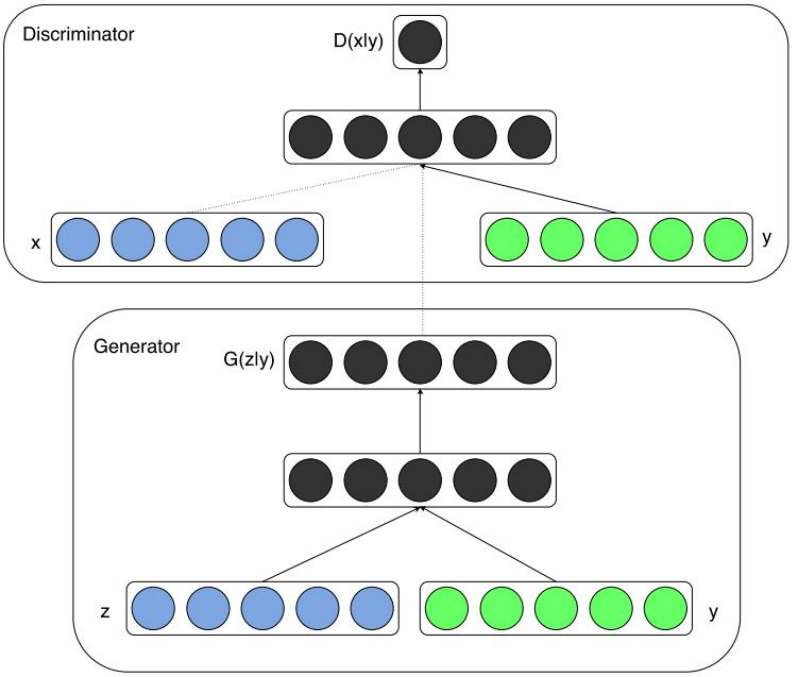


图 2.5　CGAN 基本框架示意图。[①]

入到生成网络 G 得到的生成图片。$D(x, y)$ 表示将条件信息 y 和真实图片 x 合并输入到判别网络 D 中得到的输出，其标签为 1。$D(G(z, y), y)$ 表示将条件信息 y 和生成图片 $G(z, y)$ 合输入到判别网络 D 中得到的输出，其标签为 0。

后来的工作 AC-GANs[40] 改进了条件生成对抗网络中的判别网络，将分类的功能也加入到判别网络中以提升模型的性能。最近的工作 [41] 提出了在判别网络中使用投影分类的方法，进一步增加判别网络的能力，从而使生成网络的性能进一步提升。

为了使变分自编码器满足有条件的生成，Xinchen Yan 等人提出 CVAE[26] 的框架，该框架将条件信息作为编码器和解码器的输入，这样在图像合成的时候通过对输入条件的改变便可以完成有条件的生成。同时 Aaron Van dne Oord 等人提出基于 PixeCNN 的条件生成模型 [42]。

2.2　生成对抗网络的改进

这本章节中，本文将介绍基于生成对抗网络的一些改进共工作。由于生成对抗网络训练中的稳定性存在问题，所以在2.2.1章节中，论文介绍改进损失函数以提高生成对抗网络训练稳定性的方法。在2.2.2章节中, 论文将介绍改进生成网络和判别网络结构的方法提升训练稳定性与合成图片质量的方法。在2.2.3章节中, 论文将介绍改进生成对抗网络的训练方法以提升合成图片质量的方法。

[①]本图片引用自论文 [27]。

2.2.1　损失函数的改进

由于生成对抗网络原始的损失函数容易导致梯度消失的问题，所以后面陆续出来很多工作如：EBGAN[43]、BEGAN[44]、Loss-Sessitive GAN[45]、info-GAN[46]、WGAN[47]、Least Square GAN[48]、WGAN-gp[49]、f-GAN[50]、MA-GAN[51]、iGAN[52]、McGAN[53] 针对生成对抗网络的损失函数进行改进。下面将选取其中典型工作进行介绍。

1. Wassserstein GAN

Wassserstein GAN(WGAN)[47] 第一次从理论上解释了生成对抗网络训练中出现不稳定的原因。当判别网络为最优分类器且生成图片分布和真实图片分布没有重叠区域的时候，判别网络 D 回传给生成网络 G 的梯度为 0。这个时候也就产生了梯度消失的问题。为了解决这个问题，Wasserstein GAN 提出了使用 Wasserstein 距离进行训练的方法，Wasserstein 距离定义如下：

$$W(P_r, P_g) = \inf_{\gamma \sim \Pi(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||], \tag{2.5}$$

其中 $\Pi(P_r, P_g)$ 是真实图片分布 P_r 和生成图片分布 P_g 组合起来的所有可能的联合分布的集合。

与 KL 散度距离和 JS 散度距离相比起来，Wasserstein 距离的优越性在于：当真实图片分布 P_r 和生成图片分布 P_g 没有重叠时，Wasserstein 距离仍然能够反映它们的远近。所以这可以解决在生成对抗网络早期训练时真实图片分布 P_r 和生成图片分布 P_g 无重叠时出现梯度消失的情况。

虽然可以解决生成对抗网络训练中的梯度消失问题，但是由于 Wasserstein 距离定义（公式2.5）中的 $\inf_{\gamma\sim\Pi(P_r,P_g)}$ 没法直接求解，所以 WGAN 用了一个已有的定理把它变换为如下形式：

$$W(P_r, P_g) = \frac{1}{K} \sup_{||f||_L \leqslant K} \mathbb{E}_{x\sim P_r}[f(x)] - \mathbb{E}_{x\sim P_g}[f(x)], \tag{2.6}$$

其中 $||f_w||_L \leqslant K$ 这个限制的意思是要求存在一个常数 $K \geqslant 0$ 使得定义域内的任意两个元素 x_1 和 x_2 都满足

$$|f(x_1) - f(x_2)| \leqslant K|x_1 - x_2|. \tag{2.7}$$

此时称函数 f 满足 Lipschitz 连续的条件，它的 Lipschitz 常数为 K。

为了使得判别网络 D 满足 $||f_w||_L \leqslant K$ 这个限制。WGAN 采取了一个非常简单的方法，即在生成对抗网络的训练过程中，限制判别网络 D 中所有参数 w_i 不超过某个范围 [-c, c]。在 WGAN 实验中，c 取值为 0.01，此时关于输入样本 x 的导数 $\frac{\partial f_w}{\partial x}$ 不会超过某个范围，所以一定存在某个常数 K 使得 f_w 的局部变动幅

度不会超过 K，判别网络 D 的 Lipschitz 连续条件得以满足。具体在 WGAN 的算法实现中，在每次更新完判别网络 D 的参数 w 后，再把 w 限制到 [-0.01, 0.01] 这个范围就可以了。这就得到了 WGAN 中的对于生成网络 G 和判别网络 D 的两个损失函数。对于判别网络 D 的损失函数：

$$\mathcal{L}_{\text{WGAN}}(D) = \mathbb{E}_{x\sim P_g}[D(x)] - \mathbb{E}_{x\sim P_r}[D(x)]. \tag{2.8}$$

对于生成网络 G 的损失函数：

$$\mathcal{L}_{\text{WGAN}}(G) = -\mathbb{E}_{x\sim P_g}[D(x)]. \tag{2.9}$$

2. 最小平方法生成对抗网络（Least Square GAN）

最小平方法生成对抗网络（Least Square GAN）[48] 由 Xudong Mao 等人在 2016 年提出，它提出生成对抗网络训练一个新的损失函数：最小平方（least square）损失的函数。使用最小平方损失函数可以提升生成对抗网络训练的稳定性，同时提升生成模型合成图片的质量，其对于判别网络 D 的损失函数如下：

$$\mathcal{L}_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{x\sim P_r}[(D(x)-1)^2] + \frac{1}{2}\mathbb{E}_{z\sim P_z}[(D(G(z))+1)^2]. \tag{2.10}$$

其对于生成网络 G 的函数如下：

$$\mathcal{L}_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{z\sim P_z}[(D(G(z)))^2]. \tag{2.11}$$

3. 带梯度惩罚项的 WGAN（WGAN-gp）

在上面 WGAN 中的分析可以知道，WGAN 为了使判别网络 D 满足 Lipschitz 条件，WGAN 采取了一个非常简单的方法，即在训练过程中限制判别网络 D 中所有参数 w_i 不超过某个范围 [-0.01, 0.01]。但是实验证明这会导致判别网络 D 中的参数全部集中在 0.01 或者 -0.01 附近，这导致判别网络 D 的判别能力受到影响并最终影响生成网络生成的效果。所以在 WGAN-gp 中，WGAN-gp 不再使用限制判别网络 D 中所有参数 w_i 的方法。WGAN-gp[49] 提出使用一个梯度惩罚的方式使得判别网络 D 中满足 Lipschitz 条件，WGAN-gp 考虑直接限制 WGAN 损失函数传递的梯度，即如下式：

$$\mathcal{L}_{\text{WGAN-gp}}(D) = \mathbb{E}_{x\sim P_r}[D(x)] - \mathbb{E}_{x\sim P_g}[D(x)] - \lambda\mathbb{E}_{\hat{x}\sim P_{\hat{x}}}(||\nabla_{\hat{x}}D(\hat{x})||_2 - 1)^2, \tag{2.12}$$

其中 $\mathbb{E}_{x\sim P_r}[D(x)] - \mathbb{E}_{x\sim P_g}[D(x)]$ 为输入判别函数 D 中的损失函数，$P_r(x)$ 是真实数据分布，$P_g(x)$ 是生成数据的分布。而后面一项 $\mathbb{E}_{\hat{x}\sim P_{\hat{x}}}(||\nabla_{\hat{x}}D(\hat{x})||_2 - 1)^2$ 为梯度惩

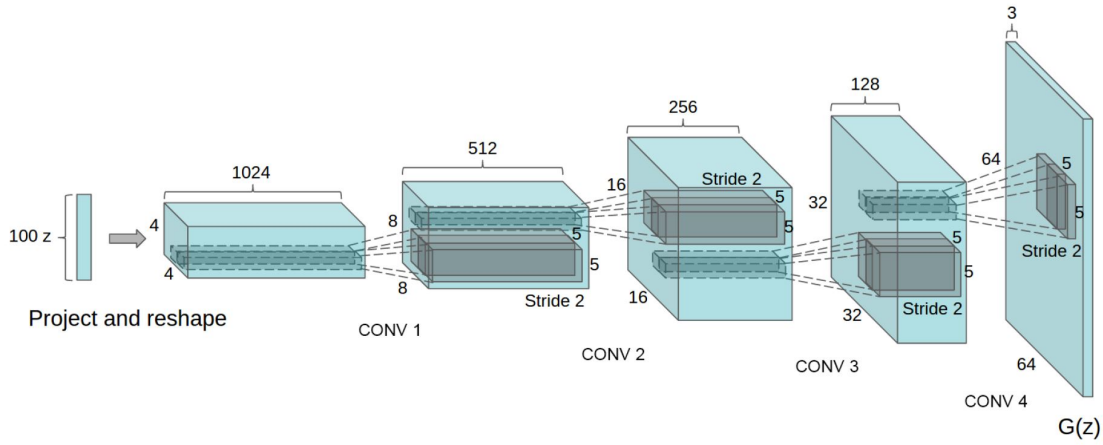


图 2.6　DCGAN 中生成网络示意图。[①]

罚项，$P_r(\hat{x})$ 是对 P_r 和 P_g 之间的空间采样，因为对整个样本空间采样，所需要的样本数量是极大的且难以做到的，所以，WGAN-gp 就提出没必要在整个样本空间上施加 Lipschitz 限制，只要在生成样本空间、真实样本空间以及在它们之间的空间满足即可。有了这样的梯度惩罚项，使得判别网络的能力可以被充分挖掘，同时使得整体的生成对抗网络训练更加稳定，同时使得生成图片的质量得到提升。

2.2.2　模型结构的改进

在本章节中，论文将介绍对于生对抗网络中生成网络和判别网络的结构的改进，由于生成对抗网络训练不稳定，所以很多研究者也尝试使用不同网络结构使得生成对抗网络训练更加稳定，其中比较有代表性的工作有 DCGAN[18]，SNGAN[10]，Self-Attention GAN[11]，Style-Based GAN[54] 等等。

1. 深度卷积生成对抗网络（DCGAN）

在生成对抗网络提出的早期，研究者们没有设计生成网络 G 和判别网络 D 的结构的准则，所以在实验中无法得到一个好的生成模型 G。深度卷积生成对抗网络（DCGAN）[18] 第一次研究生成网络 G 和判别网络 D 的网络结构设计。经过大量的实验，他们发现生成网络和判别网络应该使用下列的建议：

1. 在生成模型和判别模型中不要使用池化（包括最大化池化，平均池化），使用带有步长的卷积来实现上采样或者下采样。
2. 除了生成模型的第一层和判别模型的最后一层，在网络其他地方不要使用全连接层。
3. 在生成模型和判别模型中使用批归一化方法（Batch Normalization）[55]。
4. 在生成模型中的最后一层使用 Tanh 的激活函数，其余的全部使用 ReLU 的

[①]本图片引用自论文 [18]。

激活函数。

5. 在判别模型中的所有层使用 LeakyReLU[56] 的激活函数。

遵照上面的设计准则，DCGAN 设计了生成网络的结构示意图如图2.6所示。

2. 谱归一化生成对抗网络（SNGAN）

在 WGAN 相关工作中，论文详细介绍了让生成网络免于梯度消失的问题的关键是让判别网络满足 Lipschitz 连续的条件，在此假设上，谱归一化生成对抗网络（SNGAN）[10] 提出使用谱归一化方法（Spectrum Normalization）使得判别网络满足 Lipschitz 连续的条件。

SNGAN 的谱归一化操作是对所有层的参数 W 执行以使得其满足 Lipschitz 连续 $\sigma(W) = 1$:

$$\bar{W}_{\mathrm{SN}}(W) := W/\sigma(W). \tag{2.13}$$

如果使用公式 2.13谱归一化每一个 W^l , $\sigma\left(\bar{W}_{\mathrm{SN}}(W)\right) = 1$ 会使得 $\|f\|_{\mathrm{Lip}}$ 的上界为 1。这样论文就可以使得判别函数整体满足 Lipschitz 连续的条件，那么如何求得矩阵 W 的最大奇异值呢，SNGAN 使用的是 Power Iteration [57-58] 的方法，这里不做详述。有了谱归一化层之后，在判别网络设计的时候，将判别网络中的批归一化层（Batch Normalization）层替换为谱归一化层，然后直接训练即可。

3. 自注意力生成对抗网络（SA-GAN）

近些年自注意力机制（Self-Attention）在很多自然语言处理的工作中 [59] 取得非常好的成绩，所以 Han Zhang 等人提出 Self-Attention GAN[11]，将自注意力机制引入到生成对抗网络的生成网络和判别网络中，以提升生成对抗网络的生成图片的质量。

自注意力机制如图2.7所示，对卷积的特征层（feature map）使用两个 1×1 的卷积进行线性变换和通道压缩，然后对两个张量变形（reshape）成矩阵形式，转置之后进行矩阵的相乘，再经过 softmax 操作得到注意力图（attention map）。原特征层再使用 1×1 的卷积进行线性变换（通道数保持不变），然后与注意力图（attention map）矩阵相乘，相加，得到自注意力特征层（self-attention feature maps）。最后，自注意力特征层和原卷积特征进行加权求和（权重参数是可学的），作为最后的输出。

在生成网络和判别网络中了有了自注意力机制，对网络的生成能力和判别能力均有提升。

4. 基于风格生成对抗网络（Style based GAN）

基于风格生成对抗网络（Style-Based GAN） [54] 是由 Teras Kerras 等人在 2018 年首次提出，其在原始 PGGAN[60] 的基础上提出了一个新的高质量图像合成的框架，该框架引入了 AdaIN[54, 61] 作为规范化的方法，并希望使用 AdaIN

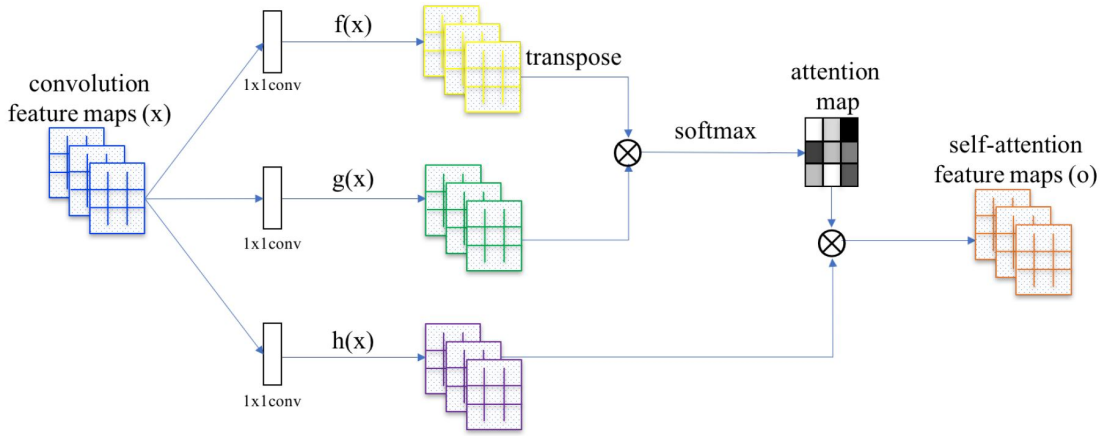


图 2.7　自注意力网络结构示意图，⊗ 表示矩阵的相乘，Softmax 是在每一列上进行操作。[①]

实现对生成图片的进行风格的控制。AdaIN 操作被定义为下式：

$$AdaIN(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}, \tag{2.14}$$

其中，x_i 为 AdaIN 的输入，$\mu(x_i)$ 为输入的均值，$\sigma(x_i)$ 为输入的方差。$y_{s,i}$ 是缩放参数，$y_{b,i}$ 是平移参数，(y_s, y_b) 既为控制合成图像风格的参数。基于风格的生成对抗网络将隐空间变量输入的位置由网络最前端变到了网络的各层 AdaIN 的参数中，基于风格的生成对抗网络学习了一个从隐变量到 (y_s, y_b) 的映射，这样可以实现改变隐变量对合成图像风格的控制。同时引入噪声放在每一个卷积层之后以实现对合成图片的微小改变，比如改变头发的发丝走向，瞳孔位置等等。实验证明这样的结构可以得到更好的图像合成效果，同时，通过改变在不同层的 AdaIN 的输入可以实现在不同粒度上编辑合成的人脸图片。

2.2.3　训练方法的改进

在生成对抗网络出现的早期，很多工作生成的图像的分辨率通常都不高，并且看起来很不真实。而 PGGAN[60] 第一次开创式的提出了由低分辨率到高分辨率逐级提升分辨率的高分辨率图像合成的训练方法，合成图像的分辨率达到了 1024×1024。如果使用原始的生成对抗网络训练方法直接训练从隐空间到 1024×1024 分辨率图片的生成对抗网络，由于高分辨率下真是图像空间分布非常复杂，所以生成网络 G 往往训练不出来。所以如图2.8所示，PGGAN 提出了逐级训练的训练方法，先试着合成低分辨率的图像，然后不断地增加分辨率。在实际中，也就是生成先合成 4×4 分辨率的图像，然后给生成网络 G 添加一次上采样操作，给判别网络 D 添加一次下采样操作，这样生成的分辨率就变为 8×8，以此类推，这样输出的分辨率每次增加两倍最终输出 1024×1024 分辨率的图像。

[①]本图片引用自论文 [11]。

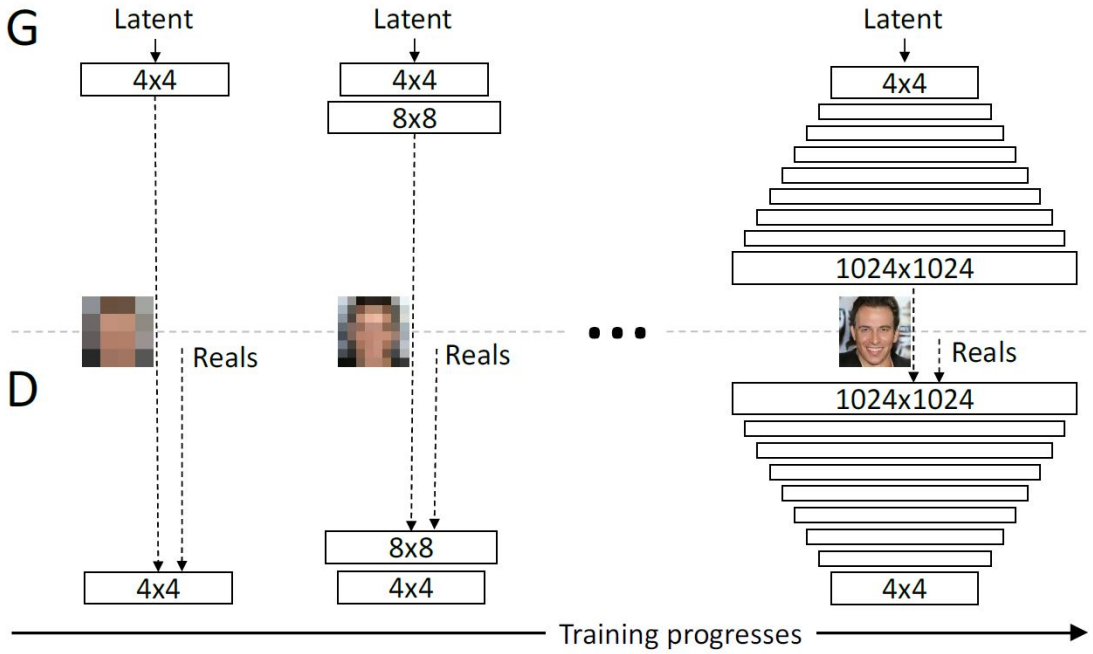


图 2.8　PGGAN 的训练方法，先训练低分辨率的图片，然后同时给生成网络 G 和判别网络 D 加卷积层，这样增加分辨率，直至可以做到 1024×1024 的分辨率。[①]

在 PGGAN 的基础上，后续出现了很多基于 PGGAN 思想的高分辨率图像合成的工作，例如 pix2pixHD[62] 使用这个思想完成从语义标注图到真实街景图的生成，人脸轮廓到真实人脸的生成等一系列图片到图片的图片转换工作。章节4中介绍的基于风格的生成对抗网络（Style Based GAN）[54] 同样食用了基于由低分辨率到高分辨率逐级提升分辨率的高分辨率图像合成的训练方法。

2.3　图像合成的应用与评价标准

图像合成技术可以被应用在各种场景中，在下面前三个章节本文将介绍图像合成几个非常经典的应用：在2.3.1章节中，论文介绍文字语句到图片的转换；在2.3.2章节中，论文介绍图片到图片的转换；在2.3.3章节中，论文介绍将图像合成应用在图片的修复，编辑，去模糊等等任务中。在2.3.4章节中，论文将介绍图像合成的评价标准。

2.3.1　文字到图片的转换

基于一段文字描述生成图像一直是计算机视觉中的一个研究热点，它首先需要将文字描述转换为特征，然后将特征输入进模型进行图像的合成。模型需要准确地抓住文字描述中的关键语义，并且在生成图像时保证生成图片符合自然图像的特点和给定的文字描述。Scott Reed 等人 [63] 第一次提出使用生成对抗

[①]本图片引用自论文 [60]。

网络完成文字描述到图像的转换。如图2.9所示，为其使用的基本的框架。该框架首先将文字描述通过 φ 转换为特征 $\varphi(t)$，然后将特征 $\varphi(t)$ 在生成网络的输入端和从 $z \sim N(0, I)$ 采样的噪声连接，然后将该连接向量输入生成网络得到图片。在判别网络中，将 $\varphi(t)$ 和判别网络中间层特征连接，再判断真假。

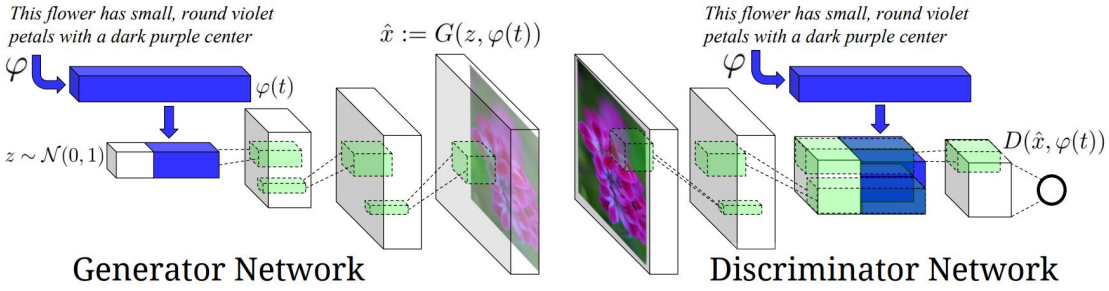


图 2.9　文字到图像转换的架构。左图为生成网络 G，它将文字描述和隐变量连接在一起输入到生成网络 G 中生成符合文字描述的图片。右图为判别网络 D，它将中间层特征和文字描述连接判断生成图片是否符合给定的文字描述。

后续有更多的工作如 GAWWN[64]、stackGAN[2]、stackGAN++[65]、AttnGAN[66]、DA-GAN[67] 基于这个工作提出改进算法继续提升文字到图片的合成质量。

2.3.2　图片到图片的转换

图像转换本质上是一个空间变换（domain transfer) 的问题：假设给定原始空间 X 的大量图片 $\{x_1, x_2, ..., x_n\}$，这些图片在统计意义上具有空间 X 的特征, 图像变换的目标就是学习一个最优的空间变换函数 f, 能够将任意空间 X 的图片 x_i 转换到目标空间 Y，使得转换后的图片 $f(x_i)$ 具备空间 Y 的特征。如图2.10所示为图片到图片转换的一些应用。最近图像合成的发展促生了很多图片到图片转换的工作。其中按照数据的组成划分，可以分为两空间有没有成对图片的转换。

Pix2Pix[68] 是两个空间有成对图片的代表性的工作。它直接使用一个生成网络完成图片到图片的转换，输入是 X 空间的图片，输出是 Y 空间的图片。利用成对的信息，对输出图片直接使用 ℓ_1 损失函数和对抗损失函数的约束，使得输出的图片效果更好。后续有很多工作如：pix2pixHD[62]、CRN[69] 等接着研究这个问题。

CycleGAN[70] 是两个空间无成对图片的代表性的工作。因为两个空间中没有了成对的图片，所以在 pix2pix[68] 中的 ℓ_1 损失函数无法使用。CycleGAN 开创性的使用了循环一致性（cycle-consistence）这一损失函数和在每个空间中使用对抗损失函数解决这一问题。后续的工作 UNIT[71]、DualGAN[72] 也在解决这一问题。

除了一对一的图片转换模型，还有一些工作注重解决一对多的图片转换模型，其中的代表工作如：StarGAN[73]、BiCycleGAN[74] 和 MUNIT[75] 等等。

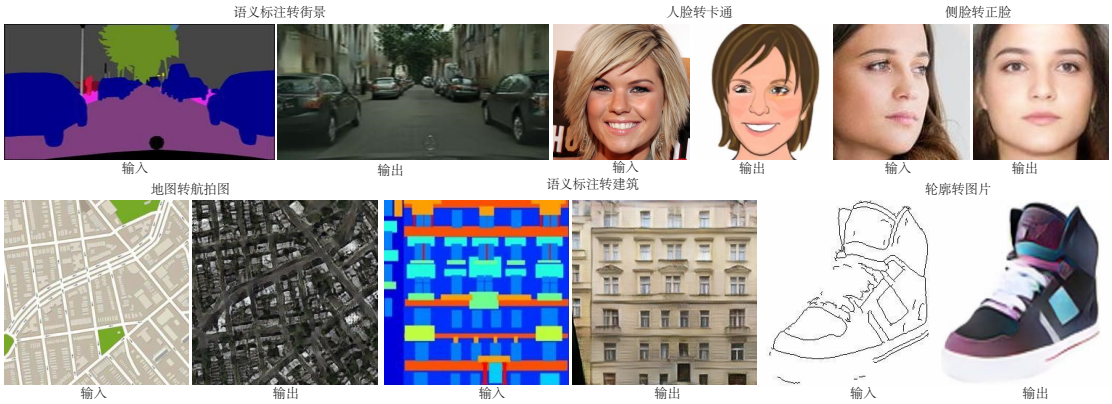


图 2.10　图片到图片转换应用的示意图。

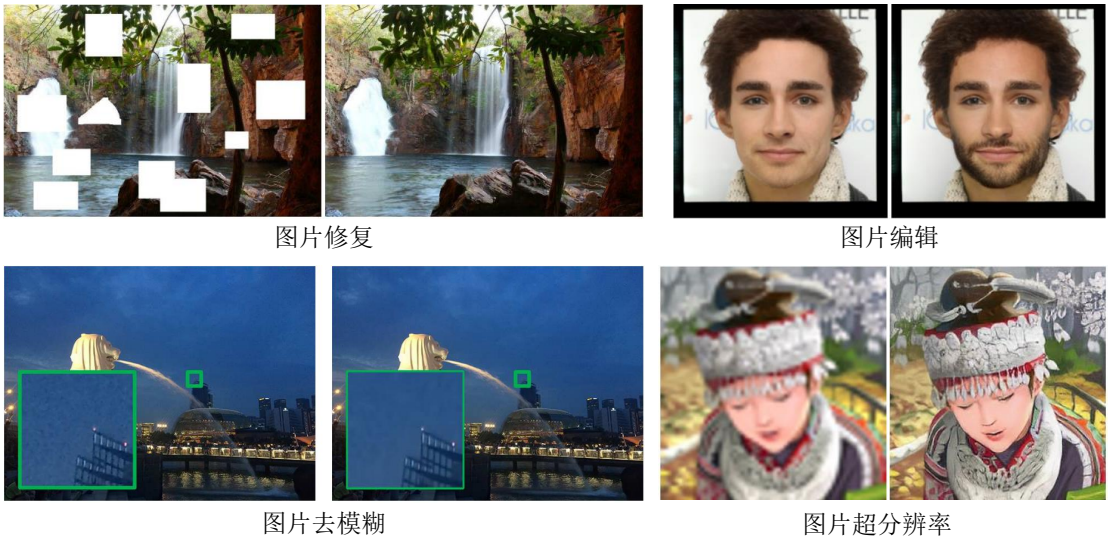


图 2.11　图像合成的应用：修复 [76]，编辑，去模糊 [77] 和超分辨率 [78]。

2.3.3　图片的修复，编辑，去模糊和超分辨率

图像合成技术的发展促进了很多图片底层任务的发展，如图片的修复、编辑、去模糊和超分辨率等等。图2.11中展示了图片合成的应用：修复 [76]，编辑，去模糊 [77] 和超分辨率 [78]。图片修复（image inpainting）[79] 是指原图片中有一块或者多块区域丢失信息，需要用合成的方法将丢失的信息补全。过去的图像修复方法往往都使用图像中已有的片段（Patch）对丢失区域进行修复，现在有了生成网络，很多工作 [76, 80-81] 直接利用大量的图像的训练学到图像中的高级语义信息对图片进行修复。

图片编辑是指用户需要编辑图片中某个或者多个区域。过去的使用的经典方法如泊松克隆 [3]，片段匹配（Patch Match）[82] 的方法注重图片偏底层的特

征的使用，如像素层面，像素梯度层面的特征。现在有了深度图像合成的模型，很多工作 [52, 83-84] 直接使用深度合成模型模拟这一过程。

图片去模糊是指使输入图片中模糊区域变清晰。传统的方法使用数学公式模拟模糊这一过程，然后通过求解该数学公式的逆过程求解出清晰的图像。现在图像合成的发展催生了在该领域的很多工作 [85-86]。

图片超分辨率是提升输入图片的分辨率使之达到更高的分辨率。现在的图像合成模型直接将该过程模拟为图片到图片转换的过程，然后使用现有的技术提升该领域的发展。典型的工作有 SRNet[87] 和 SRGAN[78] 等等。

2.3.4　图像合成的评价标准

自图像合成技术出现，图像质量的评价一直是一个比较有挑战的问题。因为合成图像的质量受到多个因素的影响，如真实性，多样性，与输入条件一致性等等。这几个因素都是非常主观的评价，所以很难找到特别好的数值评价方法。尽管如此，很多工作还是致力于寻找数值上的评价方法。

峰值信噪比（PSNR）[88] 经常用于图像压缩等领域中图片重建质量的评价方法，它常简单地通过均方误差（MSE）[89] 进行定义。两个 $m \times n$ 单色图像 I 和 K，如果一个为另外一个的噪声近似，那么它们的的均方误差定义为：$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$。峰值信噪比定义为：$PSNR = 20 \cdot \log_{10}(\frac{MAX_I}{\sqrt{MSE}})$。其中，$MAX_I$ 是表示图像点颜色的最大数值，如果每个采样点用 8 位表示，那么就是 255。PSNR 同样被用在有正确对应图片（ground-truth）的图片合成质量评价中。峰值信噪比越高，表示合成的效果越好。

但是实际检测结果显示，PSNR 的评价结果无法与人眼看到的视觉品质一致，有可能 PSNR 较高者反而比 PSNR 较低者看起来质量更差。这是因为人眼的视觉对误差的敏感度并不是绝对的，其感知结果因诸多因素的影响而产生变化。基于上面 PSNR 的问题，结构相似性指标（SSIM）[90] 提出了使用两张图片的结构相似性来衡量两张图片相似度的衡量标准。结构相似性指标的值越大，代表两张图片的相似性越高。这样也可以用来衡量合成图片的质量和其对应正确图片的距离。

上面说的两种方法都需要合成图片有对应的正确图片。然而在图片合成任务中，合成的图片常常没有一个对应正确的图片，例如从高斯噪声中合成的图片。所以这个情况下怎么衡量图片的质量呢？Tim Salimans 等人提出了使用 Inception Score[17] 来衡量合成图片的质量。Inception Score 从两个方面来考虑合成图片的质量：（1）真实性，把生成的图片 x 输入进 Inception[91] 分类网络得到其输出维的向量 y，向量的每个维度的值对应图片属于某类的概率。对于一个真实的图片，它属于某一类的概率应该非常大，而属于其它类的概率应该很小，即为

$p(y|x)$ 的熵应该很小。（2）多样性：如果一个模型能生成足够多样的图片，那么它生成的图片在各个类别中的分布应该是平均的，也就是生成图片在所有类别概率的边缘分布 $p(y)$ 熵很大。综合以上两方面得到 Inception Score 的公式为：$\mathbf{IS}(G) = \exp\Big(\mathbb{E}_{\mathbf{x}\sim p_g} D_{KL}\big(p(y|\mathbf{x})||p(y)\big)\Big)$。因为 Inception Score 只使用生成样本进行数值计算，所以 Inception Score 可以在合成图片没有对应正确图片时对合成图片的质量进行评价。

由于在计算 Inception score 时，只考虑了生成样本，没有考虑真实数据，所以 Inception Score 无法反映真实数据和样本之间的距离。所以 Martin Heusel 等人提出了 Fréchet Inception Distance(FID)[92] 的衡量方法。FID 计算了真实图片和假图片在高层特征上的距离。FID 的公式为 $\text{FID} = ||\mu_r - \mu_g||^2 + Tr\big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\big)$，其中 μ_r 为真实图片的特征的均值，μ_g 为生成的图片的特征的均值，Σ_r 为真实图片的特征的协方差矩阵 Σ_g 为生成图片的特征的协方差矩阵。其中使用的特征为 Inception V3[93] 中的全连接层之前的 2048 维特征向量。

为了衡量合成图片与输入条件一致性，很多工作提出使用训练好的图像理解模型作为判别的标准，例如 pix2pix 使用衡量图像分割的方法衡量生成的图片是否符合输入的语义标注。其将生成图片输入到训练好的图像分割网络中得到其预测的语义的标注，将预测的语义标注和输入的语义标注进行比较。如果相差越小，则表示合成图片与输入语义标注的一致性越高。该方法现在已经广泛使用在有条件输入的图像合成的评价标准中。

第 3 章　基于特征匹配条件生成对抗网络的图像合成

如第 2 章中所介绍，生成对抗网络在图像合成任务中取得了史无前例的成功，展现了比其他的合成框架更强的图像合成能力，在很多图像合成应用中展现出非常大的潜力。然而生成对抗网络还是有着训练不稳定 [17]、收敛状态无法判断 [18]、模式坍缩（mode collapse）[17] 等问题。

为了解决生成对抗网络训练不稳定的问题，本章提出了特征匹配损失函数以解决生成对抗网络中训练不稳定的问题。在训练中，对于判别网络，本章使用了和原始生成对抗网络中一样的二元交叉熵损失函数，使其保持判别能力。而对于生成网络，本章使用了特征匹配的损失函数，该损失函数要求生成图像和真实图像在判别网络中的特征中心靠近，这样解决了生成对抗网络原始损失函数中的梯度消失的问题，也就使得生成对抗网络的训练更加稳定。同时，为了满足有条件输入的图像生成，本章在生成对抗网络框架中加入了分类网络，该分类网络被用来约束生成网络合成的图片，使其满足给定的条件输入。在使用分类网络更新生成网络时，本章同样使用了特征匹配的损失函数，该损失函数要求生成的某一类的图片和真实的某一类图片在分类网络 C 中的特征中心靠近。它帮助合成模型生成更加符合输入条件的图片。实验结果表明特征匹配损失函数使得生成对抗网络的训练更加稳定，同时生成图片的质量得到提升。分类网络中的特征匹配损失函数使得生成模型可以更好地保持与输入条件的一致性。

3.1　背景介绍

在原始的生成对抗网络中，生成网络 G 和判别网络 D 进行一个博弈的过程。判别网络 D 尝试将对生成图片和真实图片进行“真或假”分类，而生成网络 G 尝试生成图片使判别网络 D 判别为“真”。所以总的损失函数为：

$$\min_{D}\max_{G} V(D,G) = -\mathbb{E}_{x\sim P_r(x)}[\log D(x)] - \mathbb{E}_{x\sim P_g(x)}[\log(1-D(x)], \tag{3.1}$$

其中 $P_r(x)$ 是真实数据分布，$P_g(x)$ 是生成数据分布。若将其对判别网络 D 和生成网络 G 的损失函数拆开，则对于判别网络 D 的损失函数为：

$$\mathcal{L}_{\text{GAN}}(D) = -\mathbb{E}_{x\sim P_r(x)}[\log D(x)] - \mathbb{E}_{x\sim P_g(x)}[\log(1-D(x))]. \tag{3.2}$$

同时对生成网络 G 的损失函数为：

$$\mathcal{L}_{\text{GAN}}(G) = \mathbb{E}_{x\sim P_g(x)}[\log(1-D(x))]. \tag{3.3}$$

如果假设判别网络 D 到达最优分类器的状态，那么令等式3.2关于 $D(x)$ 的导数为 0，则可得：

$$-\frac{P_r(x)}{D(x)}+\frac{P_g(x)}{1-D(x)}=0. \tag{3.4}$$

化简得最优判别器 $D^*(x)$ 为：

$$D^*(x)=\frac{P_r(x)}{P_r(x)+P_g(x)}. \tag{3.5}$$

给生成器损失函数加上一个不依赖于生成网络 G 的项使之变成：

$$\mathcal{L}'_{\mathrm{GAN}}(G)=\mathbb{E}_{x\sim P_r(x)}[\log(D(x))]+\mathbb{E}_{x\sim P_g(x)}[\log(1-D(x))]. \tag{3.6}$$

将最优化的 $D^*(x)$ 带入到式3.6中得到生成网络 G 的损失函数为：

$$\mathbb{E}_{x\sim P_r}\log\frac{P_r(x)}{\frac{1}{2}[P_r(x)+P_g(x)]}+\mathbb{E}_{x\sim P_g}\log\frac{P_g(x)}{\frac{1}{2}[P_r(x)+P_g(x)]}-2\log 2. \tag{3.7}$$

又由 JS 散度等于：

$$JS(P_1||P_2)=\frac{1}{2}KL(P_1||\frac{P_1+P_2}{2})+\frac{1}{2}KL(P_2||\frac{P_1+P_2}{2}). \tag{3.8}$$

所以生成网络 G 的损失函数 $\mathcal{L}'_{\mathrm{GAN}}(G)$（式3.6）可以等价于下式：

$$2JS(P_r||P_g)-2\log 2. \tag{3.9}$$

此时，研究者可以把原始生成对抗网络定义的生成网络 G 的损失函数等价变换为最小化真实分布 P_r 与生成分布 P_g 之间的 JS 散度（忽略其中的常数项 $-2\log 2$）。越训练判别器，它就越接近最优，最小化生成器的损失函数也就会越近似于最小化 P_r 和 P_g 之间的 JS 散度。问题就出在这个 JS 散度上，当两个分布没有重叠区域的时候，JS 散度为零，所以此时判别网络 D 传给生成网络 G 的梯度也为 0。所以生成网络 G 的训练是不稳定的。

如图3.1中所示，左图中蓝色部分表示真实数据分布，红色部分表示生成数据分布，右图表达的是生成对抗网络的生成网络 G 的损失函数 $\mathcal{L}'_{\mathrm{GAN}}(G)$ 在真实数据分布和生成数据分布在不同距离时的损失函数值的大小。可以看到，在真实数据分布和生成数据分布没有重叠区域的时候，生成网络 G 的损失函数 $\mathcal{L}'_{\mathrm{GAN}}(G)$ 的值都为 0，所以这个时候是没有梯度可以回传给生成网络 G 的。也就是这个时候产生了梯度消失的问题。

最近的工作 WGAN[47, 94] 也在理论上解释了生成对抗网络训练中出现的生成网络 G 梯度不稳定的原因。针对这一问题，WGAN 提出了使用 Wasserstein 距

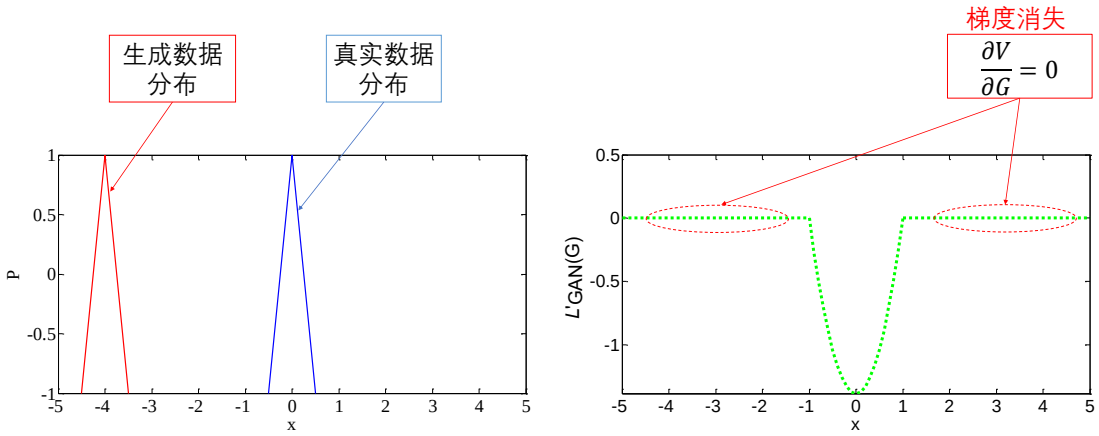


图 3.1　原始生成对抗网络中，生成网络 G 的损失函数 $\mathcal{L}'_{\mathrm{GAN}}(G)$ 在生成数据和真实数据分布距离不同时的损失函数值的变化，左图中蓝色部分表示真实数据分布，红色部分表示生成数据分布，右图表达的是生成网络 G 的原始损失函数 $\mathcal{L}'_{\mathrm{GAN}}(G)$ 在真实数据分布和生成数据分布在不同距离时值的大小。

离作为生成对抗网络训练的损失函数，并理论推导出生成对抗网络训练中判别网络 D 回传给生成网络 G 不出现梯度消失的关键是判别网络 D 对输入 x 的输出 $D(x)$ 应该满足 Lipschitz 连续，为了使判别网络 D 满足这个条件，WGAN 采取了一个非常简单的方法，即在训练过程中限制判别网络 D 中所有参数 w_i 不超过某个范围 [-c, c]，比如 $w_i \in [-0.01, 0.01]$。这种简单的做法虽然使判别网络 D 的映射表达 $D(x)$ 满足了 Lipschitz 连续条件，但是这种限制网络参数的方法会导致判别网络 D 能力受到限制，其带来的负面影响是会使生成对抗网络的训练变得非常慢。

针对生成对抗网络训练不稳定的问题和 WGAN 的局限性，本章提出新的损失函数。该损失函数基于特征匹配（Feature Matching）的方法，它改善了生成对抗网络的训练稳定性。在训练中，对于判别网络 D，本章使用了和原始生成对抗网络中一样的二元交叉熵损失函数，使其保持判别能力。而对于生成网络 G，本章使用了特征匹配的损失函数，该损失函数要求生成数据和真实数据在判别网络 D 中的特征中心靠近，这样解决了原始损失函数中的梯度消失的问题，也就使得生成对抗网络的训练更加稳定。

另外，为了使生成模型可以根据输入条件进行图片合成，之前的工作条件生成对抗网络（CGAN）[27] 将条件输入和判别网络中的某一层特征连接进行“真或假”的判别，这种将条件直接和特征连接的方式并未能充分地使用条件信息。所以本章提出在生成对抗网络框架中加入分类网络。分类网络使用图片和其对应的条件进行分类训练，这样分类网络便充分地使用了条件信息。然后框架可使用分类网络约束生成模型，使其合成更加符合条件的图片。在使用分类网络更新生成网络时，同样可使用特征匹配的损失函数，该损失函数要求生成的某一类的

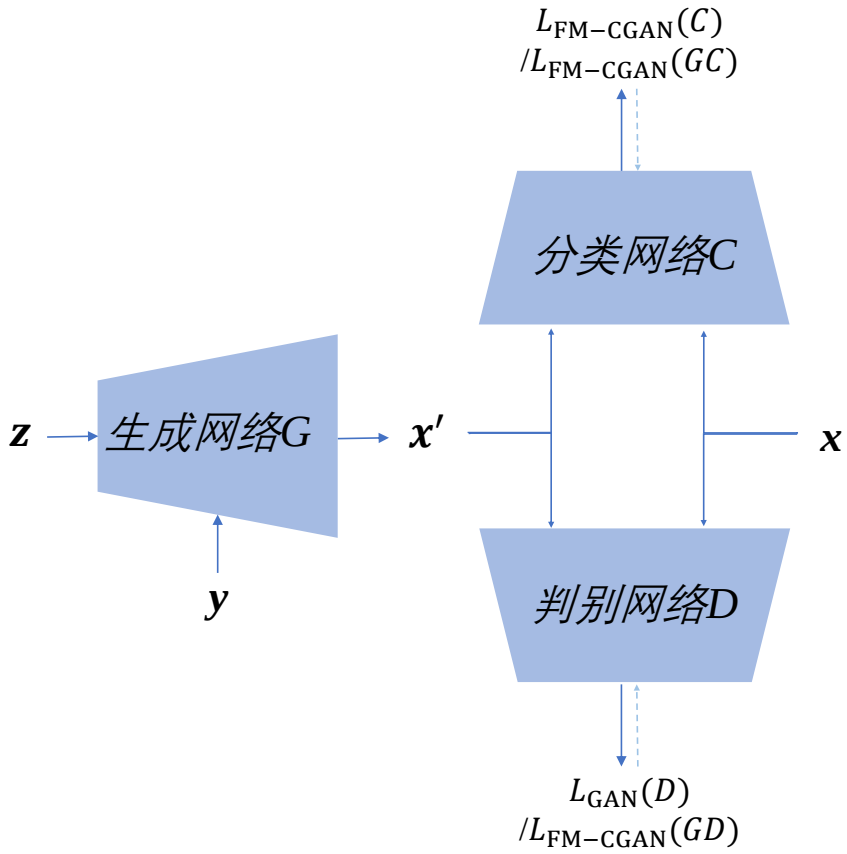


图 3.2　特征匹配条件生成对抗网络（FM-CGAN）框架示意图，它包含三个部分：（1）生成网络 G 将随机噪声 z 和标签信息 y 作为输入得到输出 $x' = G(z, y)$；（2）判别网络 D 对输入的真实数据 x 和生成数据 x' 进行二分类；（3）分类网络 C 约束生成图片 x' 使其满足给定的标签输入 y。其中损失函数的定义参见3.2.2章节和3.2.3章节。

图片和真实的某一类图片在分类网络中的特征中心靠近。它进一步帮助合成模型生成更加符合输入条件的图片。图3.2为框架的示意图，本章的框架包括三个部分：（1）生成网络 G；（2）判别网络 D；（3）分类网络 C。由于训练中使用了特征匹配损失函数，所以本章的合成框架被命名为特征匹配条件生成对抗网络（FM-CGAN）。

本章其余部分组织如下，在3.2小节中将具体介绍特征匹配条件生成对抗网络的框架结构与损失函数；在3.3小节中介绍特征匹配条件生成对抗网络的实现细节；在3.4小节中用实验证明特征匹配条件生成对抗网络的实用性;3.5小节进行小结与讨论。

3.2　特征匹配条件生成对抗网络

3.2.1　算法框架

本章节将介绍特征匹配条件生成对抗网络的算法框架。该框架主要专注于解决使用标签作为输入的图像合成的问题。如图3.2所示，论文提出的框架包括三个部分：（1）生成网络 G；（2）判别网络 D；（3）分类网络 C。

生成网络 G 的作用是接受输入的随机噪声 z 和输入的标签信息 y 得到输出 $x' = G(z, y)$，它尝试学到的由标签和隐空间输入到真实数据 x 的分布的图像合成。判别网络 D 对输入的真实数据 x 和生成数据 x' 进行二分类，所以判别网络 D 和生成网络 G 进行对抗学习的过程。他自身使用的损失函数 $\mathcal{L}_{\text{GAN}}(D)$ 和其回传给生成网络 G 的损失函数 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 将在3.2.2章节中进行具体介绍。分类网络 C 使用真实数据 x 的标签 y 进行分类训练，然后利用学到的分类特征来约束生成图片 x' 使其满足给定的标签输入。其使用的分类损失函数 $\mathcal{L}_{\text{FM-CGAN}}(C)$ 和其回传给生成网络 G 的损失函数 $\mathcal{L}_{\text{FM-CGAN}}(GC)$ 将在3.2.3章节中进行具体介绍。

3.2.2　判别网络 D 中的特征中心匹配

为了解决生成对抗网络训练不稳定的问题，论文提出了在判别网络 D 中使用特征中心匹配的方法训练生成网络 G。该损失函数要求生成数据和真实数据在判别网络 D 中的特征中心接近。假设 $f_D(x)$ 表示判别网络中的某一层特征，生成图片的表达式为 $G(z, y)$（其中 z 为隐空间变量，y 为输入标签），那么此时生成网络 G 尝试去最小化该损失函数：

$$\mathcal{L}_{\text{FM-CGAN}}(GD) = -\frac{1}{2}||\mathbb{E}_{x\sim P_r} f_D(x) - \mathbb{E}_{z\sim P_z} f_D(G(z, y))||_2^2. \tag{3.10}$$

在本章的实验中，论文选择判别网络中最后一层卷积层的输入作为特征 f_D。同样如果使用多层特征中心的匹配可以使生成对抗网络的训练更加稳定，同时收敛的更快。在训练中，论文使用滑动平均的方法求得真实数据和生成数据的特征中心。

这个时候如果使用在背景介绍中对于判别网络 D 回传到生成网络 G 的梯度的分析方式可以知道，特征匹配损失函数在生成数据分布和真实数据分布没有重叠区域时依然存在梯度，这也就解决了梯度消失的问题。如图3.3中所示，和上面一样的设置，左图中蓝色部分表示真实数据分布，红色部分表示生成数据分布，右图表达的是生成对抗网络中生成网络 G 的损失函数 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 在真实数据分布和生成数据分布在不同距离时的特征匹配的损失函数值的大小。可以看到，在真实数据分布和生成数据分布没有重叠区域的时候，生成网络 G 的损

失函数 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 回传给生成网络 G 仍然存在梯度，这样生成网络 G 可以得到由判别网络 D 中回传的梯度的更新。所以生成对抗网络训练中出现的梯度消失问题得以解决。

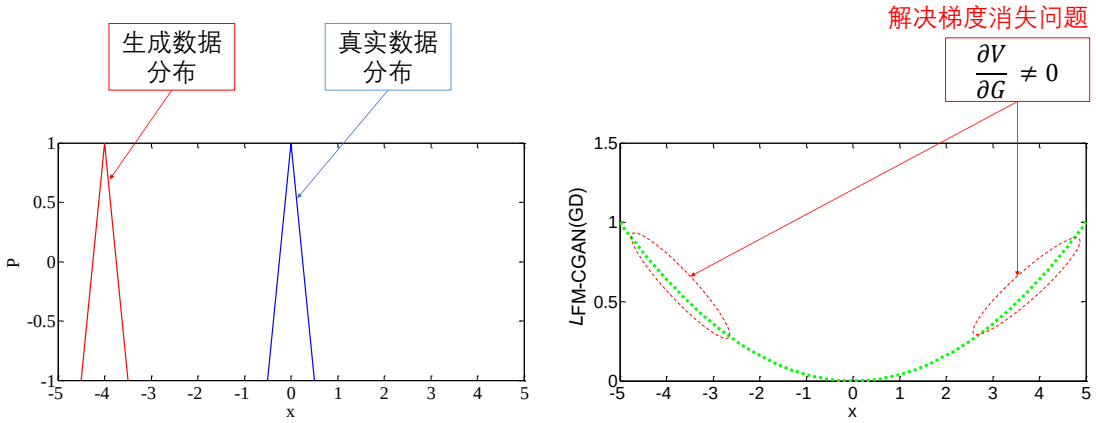


图 3.3　特征匹配生成对抗网络中，生成网络 G 的损失函数 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 在生成数据和真实数据分布距离不同时的值的变化，左图中蓝色部分表示真实数据分布，红色部分表示生成数据分布，右图表达的是判别网络 D 回传给生成网络 G 的损失函数 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 在真实数据分布和生成数据分布在不同距离时值的大小。

因此，在特征匹配条件生成对抗网络框架的训练中，用公式3.2更新判别网络 D 的参数，用公式3.10更新生成网络 G。这种非对称的训练方式会给特征匹配条件生成对抗网络框架的训练带来以下三点好处：

1. 使用特征中心匹配损失函数 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 使得特征匹配条件生成对抗网络框架的训练更加稳定，因为特征中心匹配损失函数不会导致判别网络 D 回传给生成网络 G 的梯度为零的情况。
2. 由于网络中特征的优秀表达能力，生成数据和真实数据的在判别网络 D 中的特征中心靠近使得生成模型生成图片质量得到提升。
3. 与 WGAN[47] 对比，特征中心匹配不需要对判别网络中的参数进行限制，从而判别网络的能力得到了保证，所以该方法不会影响训练的速度。

3.2.3　分类网络 C 中的特征中心匹配

在本章节中，论文将介绍特征匹配条件生成对抗网络框架中分类网络 C 中的特征中心匹配损失函数。假设有一个 K 个类 $\{y_1, y_2, \cdots, y_K\}$ 的图片集合，论文使用一个分类网络 C 去衡量其中的图片是不是属于其中的某一个类。如果使用最简单的方法去做分类，分类网络 C 使用图片 x（其标签为 y）作为输入，然后输出一个 K 维的向量，然后使用一个 softmax 函数将其转化为属于 K 类的概率。该向量的每一维表示的是其属于某个类概率。在训练阶段，分类网络 C 尝试去最小化 softmax 损失：

$$\mathcal{L}_{\text{FM-CGAN}}(C) = -\mathbb{E}_{x\sim P_r}[\log P(y|x)], \tag{3.11}$$

其中 P_r 是真实数据的分布，y 表示为每一个输入图片 x 对应的标签。$P(y|x)$ 是分类网络估计输入图片 x 为标签 y 的概率。

由于分类网络中特征优秀的表达能力 [95-97]，论文在使用分类网络给生成网络 G 回传梯度时同样使用特征中心匹配的损失函数。假设 $f_C(x)$ 代表分类网络 C 在输入图片为 x 的某一层的特征。则对于所有的 K 个类 $\{y_1, y_2, \cdots, y_K\}$，可以得到第 k 个类真实图片特征中心为 $\mathbb{E}_{x\sim P_r^k}[f_C(x)]$，其中 P_r^k 表示图片中所有真实图片中属于第 k 类的的图片。对生成图片 $G(z, y_i), y_i \in \{y_1, y_2, \cdots, y_K\}$，$z$ 表示隐空间变量的分布，y_i 表示输入标签。生成图片的第 k 个类的特征中心为 $\mathbb{E}_{z\sim P_z}[f_C((G(z, y_k))]$。则分类网络特征中心的损失函数定义为下式：

$$\mathcal{L}_{\text{FM-CGAN}}(GC) = \frac{1}{2}\sum_k ||\mathbb{E}_{x\sim P_r^k} f_C(x) - \mathbb{E}_{z\sim P_z} f_C(G(z, y_k))||_2^2. \tag{3.12}$$

在实验中论文使用了分类网络最后一个全连接层（Fully-Conneted Layer）的输入作为分类网络中的特征。论文发现在分类网络 C 中使用多层的特征中心匹配会使结果更好。在实验中每一批（Batch）训练数据中中属于一个类的图片很少，所以论文对于每个类的特征中心使用了滑动平均统计的方法。

3.2.4　整体损失函数

综合上面的分析，可以知道训练特征匹配条件生成对抗网络的损失函数有 $\mathcal{L}_{\text{GAN}}(D)$，$\mathcal{L}_{\text{FM-CGAN}}(GD)$，$\mathcal{L}_{\text{FM-CGAN}}(C)$，$\mathcal{L}_{\text{FM-CGAN}}(GC)$。所以总的损失函数为：

$$\mathcal{L}_{\text{FM-CGAN}} = \mathcal{L}_{\text{FM-CGAN}}(C) + \mathcal{L}_{\text{GAN}}(D) + \mathcal{L}_{\text{FM-CGAN}}(GD) + \lambda_1\mathcal{L}_{\text{FM-CGAN}}(GC). \tag{3.13}$$

其中只有 $\mathcal{L}_{\text{FM-CGAN}}(GD)$ 和 $\mathcal{L}_{\text{FM-CGAN}}(GC)$ 同时给生成网络 G 回传梯度，所以需要使用一个系数 λ_1 去调节这二者的权重，在实验中，论文使用的 λ_1 为 1。

3.3　实现细节与算法

在本章节中，论文将介绍特征匹配条件生成对抗网络的实现细节，首先在3.3.1章节中，论文将介绍每个网络的具体结构。在3.3.2中介绍特征匹配条件生成对抗网络训练中所使用的算法。

表 3.1　生成网络 G 的结构。

网络层	核大小/步长	输出大小
全连接	-	8192
变形 (reshape)	-	$4 \times 4 \times 512$
反卷积	$4 \times 4/2$	$8 \times 8 \times 512$
反卷积	$4 \times 4/2$	$16 \times 16 \times 256$
反卷积	$4 \times 4/2$	$32 \times 32 \times 256$
反卷积	$4 \times 4/2$	$64 \times 64 \times 128$
反卷积	$4 \times 4/2$	$128 \times 128 \times 64$
卷积	$5 \times 5/1$	$128 \times 128 \times 3$

表 3.2　判别网络 D 的结构。

网络层	核大小/步长	输出大小
卷积	$4 \times 4/2$	$64 \times 64 \times 64$
卷积	$4 \times 4/2$	$32 \times 32 \times 128$
卷积	$4 \times 4/2$	$16 \times 16 \times 256$
卷积	$4 \times 4/2$	$8 \times 8 \times 512$
卷积	$4 \times 4/2$	$4 \times 4 \times 512$
全连接	-	1

3.3.1　网络结构

实验中生成的图片的分辨率为 128×128，所以生成网络的输出为 128×128×3，判别网络 D 和分类网络 C 的输入均为 $128 \times 128 \times 3$，遵循 DCGAN[18] 中的设计准则，论文设计了如表3.1中结构的生成网络 G，如表3.2中结构的判别网络 D。对于分类网络，论文使用了一个简单的卷积神经网络结构，如表3.3所示。在各表中，网络层从上以下的层表示该网络结构中的层，核大小/步长代表了卷积层、池化层、反卷积层使用的核大小和步长。输出大小表示输入经过该层之后得到的输出特征的大小。其中前 2 维表示特征的长和高，第 3 维表示特征的通道数。在卷积层和反卷积层后面均使用了批归一化（Batch Normalization）层和激活函数层，对于判别网络论文使用了系数为 0.2 的 LeakyReLU[56] 层，生成网络和分类网络中均使用了 ReLU 激活函数层。

表 3.3　分类网络 C 的结构。

网络层	核大小/步长	输出大小
卷积	$7 \times 7/1$	$128 \times 128 \times 64$
最大池化	$2 \times 2/2$	$64 \times 64 \times 64$
卷积	$3 \times 3/1$	$64 \times 64 \times 128$
最大池化	$2 \times 2/2$	$32 \times 32 \times 128$
卷积	$3 \times 3/1$	$32 \times 32 \times 256$
最大池化	$2 \times 2/2$	$16 \times 16 \times 256$
卷积	$3 \times 3/1$	$16 \times 16 \times 512$
最大池化	$2 \times 2/2$	$8 \times 8 \times 512$
卷积	$3 \times 3/1$	$8 \times 8 \times 512$
最大池化	$2 \times 2/2$	$4 \times 4 \times 512$
卷积	$3 \times 3/1$	$4 \times 4 \times 512$
全连接	-	1024
全连接	-	类别数

3.3.2　算法流程

在框架训练过程中，论文提出的框架需要先训练分类网络 C，然后利用 C 网络的分类能力保证从标签生成的图片满足给定的标签。同时论文需要让生成网络 G 和判别网络 D 进行对抗式的训练，训练过程如算法3.1所示。

3.4　实验评估

在本章节中，本文将使用实验验证论文提出的特征匹配条件生成对抗网络的实用性。首先论文将使用一个简单数据分布的拟合实验对论文提出的特征匹配损失函数进行分析。然后论文将在具体数据集上进行具体的图片合成实验对提出的模型进行定性和定量的分析。

3.4.1　简单的例子的分析

本节使用一个简单的例子的分析特征匹配条件生成对抗网络框架的优势。在该简单的例子中，在原始特征匹配条件生成对抗网络框架的基础上，移除掉标签输入和分类网络 C，即训练一个仅使用高斯噪声作为输入拟合数据分布的生成模型。这就变成了一个标准的只有生成网络 G 和判别网络 D 的生成对抗网络模型，训练中对判别网络 D 和生成网络 G 使用的损失函数分别是论文在上文中介绍的

算法 3.1　特征匹配条件生成对抗网络（FM-CGAN）的训练算法。

Data: 训练的 Batch Size 为 m，所有的类的个数为 K，θ_G、θ_D，θ_C 分别为生成网络 G，判别网络 D 和分类网络 C 的参数。λ_1 值为 1。

while 生成网络 G 没有收敛 **do**

从 $P_r(x)$ 中采样 m 个真实数据样本 $\{x_1, ..., x_m\}$; 它们的标签为 $\{y_1, ..., y_m\}$。

$\mathcal{L}_{\text{FM-CGAN}}(C) \leftarrow -\frac{1}{m}\sum_{i=1}^{m}\log(P(y_i|x_i))$。

从 $P_z \sim N(0, I)$ 中采样 m 个样本 $\{z_1, ..., z_m\}$。

通过生成器得到生成数据 x'_i: $x'_i = G(z_i, y_i)$。

$\mathcal{L}_{\text{FM-CGAN}}(D) \leftarrow -\frac{1}{m}\sum_{i=1}^{m}[\log(D(x_i)) + \log(1 - D(x'_i))]$.

计算判别函数中的真实数据特征中心 $\frac{1}{m}\sum_{i=1}^{m} f_D(x_i)$。

计算生成数据的特征中心 $\frac{1}{m}\sum_{i=1}^{m} f_D(x'_i)$。

$\mathcal{L}_{\text{FM-CGAN}}(GD) \leftarrow \frac{1}{2}||\frac{1}{m}\sum_{i=1}^{m} f_D(x_i) - \frac{1}{m}\sum_{i=1}^{m} f_D(x'_i)||_2^2$.

计算分类网络中真实数据每个 y_i 类的中心 $f_C^{y_i}(x_i)$。

计算分类网络中生成数据每个 y_i 类的中心 $f_C^{y_i}(x'_i)$。

$\mathcal{L}_{\text{FM-CGAN}}(GC) \leftarrow \frac{1}{2}\frac{1}{K}\sum_{y_i=1}^{K}||f_C^{y_i}(x_i) - f_C^{y_i}(x'_i)||_2^2$.

使用梯度下降法更新所有网络的参数:

$\theta_C \xleftarrow{+} -\nabla_{\theta_C}(\mathcal{L}_{\text{FM-CGAN}}(C))$.

$\theta_D \xleftarrow{+} -\nabla_{\theta_D}(\mathcal{L}_{\text{FM-CGAN}}(D))$.

$\theta_G \xleftarrow{+} -\nabla_{\theta_G}(\mathcal{L}_{\text{FM-CGAN}}(GD) + \lambda_1\mathcal{L}_{\text{FM-CGAN}}(GC))$.

end

$\mathcal{L}_{\text{GAN}}(D)$ 和 $\mathcal{L}_{\text{FM-CGAN}}(GD)$，因使用了特征匹配损失函数，所以论文称该模型为特征匹配生成对抗网络（FMGAN）。如图3.4所示，假设有如（a）中的一个中心在 $(100, 100)$ 处的环形的真实数据分布。论文比较使用原始 GAN，Wasserstein GAN 和 FMGAN 拟合这个分布的结果。原始 GAN 中对判别网络 D 和生成网络 G 使用的损失函数分别是第二章中介绍的 $\mathcal{L}_{\text{GAN}}(D)$ 和 $\mathcal{L}_{\text{GAN}}(G)$。Wasserstein GAN 中对判别网络 D 和生成网络 G 使用的损失函数分别是第二章中介绍的 $\mathcal{L}_{\text{WGAN}}(D)$ 和 $\mathcal{L}_{\text{WGAN}}(G)$。三种生成对抗网络之间的唯一不同点是使用了不同的损失函数。

三种生成对抗网络使用完全一样的训练设置。生成网络 G 的输入是一个二维向量，网络结构是一个 4 层全连接层，前三层的维数分别是 32, 64, 64，最后一层是输出到 2 维的向量。判别网络 D 网络结构也是一个 4 层全连接层，前三层的维数分别是 32, 64, 64，最后一层是输出到 1 维的数值。实验中使用 RMSProp[98] 的优化方法，学习率使用的是 0.00005。实验中分别训练原始 GAN、Wasserstein

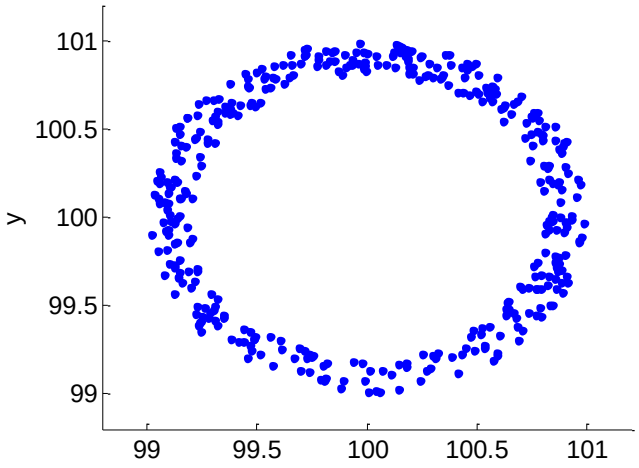


图 3.4　待拟合真实数据分布，其为中心在（100，100）处的圆环。

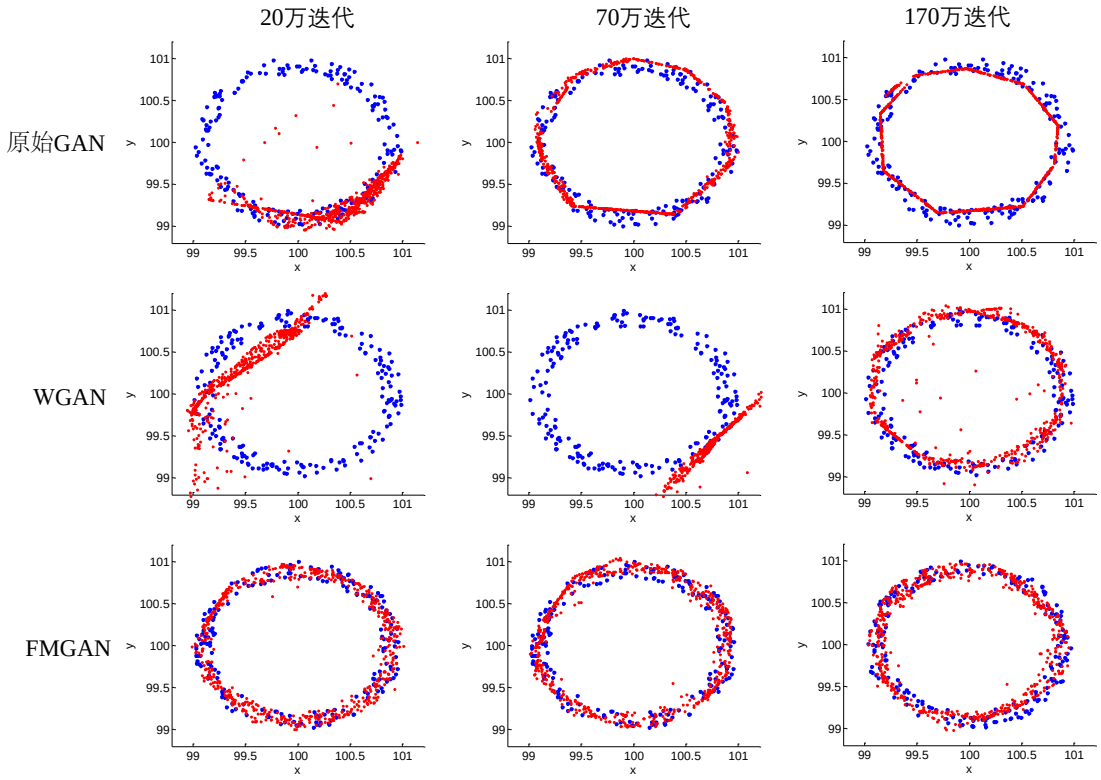


图 3.5　原始 GAN、WGAN 和 FMGAN 在不同迭代次数的拟合真实数据分布的结果。蓝色点表示待拟合的真实数据分布，红色点表示不同生成对抗网络生成的数据分布。可以看出原始 GAN 只拟合了圆环中的一条细线，出现了模式坍缩的问题，WGAN 拟合数据的速度很慢。而 FMGAN 快速拟合数据分布并且拟合的最好。

GAN、FMGAN　200 万次直到所有的模型均为收敛状态。三种生成对抗网络在不同迭代次数生成的数据分布如图3.5所示。图中蓝色点表示待拟合的真实数据分布，红色点表示不同生成对抗网络生成的数据分布，从生成的数据分布中可以观察到以下现象：

1. 对于原始的 GAN（图3.5中第一行），最终生成的点只在一条细线上，并没有覆盖整个环形区域。这就是 GAN 中常见的模式坍缩（mode collapse）问题，而且这个问题一直存在于训练过程中。
2. 对于 Wasserstein GAN（图3.5中第二行），它在训练的早期不能学到学习到真实的数据分布。本文认为这是由于 Wasserstein GAN 中的裁剪判别网络参数导致的。
3. 对于 FMGAN（图3.5中第三行），它很快学习到真实数据的分布。并且学习到的分布更加准确。

通过以上现象，论文证明了特征匹配损失函数提升了生成对抗网络训练的稳定性，同时也使得生成对抗网络收敛的更快。

3.4.2　数据集与训练设置

在本小节中，本文将介绍实验使用的数据集以及训练设置。论文在三个数据集，FaceScrub [99]、102 Category Flower [100]、和 CUB-200 [101] 上进行试验。这三个数据集包含三种完全不同种类的图片，分别是人脸，花和鸟类。

本文提出框架中的生成模型 G 在所有数据集上合成图像的分辨率均为 128×128。对于 Facescrub 数据集，论文首先使用人脸检测算法 JDA[102] 去检测人脸位置，然后根据得到脸上 5 点位置（双眼，鼻子，双嘴角）将人脸对齐到一个固定的位置上。最终裁剪出一个 128×128 分辨率的脸部区域。对于 102 Category Flower 数据集，论文使用其语义标注信息裁剪出值包含花的区域出来。然后将其分辨率缩放到 128×128 大小。对于 CUB-200 数据集，论文直接使用了其裁剪出来的图片作为训练数据。

在训练设置中，论文使用了 Adam[103] 的优化策略，学习率为 0.0002，β_1 为 0.5，β_2 为 0.999，训练中使用的 4 块 K40 卡，Batch Size 设置为 128。

3.4.3　与其他模型合成图像的质量比较

在本节中，论文比较了特征匹配条件生成对抗网络与其他条件生成模型的图像合成质量。论文比较的主要方法有条件变分自编码器（CVAE）和条件生成对抗网络（CGAN），如第二章中所介绍，条件变分自编码器 [26] 是由 Xinchen Yan 等人在 2015 年提出解决从属性信息生成图片的条件生成模型。条件生成对抗网络 [63] 是 Mehdi Mirza 等人提出的使用生成对抗网络完成条件生成的模型。为了公平的比较每种方法，论文在每种方法的训练中使用了相同的生成网络 G 结构和相同的数据。所有的框架使用相同的训练策略。在测试阶段，所有方法只使用生成网络 G 去生成图片。

论文在三个数据集 FaceScrub [99], 102 Category Flower [100], 和 CUB-

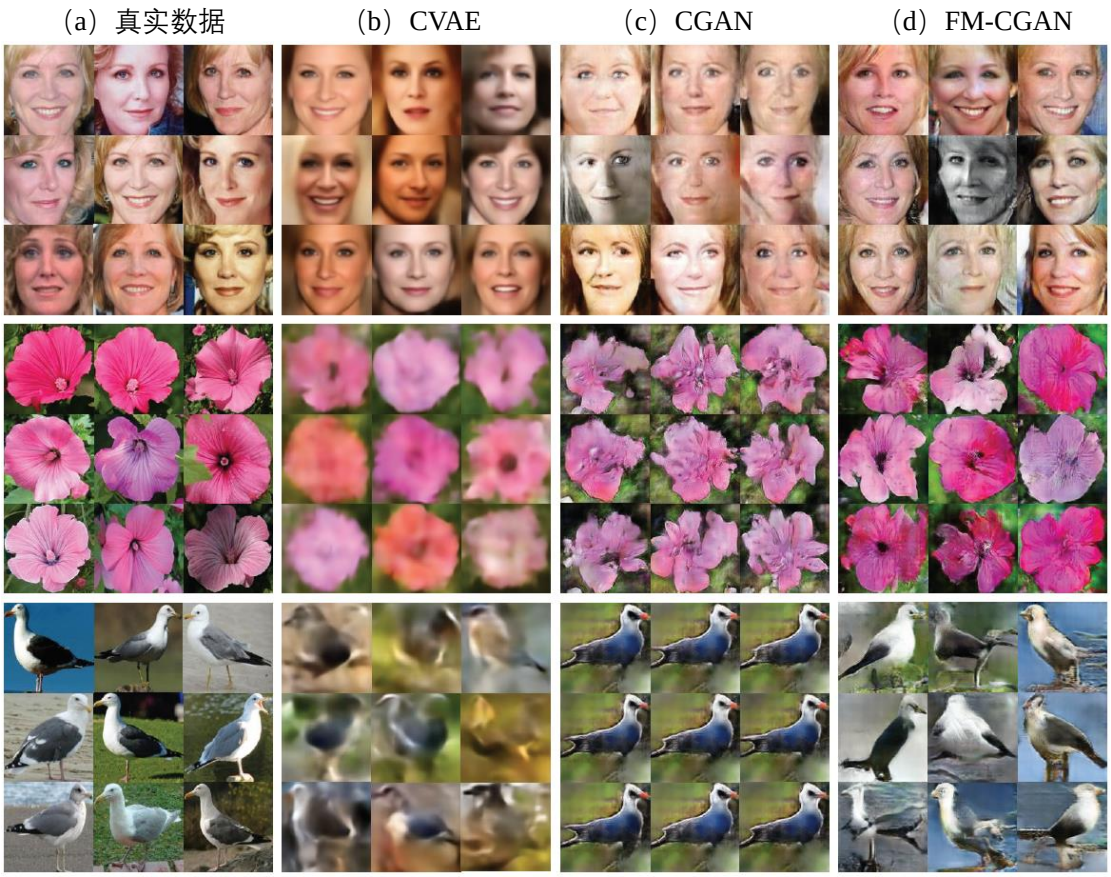


图 3.6　不同条件生成模型在 Facescrub, 102 Category Flower, 和 CUB-200 数据集上生成的图片结果比较。（a）为某一类的真实数据；（b）为条件自编码器（CVAE）[26] 生成的结果，（c）为条件生成对抗网络（CGAN）[27] 生成的结果，（d）为论文提出的特征匹配条件生成对抗网络（FM-CGAN）的结果。

200 [101] 比较生成图片的质量。在各个数据集中论文均随机选择了一个标签 y_i，然后从隐变量空间分布 $N(0, I)$ 中采样很多个隐变量，然后一起输入到生成网络 G 中得到合成图片，如图3.6所示为不同方法生成图片的视觉结果比较。其中（a）为某一类的真实数据；（b）为条件自编码器（CVAE）[26] 生成的结果；（c）为条件生成对抗网络（CGAN）[27] 生成的结果；（d）为论文提出的特征匹配条件生成对抗网络（FM-CGAN）的结果。

比较不同方法生成图片的质量可以知道：条件自编码器（CVAE）合成的图像模糊不清，这是由于 ℓ_2 损失函数导致的。条件生成对抗网络（CGAN）在每个类种合成图片的结果的多样性非常有限，特别是在 CUB-200 数据集上合成的鸟的图片视觉上看起来一样，这就是模式坍缩（mode collapse）问题。另外一方面，论文提出的特征匹配条件生成对抗网络（FM-CGAN）框架合成的图片更加清晰，同时在每个类中图片的多样性也更大。这展示了特征匹配条件生成对抗网络模型在图像合成中的优势。

表 3.4　真实图片、条件自编码器（CVAE）合成图片、条件生成对抗网络（CGAN）合成图片与特征匹配条件匹配生成对抗网络（FM-CGAN）合成图片质量的数值结果比较。相比于条件自编码器和条件生成对抗网络模型，特征匹配条件匹配生成对抗网络框架生成的图片更好地保持了输入的标签信息，同时具有更好的真实性和多样性。

数据来源	top-1 准确率	Inception score
真实图片	99.61%	20.85
CVAE	8.09%	10.29
CGAN	61.97%	15.79
FM-CGAN	79.76%	19.40

3.4.4　与其他模型合成图像的数值比较

图像合成模型的数值比较一直是一个非常挑战的问题，如2.3.4章节中介绍的一样，论文尝试在三个尺度上衡量合成模型合成的图片质量：与输入标签一致性，多样性，真实性。

论文使用 Facescrub 人脸数据集来做这个数值比较的实验，首先，论文分别用条件自编码器（CVAE），条件生成对抗网络（CGAN）和特征匹配条件生成对抗网络（FM-CGAN）随机的生成 53000 张图片（每个类 100 张）。

为了衡量生成图片与输入标签的一致性。论文先使用人脸数据训练好一个人脸识别网络，然后论文使用这个人脸识别网络来衡量生成图片的标签是否是输入标签，即论文比较将生成图片 $G(z, y_i)$ 输入人脸识别网络得到概率最高的标签是不是 y_i，即得到 top-1 准确率。该 top-1 准确率反应了生成图片是不是更好地保持了输入的标签信息。如表3.4所示，与条件自编码器和条件生成对抗网络模型比较，特征匹配条件生成对抗网络模型生成的图片更好地保持了输入的标签信息。

为了衡量生成图片的多样性和真实性，如2.3.4小节中介绍，论文使用 Inception Score([17]) 来衡量。首先论文使用 CASIA 人脸数据集 [104] 训练了一个人脸识别模型。然后使用 Inception Score 的衡量指标 $\mathbf{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g} D_{KL}\left(p(y|\mathbf{x})||p(y)\right)\right)$ 来计算不同生成模型得到的数值结果。高的 Inception Score 数值代表了模型生成了真实性更高和多样性更好的图片。如表3.4所示，相比条件自编码器框架和条件生成对抗网络框架，特征匹配条件生成对抗网络框架生成的图片具有更高的真实性和多样性。

3.5　小结与讨论

在本章中，论文分析了生成对网络训练不稳定问题的成因。针对该问题，论文提出了特征匹配损失函数。在训练中，对于判别网络，论文使用了和原始生成对抗网络中一样的二元交叉熵损失函数，使其保持判别能力。而对于生成网络，论文使用了特征匹配的损失函数，该损失函数要求生成图像和真实图像在判别网络中的特征中心靠近，这样解决了生成对抗网络原始损失函数中的梯度消失的问题，也就使得生成对抗网络的训练更加稳定。同时了满足指定标签的图像合成，论文在生成对抗网络框架中加入分类网络，并且在分类网络中使用了特征损失函数。在简单数据分布和三个数据集上的图像合成结果表明我们提出算法的有效性。和其他现有的图像合成算法相比，本章提出的方法也取得了明显的性能提升。

第 4 章　基于条件变分生成对抗网络的图像合成

通过第 2 章的介绍和第 3 章的讨论，研究者们可以发现图像合成随着生成对抗网络的发展得到快速的进步。但是生成对抗网络的问题如训练不稳定，收敛状态无法判断，模式坍缩（mode collapse）等问题仍然困扰着研究者们。在第 3 章中提出的特征匹配条件生成对抗网络（FM-CGAN）框架中，论文提出的特征匹配损失函数改进了生成对抗网络训练的稳定性问题。但是生成对抗网络中模式坍缩问题仍然亟需解决。

为了解决模式坍缩问题，本章提出一个新的框架：条件变分生成对抗网络（CVAE-GAN）。该框架将编码网络加入生成对抗网络的框架中，以解决生成对抗网络中的模式坍缩问题。编码网络将图片空间映射到隐空间，再使用生成网络将隐空间映射回图片空间，因为原图片空间的分布中的图片是多样的，所以生成网络生成的图片也是多样的。这样便解决了生成对抗网络中的模式坍缩问题。实验结果表明，条件变分生成对抗网络框架相对其他框架生成了多样性更高的图片，从而证明该框架有效解决了模式坍缩问题。同时该框架可以完成很多应用：细粒度图片合成、图片修复、图片渐变、图片属性检索、数据增强等。

4.1　背景介绍

生成对抗网络中出现的模式坍缩问题在实际应用中经常困扰着研究者们。模式坍缩问题是指生成对抗网络训练完成后，其在隐空间中的采样的多个点在输入到生成网络 G 后得到的图片的内容非常相似，此时生成网络生成的图片没有多样性。如图4.1中所示，左边为隐变量空间，在隐变量空间中采样的点 z_1，z_2，z_3 经过生成模型后得到的图像内容相似。

为了解决生成对抗网络中的模式坍缩问题。受到变分自编码器的启发，变分自编码器中编码网络将图片空间映射到隐变量空间中，然后用解码网络将隐变量空间再映射回图片空间。在此过程中，重构损失函数约束重构图片和输入图片一致，所以变分自编码器的生成图片空间和真实图片空间是一致的。所以一个直接的想法是将变分自编码器中的编码器加入到生成对抗网络框架中，同时将变分自编码器中的解码网络和生成对抗网络的生成网络合并来解决生成对抗网络训练中出现的模式坍缩的问题。如图4.2所示，在加入了编码网络后，真实图片空间中的每个点都被映射到隐空间中，然后再由生成网络将隐空间中的点映射回图片空间中。因真实图片空间中的图片是多样的，所以生成的图片空间也是多样性的，这也就解决了模式坍缩的问题。

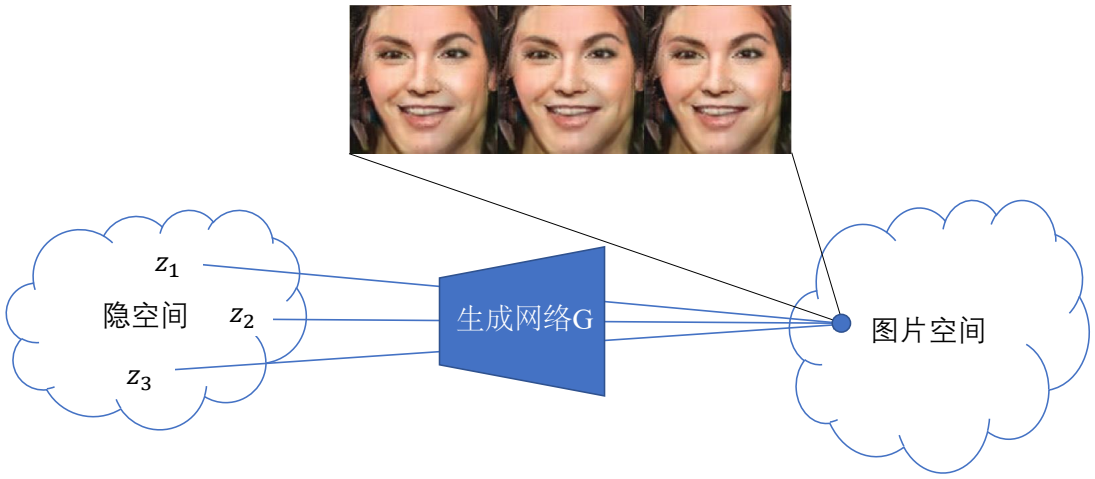


图 4.1　生成对抗网络训练中出现的模式坍缩的问题：左边为隐变量空间，在隐变量空间中采样的点 z_1，z_2，z_2 经过生成模型后得到的图像内容相似。

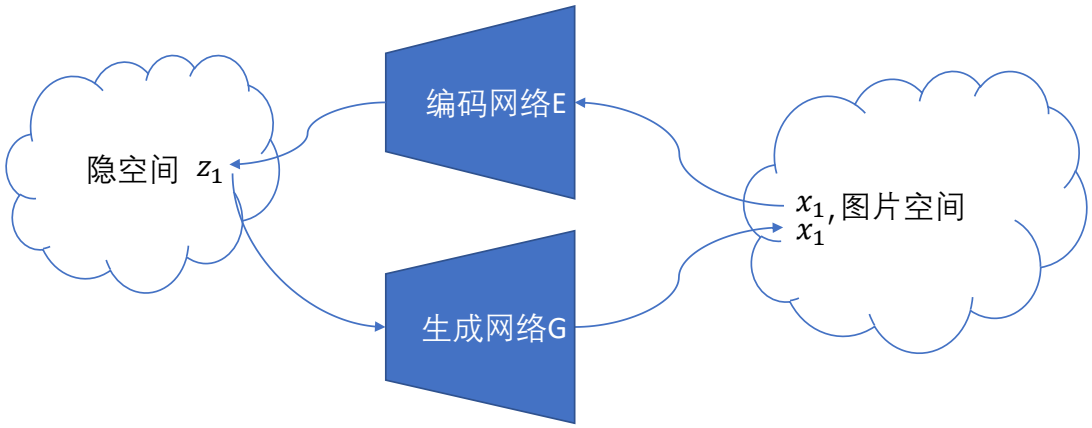


图 4.2　利用编码网络解决模式坍缩的问题，左边为隐变量空间，右边为图片空间，编码网络将图片空间中的点 x_1 映射到隐空间中的 z_1 点，然后生成网络将 z_1 映射回图片空间得到 x_1'，因重构损失函数约束 x_1' 和 x_1 一样，即约束了生成网络生成的图片空间和原图片空间一致，从而解决了模式坍缩的问题。

但是在实际中，这种简单的合并并不能直接解决问题。因为从变分自编码器中生成的图片常常是模糊的，所以生成对抗网络中的判别网络可以很快的学到如何判别“真或假”并收敛。由第 3 章中分析的，这会导致判别网络回传给生成网络的梯度为 0，从而导致生成模型在训练过程中得不到正确的更新。受到第 3 章中特征匹配条件生成对抗网络的启发，本章同样在合并的框架中使用特征匹配的算法，由于有了图片空间到隐空间和隐空间再回到图片空间的映射，该框架建立了一个输入图片 x 和生成图片 x' 的成对信息。所以本章改进了第四章中的特征中心匹配，使之升级为成对的特征匹配，即要求对于每个输入图片和其重构的图片的在判别网络中的特征一致。

本章同样致力于解决有条件输入的图像合成，所以在有了编码网络的生成对抗网络框架中加入了分类网络。有了成对的输入图片和其重构的图片，框架同样在分类网络中要求它们的特征一致，这样使合成模型可以合成更加符合条件的图片。

如图4.3中所示，论文比较了本章提出的条件变分生成对抗网络（CVAE-GAN）框架与现有的框架如：变分自编码器（VAE）[4]、生成对抗网络（GAN）[15]、变分自编码器/生成对抗网络（VAE/GAN）[23]、条件变分自编码器（CVAE）[26]、条件生成对抗网络（CGAN）[27]、即插即用生成网络（PPGN）[105]、第三章提出中的特征匹配条件生成对抗网络（FM-CGAN）的区别与联系。本章提出的框架包含 4 个部分：（1）编码网络 E，（2）生成网络 G；（3）判别网络 D；（4）分类网络 C。本章提出的框架结合了条件变分自编码器和条件生成对抗网络的框架，所以该框架名字被命名为条件变分生成对抗网络（CVAE-GAN）。

本章其余部分组织如下，在4.2小节中将具体介绍条件变分生成对抗网络的框架结构、损失函数与算法流程；4.3小节将介绍条件变分生成对抗网络的实验评估；在4.4小节中将展示条件变分生成对抗网络可以被应用在多个任务中;4.5小节进行小结与讨论。

4.2　条件变分生成对抗网络

4.2.1　基本框架

条件变分生成对抗网络（CVAE-GAN）框架如图4.4所示，主要包括 4 个部分：（1）编码网络 E；（2）生成网络 G；（3）判别网络 D；（4）分类网络 C。

编码网络 E 的作用是将输入图片 x 和其对应标签 y 映射到隐空间得到 $z = E(x, y)$，生成网络 G 的作用是接受输入的隐空间表达 z 和标签 y 得到输出 $x' = G(z, y)$，它尝试学到的真实数据 x 的分布。判别网络 D 对输入的真实数据 x 和生成数据 x' 进行二分类。分类网络 C 使用真实数据 x 的和其标签 y 进行分类，然后用学到的分类特征来约束生成图片 x' 使其满足给定的标签输入。其中涉及的损失函数将在4.2.2章节中进行具体介绍。

4.2.2　损失函数

条件变分生成对抗网络框架使用编码网络 E 得到从图片 x 和其对应标签 y 到隐空间变量 z 的映射，即 $z = E(x, y)$。生成网络 G 的作用是接受输入的隐空间表达 z 和标签 y 得到输出 $x' = G(z, y)$，因此也得到了从隐空间到图片空间的映射。

采用和变分自编码器相同的做法，条件变分生成对抗网络框架使用编码器

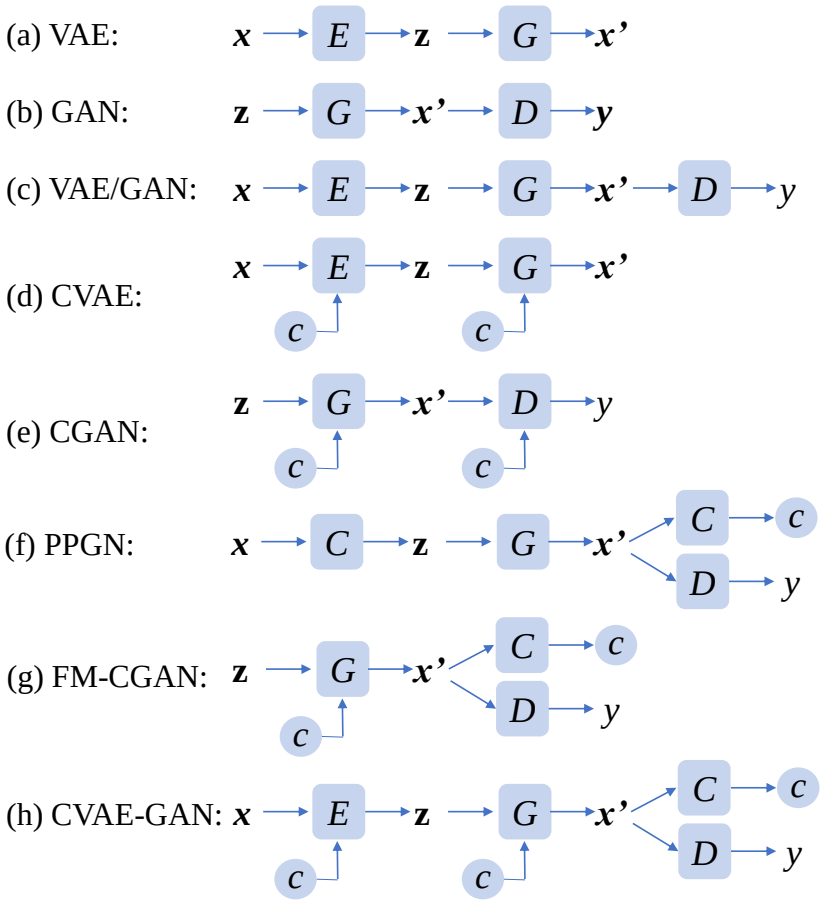


图 4.3　经典的图像合成模型的示意图，包括：变分自编码器（VAE）[4]、生成对抗网络（GAN）[15]、变分自编码器/生成对抗网络（VAE/GAN）[23]、条件变分自编码器（CVAE）[26]、条件生成对抗网络（CGAN）[27]、即插即用生成网络（PPGN）[105]、第三章提出中的特征匹配条件生成对抗网络（FM-CGAN）以及本章提出的条件变分生成对抗网络（CVAE-GAN）。这里 x 和 x' 分别表示输入图像和生成图像；E，G，C，D 分别表示编码网络，生成网络，分类网络和判别网络；z 是隐空间变量；y 是判别网络的输出，表示“真或假”；c 是表示条件，可以是标签也可以是属性。

得到输入图片在隐空间中的均值和方差的估计，然后利用 KL 损失函数去约束隐空间变量 z 的分布与一个先验分布 $P_z \sim N(0, I)$ 的距离：

$$\mathcal{L}_{\text{CVAE-GAN}}(KL) = \frac{1}{2}\left(\mu^T\mu + sum(\exp(\epsilon) - \epsilon - 1)\right), \tag{4.1}$$

其中 μ 是编码网络估计的均值，ϵ 是编码网络估计的方差的对数值 $log(\sigma^2)$。由于直接使用均值和方差无法进行求导，所以这里使用重新参数化的技巧（reparameterization trick）。可以使用公式 $z = \mu + r \odot exp(\epsilon)$ 采样一个 z，其中 r 是一个从 $N(0, I)$ 中随机采样的点，$\odot$ 表示元素位置上的相乘。在得到一个从 x 到 z 的映射之后，生成网络 G 使用 z 生成图片 x'。然后使用像素级别的 ℓ_2 重构损失函数约束 x 和 x'。同时由于 x 和 x' 的成对关系，在判别网络 D，分类网络 C 中可以

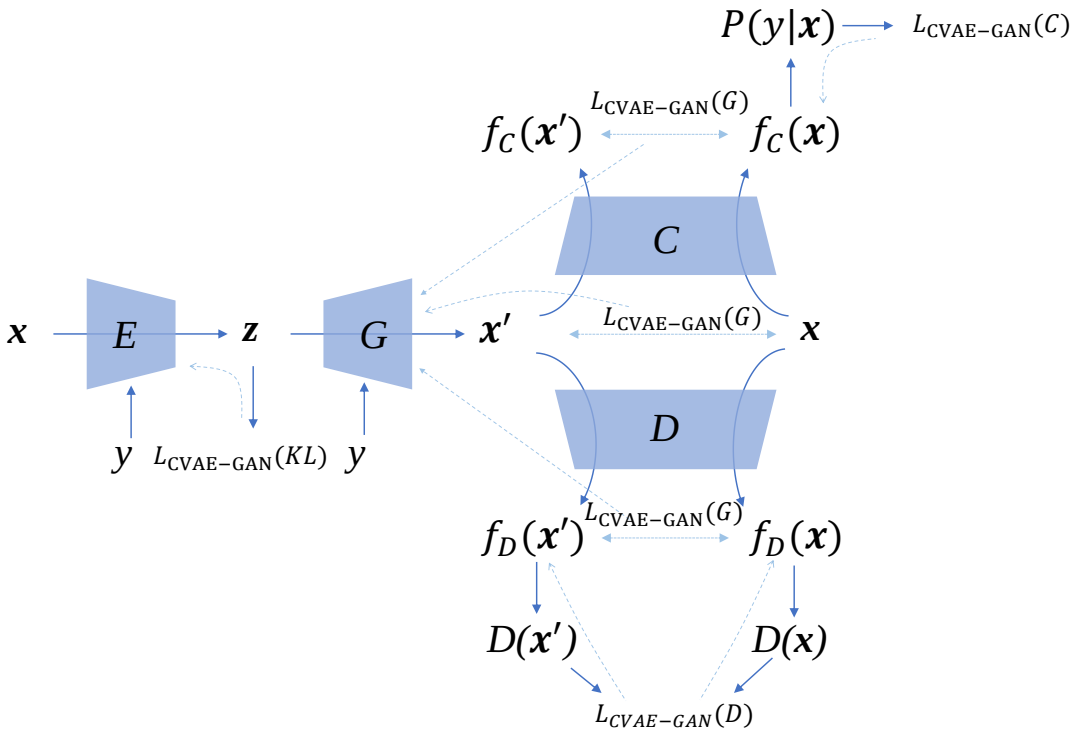


图 4.4　条件变分生成对抗网络（CVAE-GAN）框架示意图。框架包括四个部分：（1）编码网络 E，（2）生成网络 G；（3）分类网络 C；和（4）判别网络 D。实线表示网络前向传播，虚线表示梯度回传的方向。具体细节参见4.2章节。

使用特征匹配损失函数，所以对 x' 的损失函数如下：

$$\begin{aligned}\mathcal{L}_{\text{CVAE-GAN}}(G) = \frac{1}{2}(||x - x'||_2^2 + \xi_1||f_D(x) - f_D(x')||_2^2 \\ + \xi_2||f_C(x) - f_C(x')||_2^2),\end{aligned} \tag{4.2}$$

其中 f_D 和 f_C 分别是判别网络 D 和分类网络 C 中的某一个特征层。ξ_1 和 ξ_2 分别是判别网络和分类网络中使用成对特征匹配损失函数的系数大小，在实验中都被设置为 1。

对于条件变分生成对抗网络中的判别网络 D 而言，其功能和原始 GAN 的相同，其尝试将对生成图片和真实图片进行分类，所以，对于判别网络 D 的损失函数为：

$$\mathcal{L}_{\text{CVAE-GAN}}(D) = -\mathbb{E}_{x\sim P_r}[\log D(x)] - \mathbb{E}_{z\sim P_z}[\log(1 - D(G(z, y))]. \tag{4.3}$$

对于分类网络 C 而言，其和特征匹配条件生成对抗网络中的分类器相同，它使用图片 x（其标签为 y）作为输入，然后输出一个 K 维的向量，然后使用一个 softmax 函数将其转化为属于 K 类的概率。该向量的每一维表示的是其属于某个类概率。在训练阶段，分类网络 C 尝试去最小化 softmax 损失：

$$\mathcal{L}_{\text{CVAE-GAN}}(C) = -\mathbb{E}_{x\sim P_r}[\log P(y|x)], \tag{4.4}$$

其中 P_r 是真实数据的分布，y 表示为每一个输入图片 x 对应的标签。$P(y|x)$ 是分类网络估计输入图片 x 为标签 y 的概率。

所以所有的损失函数为：

$$\mathcal{L}_{\text{CVAE-GAN}} = \mathcal{L}_C + \mathcal{L}_{\text{CVAE-GAN}}(D) + \mathcal{L}_{\text{CVAE-GAN}}(G) + \xi_3\mathcal{L}_{\text{CVAE-GAN}}(KL), \tag{4.5}$$

其中损失函数 $\mathcal{L}_{\text{CVAE-GAN}}(KL)$ 和 $\mathcal{L}_{\text{CVAE-GAN}}(G)$ 的梯度都会回传到编码网络 E 中。所以 ξ_3 为平衡损失函数 $\mathcal{L}_{\text{CVAE-GAN}}(KL)$ 和 $\mathcal{L}_{\text{CVAE-GAN}}(G)$ 使用的权重系数。在实验中设置为 3。

4.2.3　算法流程

在框架训练过程中，条件变分生成对抗网络框架需要先训练分类网络，然后利用分类网络 C 的分类能力保证从标签生成的图片满足给定标签。同时框架需要让生成网络 G 和判别网络 D 进行对抗式的训练，训练过程如算法4.1所示。在训练设置中，本章使用了 Adam[103] 的优化策略，学习率为 0.0002，β_1 为 0.5，β_2 为 0.999，训练中使用的 4 块 K40 卡，Batch Size 设置为 128。

4.3　实验评估

本实验中同样使用的是三个数据集，FaceScrub [99]，102 Category Flower [100], 和 CUB-200 [101] 来验证本文提出的条件变分生成对抗网络框架的实用性。这三个数据集包含三种完全不同的物体，分别是人脸，花和鸟类。对数据的处理和实验设置均和第三章中的实验设置相同。

在实验中，框架的网络结构有编码网路 E，生成网络 G，判别网络 D 和分类网络 C。其中生成网络 G，判别网络 D 和分类网络 C 均使用了和第三章中特征匹配生成对抗网络中相同的结构，其中编码网络论文使用了一个 GoogleNet [91] 结构，其中将输入换成 128×128 分辨率的。所以最后一层的平均池化的大小由原来的 7×7 换为 4×4，同时将原来的单个输出，改为输出两个输出，两个分别输出估计的均值 μ 和方差的对数值 $log(\sigma^2)$。

4.3.1　指定标签的图像合成

在这一章节中，论文展示条件变分生成对抗网络框架可以用于指定标签的图像合成，比如指定一个人，一种花或者是一种鸟类。在条件变分生成对抗网络

算法 4.1　条件变分生成对抗网络（CVAE-GAN）的训练算法。

Data: 训练的 Batch Size 为 m，所有的类的个数为 K，θ_E，θ_G、θ_D，θ_C 分别为编码网络 E，生成网络 G，判别网络 D 和分类网络 C 的参数；ξ_1 和 ξ_2 分别是判别网络 D 和分类网络 C 中的特征匹配损失函数的权重，实验中均设为 1；ξ_3 为 KL 损失函数的权重，在实验中设为 3；P_z 是假设隐变量空间分布 $N(0, I)$。

while 生成网络 G 没有收敛 **do**

从 $P_r(x)$ 中采样 m 个真实数据样本 $\{x_1, ..., x_{(m)}\}$; 它们的标签为 $\{y_1, ..., y_m\}$。

$\mathcal{L}_{\text{CVAE-GAN}}(C) \leftarrow -\frac{1}{m}\sum_{i=1}^{m}\log(P(y_i|x_i))$.

通过编码网络 E 得到生成数据 z_i: $z_i = E(x_i, y_i)$。

$\mathcal{L}_{\text{CVAE-GAN}}(KL) \leftarrow \frac{1}{m}\sum_{i=1}^{m} KL(z_i)||P_z)$.

通过生成网络 G 得到生成数据 x_i': $x_i' = G(z_i, y_i)$。

$\mathcal{L}_{\text{CVAE-GAN}}(D) \leftarrow -\frac{1}{m}\sum_{i=1}^{m}[\log(D(x_i)) + \log(1 - D(x_i'))]$.

$\mathcal{L}_{\text{CVAE-GAN}}(G) \leftarrow$
$\frac{1}{2}\frac{1}{m}\sum_{i=1}^{m}(||x_i - x_i'||_2^2 + \xi_1||f_D(x_i) - f_D(x_i')||_2^2 + \xi_2||f_C(x_i) - f_C(x_i')||_2^2)$.

使用梯度下降法更新所有网络的参数:

$\theta_C \xleftarrow{+} -\nabla_{\theta_C}(\mathcal{L}_{\text{CVAE-GAN}}(C))$.

$\theta_D \xleftarrow{+} -\nabla_{\theta_D}(\mathcal{L}_{\text{CVAE-GAN}}(D))$.

$\theta_E \xleftarrow{+} -\nabla_{\theta_E}(\xi_3\mathcal{L}_{\text{CVAE-GAN}}(KL) + \mathcal{L}_{\text{CVAE-GAN}}(G))$.

$\theta_G \xleftarrow{+} -\nabla_{\theta_G}(\mathcal{L}_{\text{CVAE-GAN}}(G))$.

end

框架中，只需要指定标签信息作为条件输入，将条件输入到生成模型 G 中即可完成指定标签的图像合成。如图4.5所示，为条件变分生成对抗网络在指定的标签上合成的图片（每组图像下方为标签）。可以看到合成的图片真实且在一个类中呈现多样性。这展示了论文提出的条件变分生成对抗网络的模型的实用性。

4.3.2　损失函数的消融实验

在条件变分生成对抗网络中，在更新生成网络 G 时使用了判别网络 D 和分类网络 C 中成对的特征匹配损失函数和图像像素层面的重构损失函数。为了解释这其中每一个损失函数的作用，论文将损失函数 $\mathcal{L}_{\text{CVAE-GAN}}(G)$ 分成三个部分，分别是 $\mathcal{L}_G(img)$, $\mathcal{L}_G(D)$ 和 $\mathcal{L}_G(C)$。$\mathcal{L}_G(img) = \frac{1}{2}(||x - x'||)$ 表示图像像素层面的损失函数。$\mathcal{L}_G(D) = \frac{1}{2}\xi_1||f_D(x) - f_D(x')||_2^2$ 表示判别网络 D 中的特征匹配损失函

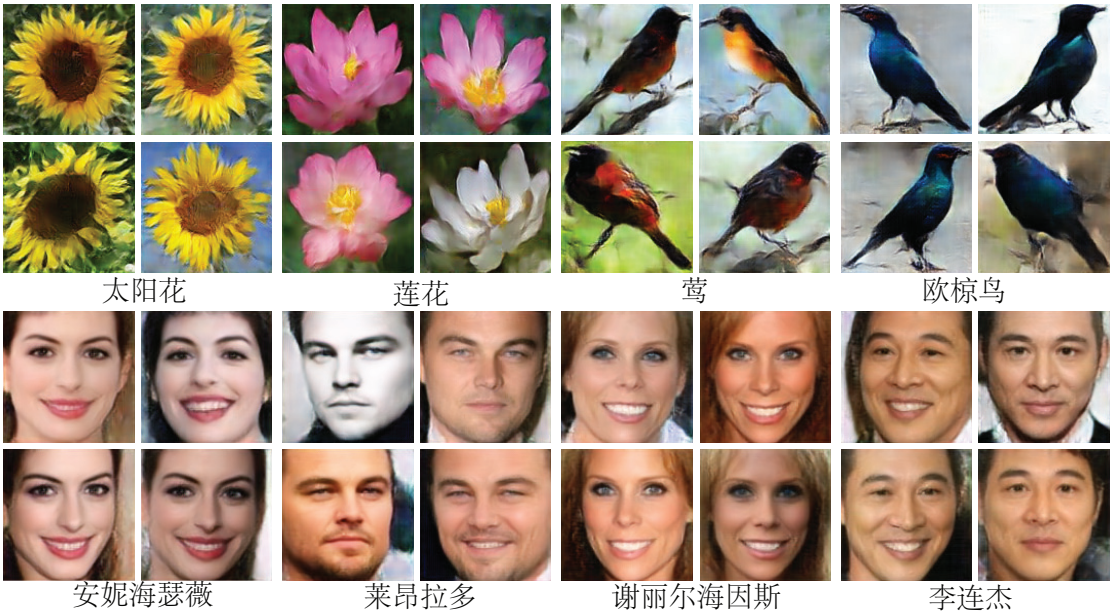


图 4.5　论文提出的条件变分生成对抗网络（CVAE-GAN）可以完成指定标签的图像合成。每组图像均是使用下面的标签作为输入条件合成的。可以看到，合成的图片真实且在一个类中呈现多样性。

数，$\mathcal{L}_G(C) = \frac{1}{2}\xi_2||f_C(x) - f_C(x')||_2^2$ 表示分类网络 C 中的特征匹配损失函数。

实验中训练了四个不同设置的条件变分生成对抗网络：第一个使用条件变分生成对抗网络中的所有损失函数；第二个使用除了 $\mathcal{L}_G(img)$ 之外的损失函数（表示为 w/o $\mathcal{L}_G(img)$）；第三个使用除了 $\mathcal{L}_G(D)$ 之外的损失函数（表示为 w/o $\mathcal{L}_G(D)$）；第四使用除了 $\mathcal{L}_G(C)$ 之外的损失函数（表示为 w/o $\mathcal{L}_G(C)$）。

图4.6中比较了四种不同损失函数组得到得到的实验结果。观察结果可以知道，移除判别网络 D 中损失函数 $\mathcal{L}_G(D)$ 使得生成图片变得更加模糊；移除分类网络 C 中损失函数 $\mathcal{L}_G(C)$ 使得生成图片无法保持输入的身份信息；移除像素层面的损失函数 $\mathcal{L}_G(img)$ 使得生成图片丢失了很多细节。而在另外一个方面，使用了所有损失函数的条件变分生成对抗网络生成了真实且保持身份信息的人脸图片。

4.3.3　与其他方法的合成质量比较

论文在三个数据集 FaceScrub [99]，102 Category Flower [100]，和 CUB-200 [101] 比较生成图片的结果。在各个数据集中均随机选择了一个标签 y_i，然后从隐变量空间分布 $N(0, I)$ 中采样很多个隐变量，然后一起输入到生成网络 G 中得到合成图片，如图3.6所示为不同方法生成图片的视觉结果比较。其中（a）为某一类的真实数据；（b）为第三章中提出的特征匹配条件生成对抗网络生成的图片结果；（c）为条件变分生成对抗网络生成的图片结果。

比较条件变分生成对抗网络与特征匹配条件生成对抗网络的合成图片结果，

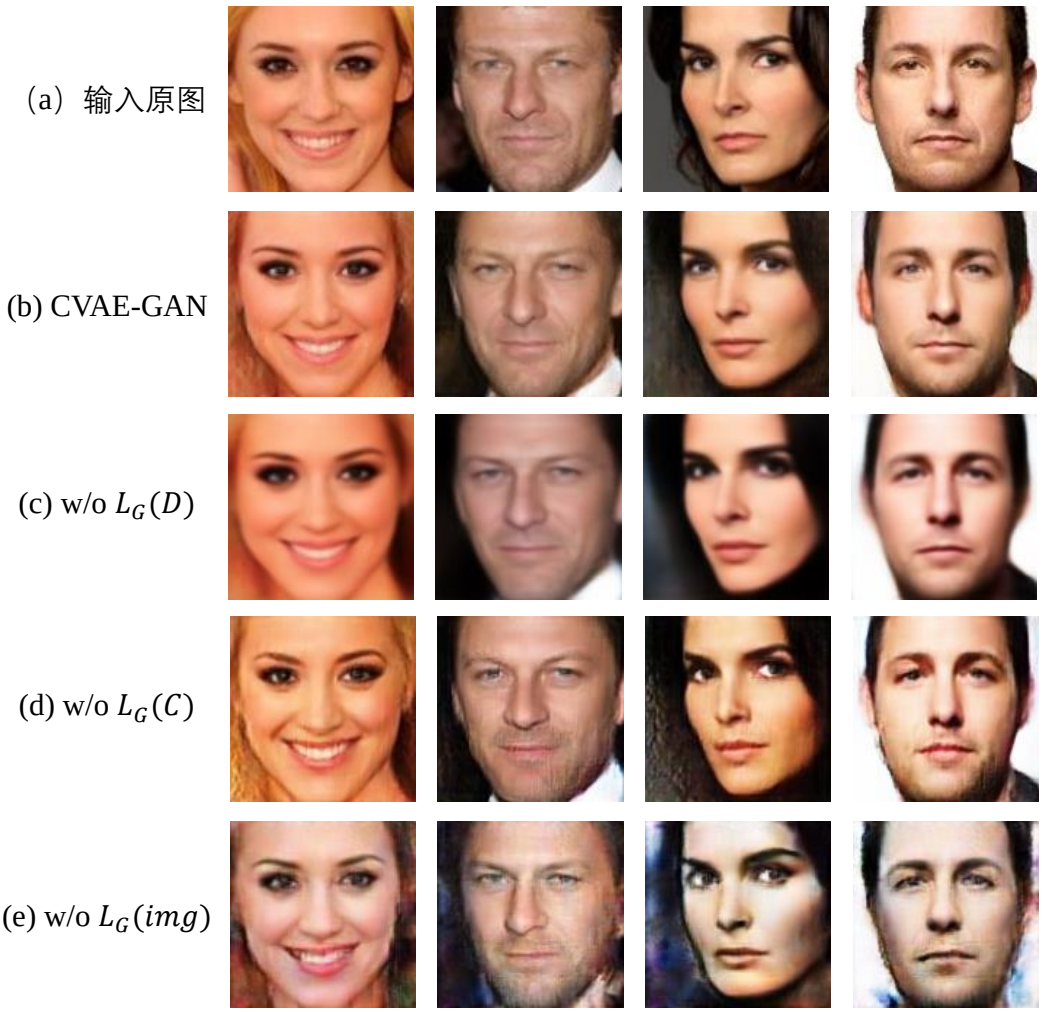


图 4.6　使用不同的损失函数得到生成网络 G 的重构结果比较。（a）为输入，（b）为条件变分生成对抗网络（CVAE-GAN）重构的结果，（c）为移除判别网络 D 中损失函数的结果，（d）为移除分类网络 C 中损失函数的结果，（e）为移除像素上的损失函数的结果。

可以发现条件变分生成对抗网络比特征匹配条件生成对抗网络合成的图片更加清晰，同时类中的多样性也更大。这展示了条件变分生成对抗网络模型合成图片的优势。

4.3.4　与其他方法的数值比较

与第三章中的衡量标准相同，论文尝试在三个尺度上衡量合成模型：与输入条件一致性，多样性，真实性。论文使用 Facescrub 人脸数据做这个数值比较的实验，首先，论文分别用特征匹配条件生成对抗网络和条件变分生成对抗网络随机的生成 53000 张图片（每个类 100 张）。然后对于生成的图片

计算 top-1 准确率和 Inception score 的方式和第三章中相同，如表格4.1所示，条件变分生成对抗网络和特征匹配条件生成对抗网络的 Inception Score 得分差

图 4.7　不同条件生成模型在 Facescrub, 102 Category Flower, 和 CUB-200 数据集上生成的图片结果比较。（a）为某一类的真实数据；（b）为第三章中提出的特征匹配条件生成对抗网络生成的图片结果；（c）为条件变分生成对抗网络生成的图片结果。

不多，但是在保持输入条件信息这一指标中条件变分生成对抗网络取得了更好的结果。

4.3.5　隐空间变量分析

在条件变分生成对抗网络中，生成网络 G 可以使用隐空间变量 z 和标签重构出图片。输入相同的标签，不同的隐空间变量给合成模型会得到照片有相同的标签但是有不同的属性。所以一个问题自然产生：对于不同的标签输入，取相同的隐空间变量 z 会得到怎样的结果?

论文发现对于生成网络 G 而言，相同的隐空间变量代表了合成图片具有相

表 4.1　真实图片、特征匹配条件生成对抗网络合成图片和条件变分生成对抗网络合成图片质量的数值结果比较。相比于特征匹配条件生成对抗网络，条件变分生成对抗网络模型合成的图片更好地保持了输入的标签信息，同时具有更好的真实性和多样性。

模型	top-1 准确率	Inception score
真实数据	99.61%	20.85
FM-CGAN	79.76%	19.40
CVAE-GAN	97.78%	19.03

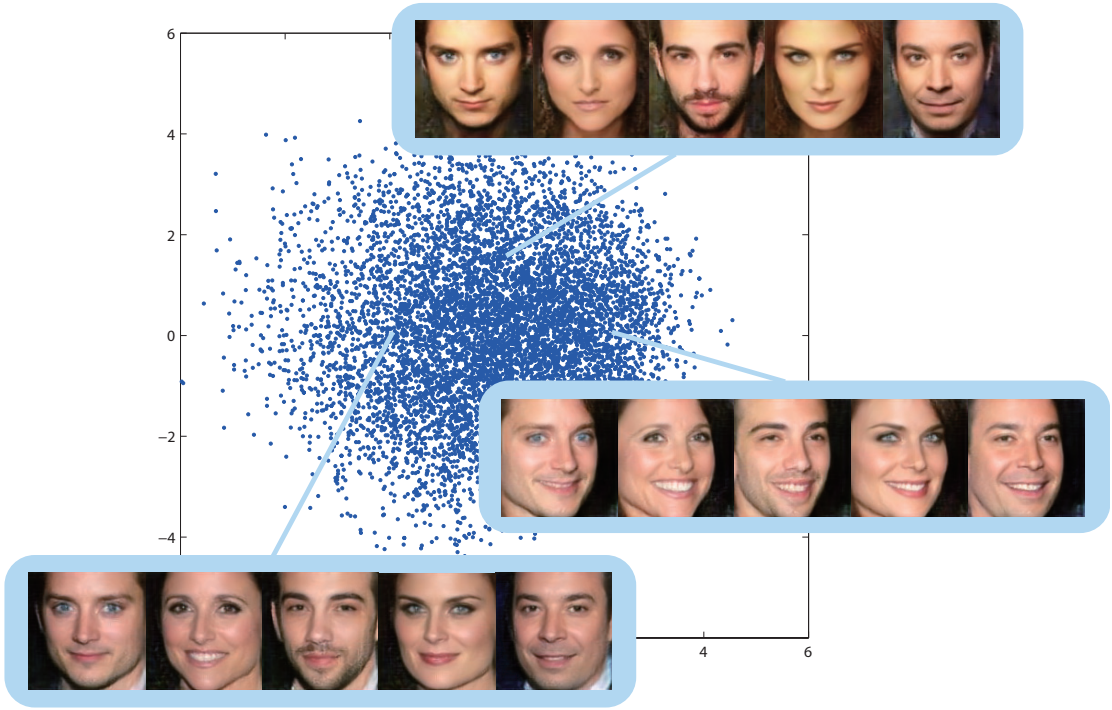
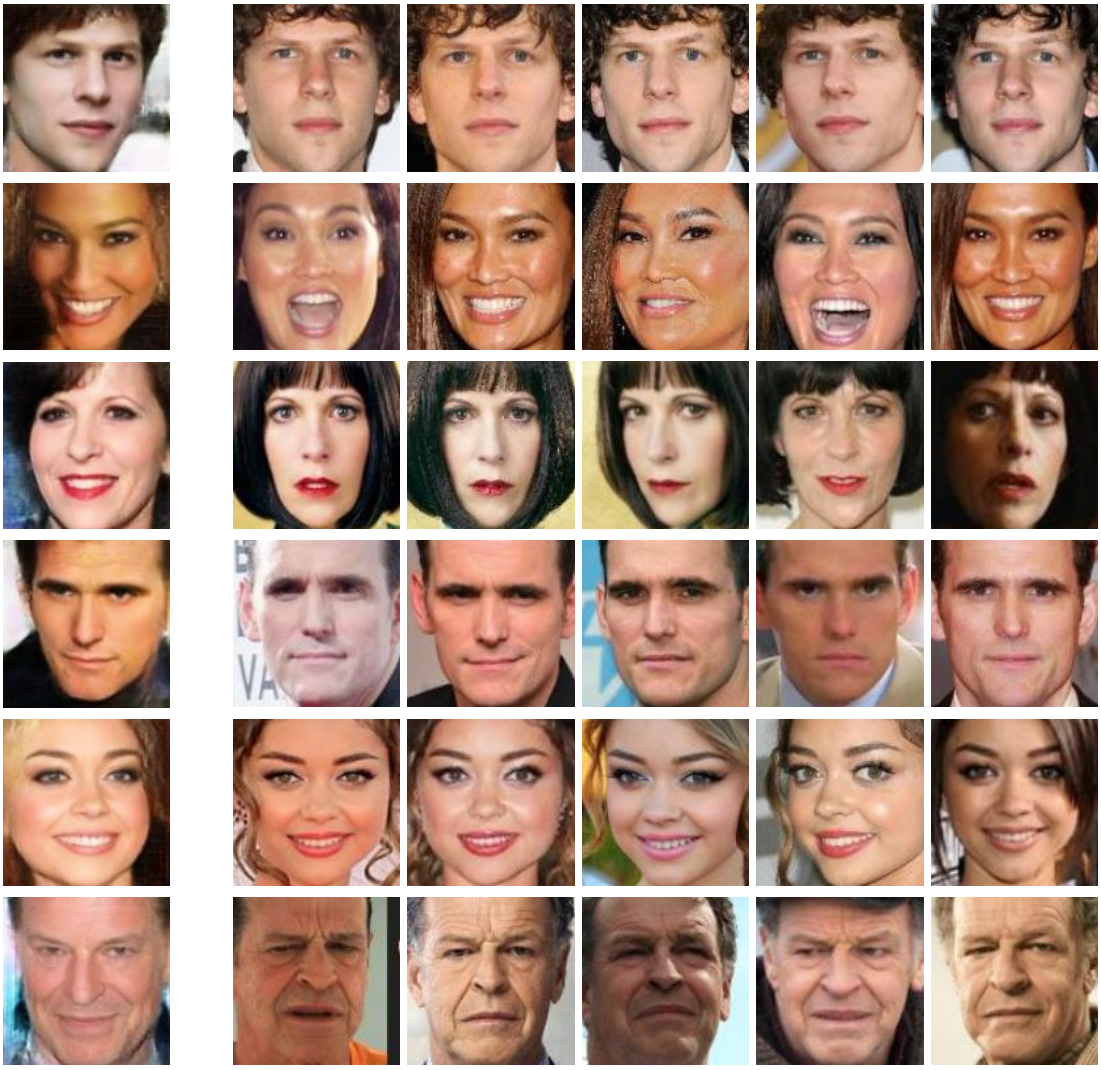


图 4.8　隐空间变量的分布示意图，图中的人脸图片是使用不同的标签和相同的隐空间变量生成的结果，可以注意到，相同的隐空间变量生成了相同属性的人脸图片，属性包括表情，背景，脸的角度等等。

同的属性（角度、光照、背景等）。论文用 Facescrub 数据集进行实验，对于一个已经训练完成条件变分生成对抗网络模型，使用编码网络得到所有真实图片的隐空间变量表达，然后使用主成分分析法（PCA）[106] 将隐空间变量的维数降为 2 维，然后在一个平面上画出所有真实图片的隐空间变量的分布（如图4.8中的蓝点所示）。再然后将相同的隐空间变量和不同的标签输入到生成网络 G 中合成图片，合成的图片如图4.8中的淡蓝色框中所示。可以看到相同的隐空间变量生成了相同属性的人脸图片，属性包括表情，背景，脸的角度等等。

从编码网络 E 的角度考虑，隐空间变量代表了合成图片的属性信息。也就意味着编码网络 E 将图片中的属性信息编码到隐空间变量中，编码网络有了提取属性信息这样的性质，所以可以用编码器完成相同属性图片的检索。论文将



(a) 合成图片　　(a) 在真实图片中前五个最近邻搜索

图 4.9　条件变分生成对抗网络合成图片在训练集图片中最近邻搜索结果，(a) 中为条件变分生成对抗网络的合成图片，(b) 中为在训练集图片中搜索结果的前五名。可以看出条件变分生成对抗网络可以生成真实图片中不存在的图片。

在4.4.3中作具体的介绍。

4.3.6　生成图片的最近邻搜索

在本章节中，论文将验证本章提出的条件变分生成对抗网络模型不是仅仅在训练过程中“记住”了所有的真实图片。它可以生成真实图片分布以外的图片。论文在人脸数据集 Facescrub 中进行验证。首先使用生成模型 G 随机生成 6 张图片，然后在真实数据中进行图片检索，检索结果如图4.9所示，(a) 中为合成图片，(b) 为在训练集图片中检索到的真实图片的前五名。可以观察到条件变分生成对抗网络框架合的成图片和检索的结果并不相同，也就表示条件变分生成对抗网络模型在图片合成过程中不是仅仅“记住”了真实数据，其有一定的泛

化能力。

4.4　条件变分生成对抗网络的应用

在这一小节中，论文将展示本章提出的条件变分生成对抗网络模型可以被应用在很多任务中，例如：图片修复，图片渐变，图片的属性检索，数据增强等等。

4.4.1　图片修复

论文在 FaceScrub [99], 102 Category Flower [100], 和 CUB-200 [101] 三个数据集上进行图片修复的实验。首先在 128×128 分辨率的原图中（图4.10(a)），破坏一块 50×50 的区域（图4.10(b)）。然后将待修复的图片 x 输入进编码网络 E 中得到其在隐空间中的表达 z，然后可以通过生成网络 G 合成图像 $x' = G(z, y)$，其中 y 表示待修复图片 x 的标签。然后根据图像修复的方法得到修复后的图像 x_r：

$$x_r = \boldsymbol{M} \odot x' + (1 - \boldsymbol{M}) \odot x, \tag{4.6}$$

其中 $\boldsymbol{M}$ 是一个 0 或者 1 的二元矩阵，指示原图中哪一块待修复。$\odot$ 表示元素位置上的相乘。所以 $(1 - \boldsymbol{M}) \odot x$ 表示原图中不用修复的区域，$\boldsymbol{M} \odot x'$ 表示用条件变分生成对抗网络修复的区域。

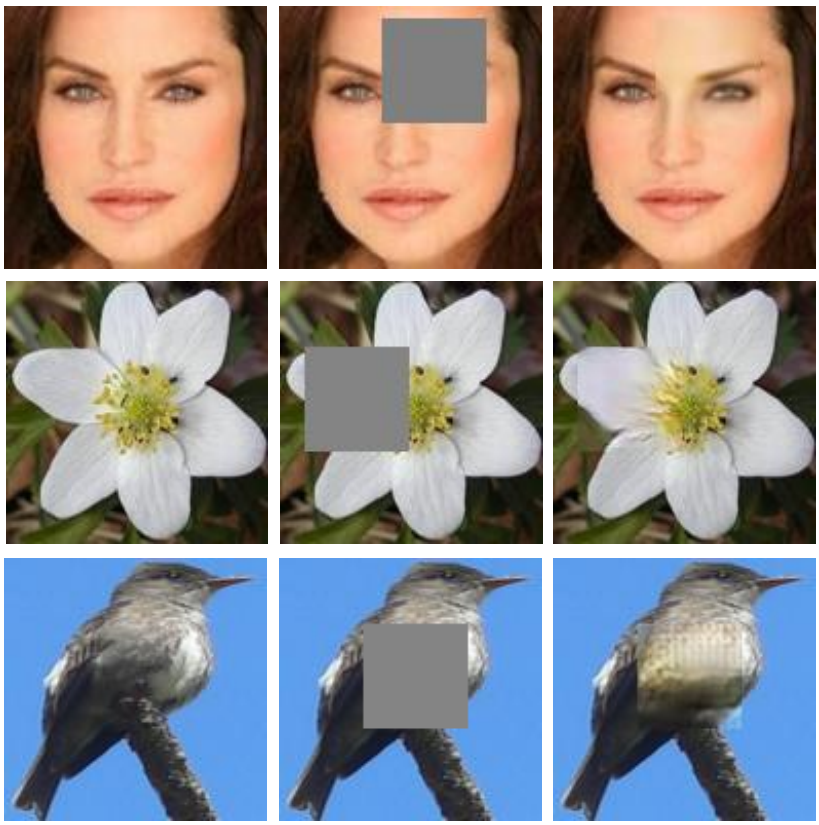
图像修复的结果如图4.10所示，可以看到本章提出的条件变分生成对抗网络模型可以得到不错修复结果。

4.4.2　图片渐变

在本节中，论文展示条件变分生成对抗网络模型可以被用在图片渐变这个任务中，论文在 FaceScrub [99], 102 Category Flower [100], 和 CUB-200 [101] 三个数据集上进行图片渐变的实验。首先在每个数据集中选择一对在相同标签中的图片 x_1 和 x_2，然后使用编码网络 E 提取器隐空间的表达 z_1 和 z_2。这样之后可以使用线性插值的方式得到一系列的隐空间变量 z。线性插值的表达式为：

$$z = \alpha z_1 + (1 - \alpha) z_2, \alpha \in [0, 1]. \tag{4.7}$$

最后将得到的线性插值的隐空间变量 z 和其标签输入生成网络 G 中得到其图片渐变的结果。如图4.11所示为图片渐变的结果。可以看到，图片中的属性例如：人脸的角度，表情，背景；花的数量与颜色；鸟的角度和颜色均会发生渐变。同时在渐变的过程中生成图片的标签信息还是被保持的很好。



a) 原图　　b) 待修复图片　　c) 修复结果

图 4.10　条件变分生成对抗网络应用在图像修复的任务中。（a）为原图，（b）为待修复图片，（c）是修复的结果。

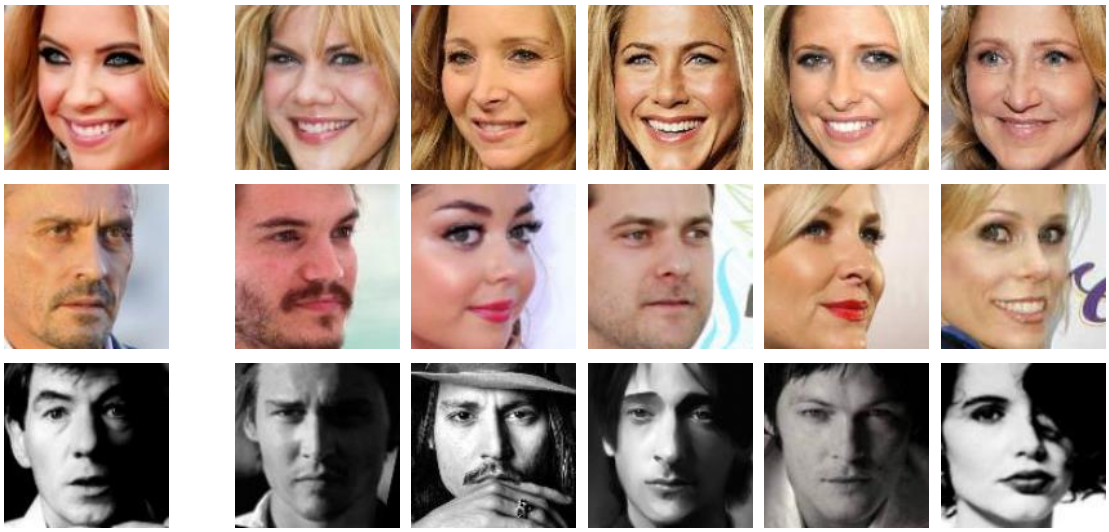
（a）原图x_1　　（b）渐变结果　　（c）原图x_2

图 4.11　条件变分生成对抗网络应用在图片渐变任务中。（a）为原图 x_1，（c）为原图 x_2，（b）为在 x_1 与 x_2 中图像渐变的结果。

4.4.3　图片属性检索

如4.3.5节中介绍的，可以得到一个推论：相似的隐空间变量代表了图片空间中相似的属性。因此，可以使用编码网络去提取图片中的属性特征，然后利用编

码网络 E 实现相似属性特征图片的搜索。论文使用人脸数据集 Facescrub 进行这个实验，首先用编码网络 E 对所有真实的图片提取了属性特征，然后选取其中的一些图片作为检索图片，用其属性特征在所有的图片的属性特征中进行检索。论文使用 ℓ_2 距离进行排序，如图4.12所示，为相同属性图片检索的结果。从结果可以看出，利用编码器 E 提取的特征检索到的结果都为属性特别相似的图片。



a) 检索图片　　b) 检索结果的前五名

图 4.12　条件变分生成对抗网络应用在图片属性相似的检索任务中。(a) 为检索图片，(b) 为图像属性相似的检索结果。

4.4.4　数据增强

生成对抗网络合成的图片究竟能不能用于训练数据呢? 在本节中，论文进一步展示本章提出的条件变分生成对抗网络模型生成的数据可以被用作是数据增强的一种方法。论文使用 Facescrub 数据集作为训练数据进行训练，测试数据使用 LFW[107]。

论文在实验中使用了两种数据增强的策略：第一种是对于训练数据集中已有的标签生成更多的图片作为训练数据；第二种是使用已有标签线性插值出新的标签，然后在新的标签中生成图片数据作为增加的训练数据。论文测试了两种数据增强的策略。对于第一种情况。对于每一个已经存在的标签随机生成 200 张人脸图片，总共接近 10 万张图片。对于第二种情况，论文使用不同标签线性插值的方法得到了新的 5 千个标签，然后对于每个标签生成 100 张图片，总共 50 万张图片。在两种情况中，论文均将新生成的图片和 Facescrub 中的原图一起训练人脸识别模型。

在测试阶段，论文直接使用特征的余弦距离去衡量两张脸的身份相似性。在 LFW 上做了 10 个部分的交叉验证实验。在表4.2中，论文比较了不同数据增强方

法与只用原始数据的实验比较结果。使用已有标签的数据增强可以提升分类模型的准确率，而使用 5 千新标签的数据增强可以进一步提升分类模型的准确率。

表 4.2　数据增强的结果。

方法	训练数据量	准确率
无数据增强	8 万	91.87%
已有标签的数据增强	8 万 +10 万	92.77%
5 千新标签的数据增强	8 万 +50 万	92.98%

4.5　小结与讨论

在本章中，论文提出了条件变分生成对抗网络框架，该框架将编码网络加入到生成对抗网络的训练中，编码网络将图片空间映射到隐空间，再使用生成网络将隐空间映射回图片空间，因为原图片空间的分布中的图片是多样的，所以生成网络生成的图片也是多样的。这样解决了生成对抗网络中的模式坍缩问题。实验结果表明，加入了编码网络的生成对抗网络框架生成了更加富有多样性的图片，从而证明该框架有效解决了模式坍缩问题。同时该框架可以完成很多应用：细粒度图片合成、图片修复、图片渐变、图片属性检索、数据增强等。

在第 3 章和本章中，论文分别提出特征匹配条件生成对抗网络框架和条件变分生成对抗网络框架完成标签作为条件输入的图片合成。特征匹配条件生成对抗网络框架使用特征中心匹配损失函数解决了生成对抗网络中的训练不稳定的问题，条件变分生成对抗网络框架改进特征中心匹配损失函数为成对特征匹配损失函数，进一步提升生成对抗网络的训练稳定性，同时编码网络的加入解决模式坍缩问题。

第 5 章　基于身份保持的生成对抗网络的人脸图像合成

在第 3 章和第 4 章中，论文提出的条件合成框架可以满足训练数据集中既有标签的图片合成。但是为了完成数据集中不存在的标签的图片合成，例如对于人脸图片的合成任务，人们的需求是合成任意指定身份的人脸图片，这时特征匹配条件生成对抗网络（FM-CGAN）和条件变分生成对抗网络（CVAE-GAN）均无法完成。

为了解决这个问题，本章提出了身份保持的生成对抗网络（IP-GAN）框架以满足面向开放集身份保持的人脸图片合成。该框架可以解耦人脸图片中的身份特征和属性特征（角度、表情、光照等），然后重组该身份特征和从另外一张人脸图片提取的属性特征得到一张新的人脸图片。该人脸图片满足给定的身份特征，同时也满足给定的属性特征。实验结果表明，该框架实现了开放集中的身份保持的人脸图片合成。同时该框架可以应用在很多任务中：侧脸图片转正脸图片、人脸识别中的对抗样本检测、人脸图片属性转换等。

5.1　背景介绍

近来，随着生成对抗网络的发展，人脸图片的合成逐渐变为一个非常热门的研究方向，最近的工作 DR-GAN[108]、FF-GAN[109] 尝试合成一张输入人脸图片不同角度的图片，它们仅仅可以完成对人脸图片的角度的修改。另外一方面，本文在第三章中提出的特征匹配条件生成对抗网络框架和第四章中提出的条件变分生成对抗网络框架只能完成训练数据集中已有身份的人脸图片合成。针对这些工作的局限性，本章希望可以完成以下目标：对于任意给定身份的人脸图片，生成模型可以合成很多不同属性的人脸图片，同时在该过程中合成图片的身份保持不变。

为了达到这个目标，论文首先要求框架可以解耦人脸图片中的身份特征和属性特征（角度、表情、光照等），然后重组该身份特征和从另外一张人脸图片提取的属性特征得到一张新的人脸图片。如图5.1所示，可以从 A 和 B 两张人脸图片中分别解耦出身份特征和属性特征，然后重新组合得到新的人脸图片 A’ 和 B’。A’ 具有 A 的身份特征和 B 的属性特征。B’ 具有 B 的身份特征和 A 的属性特征。

因此，本章设计了图5.2所示的基于生成对抗网络的框架，该框架由五个部分构成：（1）身份提取网络 I，其主要功能是从输入图片中得到身份特征。（2）属性提取网络 A，其主要功能是从输入图片中得到属性特征。（3）生成网络 G，

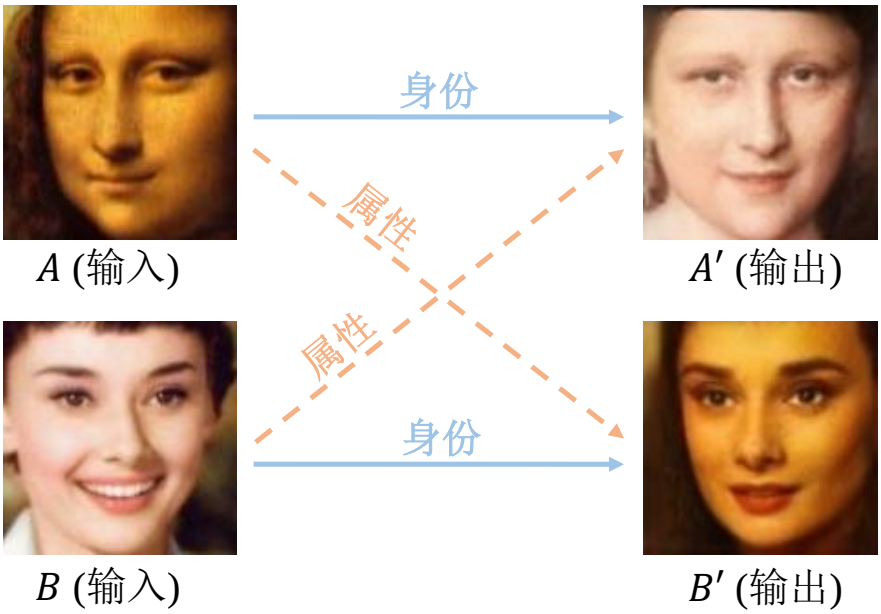


图 5.1　本章提出身份保持的生成对抗网络可以从图片中解耦出身份特征和属性特征，然后重新组合这些特征合成新的图片。图中，模型从输入 A 和 B 中分别提取出身份特征和属性特征，然后重组这些特征输入到生成网络 G 中合成新的图片 A' 和 B'，可以看到 A' 具有 A 的身份特征但是是 B 的属性特征。B' 具有 B 的身份特征但是 A 的属性特征。

其主要功能是结合身份特征和属性特征合成新的人脸图片。(4) 分类网络 C，主要是用来保证合成图片的身份与输入身份图片一致。(5) 判别网络 D，用来判别输入图片的“真或假”，同时约束生成图片尽可能为“真”。由于该框架可以完成身份保持（Identity Preserving）的人脸图片合成，故框架被命名为身份保持的生成对抗网络（IP-GAN）。

为了使身份提取网络 I 提取身份特征，论文使用了最近的人脸识别技术。利用海量图片数据和深度卷积神经网络模型的优势，训练出好的人脸识别模型作为身份提取网络 I。但是，提取人脸图片中属性信息非常困难，因为目前没有大量的含有属性标注的人脸图片，所以无法使用有监督（supervised）的方法得到属性特征。论文在框架中采用了一个简单的方法解决了这一问题，即使用网络的损失函数“迫使”属性提取网络 A 学习提取图片中的属性特征。考虑一种情况，当输入的身份特征和属性特征均来自一张人脸图片，这时框架虚虚要重构出人脸图片，由于仅仅依靠身份特征无法重构原图，所以需要属性提取网络 A 提取属性信息以恢复出原图。但是这里又有一个问题，就是生成网络 G 可以忽略身份特征直接用属性特征进行重构，所以框架对属性提取网络使用了 KL 损失函数，使得属性特征中不含任何身份特征。这样属性提取网络通过一个无监督的方式学到如何从人脸图片中提取属性特征。

同时为了解决生成对抗网络中出现的训练不稳问题，论文使用了第三章中

提出的特征匹配损失函数。论文将用实验证明本章提出的身份保持的生成对抗网络框架的实用性，另外论文将展示本章提出的身份保持的生成对抗网络可以被用在各种应用中。

本章其余部分组织如下，在5.2小节中将具体介绍身份保持的生成对抗网络框架结构、损失函数与算法流程；在5.3小节中介绍关于身份保持的生成对抗网络框架结构的实验分析；5.4小节中将展示身份保持的生成对抗网络框架结构可以被应用在多个任务中；5.5小节进行小结与讨论。

5.2　身份保持的生成对抗网络

5.2.1　框架结构

在本节中，论文将重点介绍本章提出的身份保持的生成对抗网络（Identity Preserving GAN）的框架结构，为了实现对于任意给定身份的人脸图片合成，框架需要两张输入图片，一张是身份图片 x^s 用来提取身份特征，一张是属性图片 x^a 用来提取属性特征（例如：角度、表情、光照、甚至是背景）。框架合成一张一张新的图片具有 x^s 的身份特征同时具有 x^a 的属性特征。

如图5.2所示, 本章提出的框架基于生成对抗网络。它包含五个部分，（1）身份提取网络 I，它负责从身份图片 x^s 中提取身份特征 $f_I(x^s)$。（2）属性提取网络 A，它负责从属性图片 x^a 中提取属性特征 $f_A(x^a)$。（3）生成网络 G，它负责将特征 $[f_I(x^s)^T, f_A(x^a)^T]^T$ 转换成图片。（4）分类网络 C，它只存在于训练阶段，在训练中主要是用来保证合成图片的身份信息与输入身份图片一致。（5）判别网络 D，它只存在于训练阶段，在训练中用来判别图片的“真假”，并使生成模型尽可能合成真实的图片。

在测试阶段，只需要三个网络，分别是身份提取网络 I，属性提取网络 A 和生成网络 G。这样可以对于给定任意两张人脸图片合成另外一张图片。如图5.1所示，A 图片经过身份提取网络 I 得到身份特征 $f_I(A)$，B 图片经过属性提取网络 $f_A(B)$。然后生成网络 G 使用 $f_I(A)$ 和 $f_A(B)$ 合成新的图片 A’。通过一样的方法可以合成 B’。

5.2.2　解耦身份特征和属性特征

框架中最核心的部分就是怎样从图片中解耦出身份特征和属性特征，在本节中，论文将做具体的介绍。在人脸图片训练数据集中，一般只有人脸图片的身份标注，并没有属性相关的标注。这是因为身份的标注通常的容易获得的。如很多大的数据集例如 FaceScrub [99], CASIA-WebFace [104] 和 MS-Celeb-1M [110] 数据集都有身份的标注。然而，属性特征的标注通常都是非常困难的，甚至有一

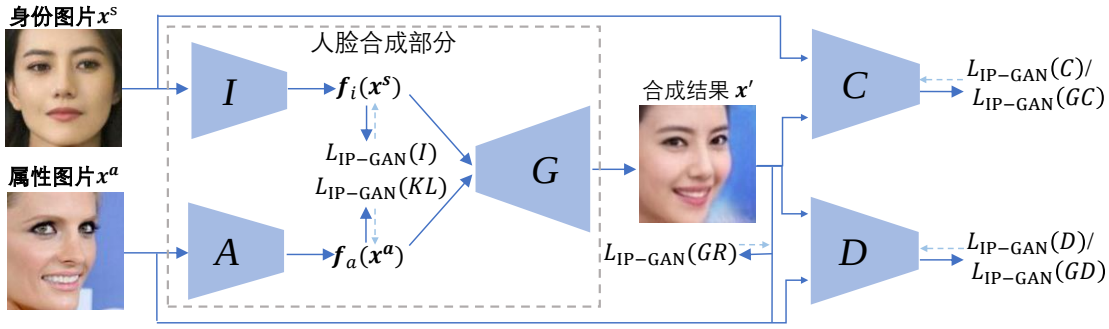


图 5.2　身份保持的人脸图片合成（IP-GAN）框架示意图。

些属性例如光照，背景是不能标注的。

从人脸图片中提取身份特征非常简单，论文充分利用了现在人脸识别技术的进步与发展 [111-112]。给定一个带有身份标注的人脸图片的集合 $\{x_i^s, c_i\}$，身份提取网络 I 完成一个人脸分类的任务，这样可以用 softmax 损失函数来训练属性提取网络 I。这样，相同身份的人脸图片输入进身份提取网络 I 会有近乎一样的身份特征。身份提取网络 I 的损失函数为：

$$\mathcal{L}_{\text{IP-GAN}}(I) = -\mathbb{E}_{x^s \sim P_r}[\log P(c|x^s)], \tag{5.1}$$

其中，$P(c|x^s)$ 表示人脸图片 x^s 的身份是 c 的概率。然后论文选取身份提取网络 I 中的最后一个池化层的特征作为身份特征。

为了使用无监督的方法使得属性提取网络 A 提取属性特征，论文提出了一个新的简单的方法给每一个人脸提取属性特征。论文使用了两个损失函数，一个是重构损失函数，另一个是 KL 散度损失函数。

重构损失函数考虑两种情况，一种是输入的身份图片 x^s 与属性图片 x^a 相同，另一种身份图片 x^s 和属性图片 x^a 不同。在两种情况中，框架都要求生成结果 x' 去重构属性图片 x^a。但是两种情况中的重构损失函数的权重不同。即损失函数为：

$$\mathcal{L}_{\text{IP-GAN}}(GR) = \begin{cases} \frac{1}{2}||x^a - x'||_2^2 & \text{if } x^s = x^a \\ \frac{\lambda}{2}||x^a - x'||_2^2 & \text{其它} \end{cases}, \tag{5.2}$$

其中 λ 是当身份图片 x^s 与属性图片 x^a 不同时的重构损失函数的系数。接下来论文将分析两种情况中的重构损失函数。

当输入给框架的身份图片 x^s 与属性图片 x^a 相同时，输出图片 x' 必须是和输入身份图片 x^s 或者属性图片 x^a 相同。所以此时的损失函数是：$\frac{1}{2}||x^a - x'||_2^2$。这时重构函数是怎么帮助属性提取网络提取属性的呢？假设某一身份有很多不同属性特征的人脸图片，则这些图片输入到身份提取网络 I 得到的身份特征 $f_I(x)$

几乎相同。如果将这些图片中的每一张输入到框架中得到输出图片。因为这些图片的身份特征相同，所以输出图片的不同必然来自于属性提取网络 A 提取的特征的不同，所以框架会“迫使”属性提取网络 A 学到提取属性特征帮助框架重构图片。

当输入给框架的身份图片 x^s 与属性图片 x^a 不相同时，框架无合成人脸图片的参考。但是框架应该约束合成的图片的属性应该和输入的属性图片相同。因此论文采用了小权重的像素级别的损失函数：$\frac{\lambda}{2}||x^a - x'||_2^2$，去约束合成图片中的一些属性和输入属性图片相同。实验中设置 $\lambda = 0.1$。论文在实验中也比较了不同的 λ 设置对实验结果的影响，将在实验部分详细分析。

KL 散度损失函数为了让属性提取网络 A 更好地学到如何提取属特征，论文同时也使用了 KL 损失函数去约束属性特征分布和一个先验概率分布 $P_z \sim N(o, I)$ 接近。KL 散度损失函数会限制属性特征的分布范围，以使得其不包含身份特征。对于每一个属性图片，属性提取网络 A 输出均值 μ 和方差的对数值 $\epsilon = \log(\sigma^2)$。KL 散度损失函数如下式所示：

$$\mathcal{L}_{\text{IP-GAN}}(KL) = \frac{1}{2}(\mu^T\mu + \sum_{j-1}^{J}(\exp(\epsilon) - \epsilon - 1)), \tag{5.3}$$

其中，j 表示向量 ϵ 中的第 j 个元素。和变分自编码器 [4] 的做法相似，在训练中使用公式 $z = \mu + r \odot exp(\epsilon)$ 采样一个属性值。r 是一个从 $N(0, I)$ 中随机采样的点，$\odot$ 表示元素位置上的相乘。

5.2.3　特征匹配损失函数

在从身份提取网络 I 中提取了身份特征 $f_I(x^s)$ 和从属性提取网络 A 中提取属性特征 $f_A(x^a)$ 之后，身份保持的生成对抗网络框架将两个特征连接在一起得到 $z = [f_I(x^s)^T, f_A(x^a)^T]^T$，并将其输入到生成网络 G 中合成一张人脸图片 x'。这一节中，论文将介绍特征匹配损失函数。它即帮助生成网络 G 做到身份保持的人脸合成，同时又使框架的训练过程变得更加稳定。

和生成对抗网络中中的判别网络 D 相同，身份保持的生成对抗网络中的判别网络同样是尝试将对生成图片和真实图片进行分类，所以，对于判别网络 D 的损失函数为：

$$\mathcal{L}_{\text{IP-GAN}}(D) = -\mathbb{E}_{x\sim P_r}[\log D(x)] - \mathbb{E}_{z\sim P_z}[\log(1 - D(G(z))]. \tag{5.4}$$

如同之前特征匹配条件生成对抗网络中分析，如果对于生成网络 G 直接使用损失函数 $\mathbb{E}_{z\sim P_z}[\log(1 - D(G(z))]$ 优化，那么会导致判别网络回传给生成网络 G 的梯度消失的问题。所以论文采用了和条件变分生成对抗网络中相同的做法，即

在更新生成网络 G 时使用判别网络 D 中的特征匹配损失函数回传的梯度。该损失函数如下式所示：

$$\mathcal{L}_{\text{IP-GAN}}(GD) = \frac{1}{2}||f_D(x') - f_D(x^a)||_2^2. \tag{5.5}$$

在实验中，论文选择了判别网络 D 中的最后卷积层的的输入作为特征 f_D。

同时，分类网络 C 和身份提取网络 I 功能相同，都是在进行人脸分类的的任务。所以它们的损失函数相同，如下式所示：

$$\mathcal{L}_{\text{IP-GAN}}(C) = -\mathbb{E}_{x^s \sim P_r}[\log P(c|x^s)]. \tag{5.6}$$

为了使生成网络 G 完成身份保持的人脸合成，论文借助于人脸识别网络优秀的分类能力。身份保持的生成对抗网络框架要求生成图片 x' 和输入身份图片 x^s 在判别网络 C 中有相同的特征，这样保证了合成的图片具有 x^s 的身份。假设 f_C 是分类网络 C 中某一层的特征。那么分类网络 C 中的特征匹配损失函数即为下式：

$$\mathcal{L}_{\text{IP-GAN}}(GC) = \frac{1}{2}||f_C(x') - f_C(x^s)||_2^2. \tag{5.7}$$

其中，论文使用分类网络 C 中的最后一层全连接层的输入作为特征层 f_C。在实验中，身份提取网络 I 和分类网络 C 共享参数。同时为了加速训练的速度，身份提取网络 I 和分类网络 C 直接使用已经训练好的人脸识别网络。

表 5.1　各网络与它们相关的损失函数。

网络	损失函数
I	$\mathcal{L}_{\text{IP-GAN}}(I)$
A	$\mathcal{L}_{\text{IP-GAN}}(I)$, $\mathcal{L}_{\text{IP-GAN}}(GR)$
G	$\mathcal{L}_{\text{IP-GAN}}(GR)$, $\mathcal{L}_{\text{IP-GAN}}(GC)$, $\mathcal{L}_{\text{IP-GAN}}(GD)$
D	$\mathcal{L}_{\text{IP-GAN}}(D)$
C	$\mathcal{L}_{\text{IP-GAN}}(C)$

最终身份保持的生成对抗网络框架的损失函数是以上介绍的公式5.1 - 公式5.7的总和。尽管这有很多的损失函数，但是如表格5.1所示，每个网络和其相关的损失函数都不多。因此总体而言，身份保持的生成对抗网络框架是非常容易训练的。

5.2.4　无监督的训练方法

生成训练数据集中不存在身份的人脸图片具有非常大的挑战。它要求生成网络 G 可以生成任意身份的图片，即在合成人脸图片过程中可以捕捉所有人的人脸身份特征。先存公开的有标注身份的数据集又很少的包括各种极限条件（例如大角度、夸张的表情、复杂的背景）的人脸图片。或者换句话说，这些数据集的多样性不足。

为了解决数据多样性不足的问题，论文从谷歌和 Flicker 中又收集了 1 百万张图片。然后使用人脸检测工具检测到人脸区域。这些图片有着比现有数据集更大的多样性，所以论文将这些数据加入到训练数据集中。这些数据是没有标签数据的，所以论文叫这个方法无监督的训练方法。

这些没有标注的图片既可以被当做是身份图片，也可以被当做属性图片参与训练。这些图片的使用增加了生成网络的能力，使其可以生成背景更复杂，表情更夸张的图片。

5.2.5　算法流程

本节介绍身份保持的生成对抗网络框架的具体算法流程。在训练中，框架先使用 softmax 损失函数将身份提取网络 I 和分类网络 C 训练好，然后再训练其他的网络。因为训练中存在身份图片和属性图片相同和不相同两种情况，所以实验中使用两种情况交替训练的方法。第一步先用身份图片和属性图片相同情况进行训练，第二步用身份图片和属性图片不同情况的训练，这样交替训练。具体步骤如算法5.1所示。

5.3　实验分析

本小节将注重介绍关于身份保持的生成对抗网络框架中的实验分析。首先论文将介绍实验的具体设置，然后将对框架进行消融实验分析，再然后将对重构函数中的 λ 进行分析，最后将分析 KL 损失函数。

5.3.1　实验设置

论文使用 MS-Celeb-1M[110] 数据集作为训练数据集，经过清理后的 MS-Celeb-1M 数据集有大约 8 万个人，大约 500 万人脸图片。对于每一张图，论文使用 JDA[102] 人脸检测算法检测其中的人脸区域。然后将人脸对齐和缩放到一个固定的位置上。

对于身份提取网络 I，属性提取网络 A 和分类网络 C，身份保持的生成对抗网络框架均使用了 VGG-16[97] 的结构，其中不同的是 VGG-16 的输入图片为

算法 5.1　身份保持的生成对抗网络（IP-GAN）的训练算法。

Data: θ_A，θ_G、θ_D 分别为属性提取网络 A，生成网络 G，判别网络 D 的参数；$iter \leftarrow 1$

while 生成网络 G 没有收敛 **do**
- 从 $P_r(x)$ 中采样真实数据作为身份图片 x^s。
- **if** $iter\ \%\ 2 = 1$ **then**
  - $x^a \leftarrow x^s$.
  - $\lambda = 1$.
- **else**
  - 从 $P_r(x)$ 中采样真实数据作为属性图片 x^a。
  - $\lambda = 0.1$.
- **end**
- $f_I(x^s) \leftarrow I(x^s)$; $f_A(x^a) \leftarrow A(x^a)$.
- $\mathcal{L}_{\text{IP-GAN}}(KL) \leftarrow KL(f_A(x^a)||P(z))$.
- $x' \leftarrow G([f_I(x^s)^T, f_A(x^a)^T]^T)$.
- $\mathcal{L}_{\text{IP-GAN}}(D) \leftarrow -[\log(D(x^a)) + \log(1 - D(x'))]$.
- $\mathcal{L}_{\text{IP-GAN}}(GR) \leftarrow \frac{1}{2}||x^a - x'_i||_2^2$.
- $\mathcal{L}_{\text{IP-GAN}}(GD) \leftarrow \frac{1}{2}||f_D(x_i) - f_D(x'_i)||_2^2$.
- $\mathcal{L}_{\text{IP-GAN}}(GC) \leftarrow \frac{1}{2}||f_C(x_i) - f_C(x'_i)||_2^2)$.
- 使用梯度下降法更新所有网络的参数：
- $\theta_D \xleftarrow{+} -\nabla_{\theta_D}(\mathcal{L}_{\text{IP-GAN}}(D))$.
- $\theta_E \xleftarrow{+} -\nabla_{\theta_E}(\mathcal{L}_{\text{IP-GAN}}(KL) + \mathcal{L}_{\text{IP-GAN}}(GR))$.
- $\theta_G \xleftarrow{+} -\nabla_{\theta_G}(\mathcal{L}_{IPGAN}(GR) + \mathcal{L}_{\text{IP-GAN}}(GD) + \mathcal{L}_{\text{IP-GAN}}(GC))$.
- iter ← iter + 1.

end

224×224 分辨率。而合成图片的分辨率为 128×128。所以将 VGG-16 中最后一个全连接层的输入大小由原来的 $512 \times 7 \times 7$ 变为 $512 \times 4 \times 4$，同时将全连接层由原来的多层改为一层。对于判别网络 D，身份保持的生成对抗网络框架使用了和 DCGAN 中 D 相同的结构。对于生成网络 G，其使用了一个“倒置”的 VGG-16 结构。所有网络中均使用 le 批归一化层（Batch Normalization）和激活函数层 ReLU 层。

5.3.2　框架的消融实验

在本节中，论文通过损失函数和训练策略的消融实验来理解身份保持的生成对抗网络框架中的损失函数和训练策略是如何帮助框架的。

论文比较了五种身份保持的生成对抗网络框架的变化：（1）训练中不使用损失函数 $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GD})$（表示为 w/o $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GD})$）；（2）训练中不使用损失函数 $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GC})$（表示为 w/o $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GC})$）；（3）训练中不使用身份图片和属性图片不同的情况（表示为 w/o T）；（4）训练中不使用无监督的训练数据（表示为 w/o U）；（5）本章提出的 IP-GAN 框架，其使用了所有的损失函数和训练技巧。不同设置中的网络结构和训练设置均相同。

图5.3中展示了这五种设置的身份保持的生成对抗网络训练得到的模型合成人脸图片质量比较。（a）为输入的身份图片；（b）为输入的属性图片；（c）不同训练设置的方法；（d）为不同训练设置的身份保持的生成对抗网络框架合成图片的结果。通过比较结果，可以观察到训练中不使用损失函数 $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GD})$（表示为 w/o $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GD})$）会使得合成的人脸图片变模糊。训练中不使用损失函数 $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GC})$（表示为 w/o $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GC})$）会使生成的图片无法保持输入身份图片的身份特征。训练中不使用身份图片和属性图片不同的情况（表示为 w/o T）会使合成的人脸图片不能保证输入的属性图片的属性特征，特别是表情特征。相比于训练中不使用无监督的训练数据（表示为 w/o U）的情况，完整的身份保持的生成对抗网络框架合成的图片的属性特征保持的更好，例如嘴巴张开的更大等等属性。

为了可以在数值上衡量不同训练设置的身份保持的生成对抗网络框架的结果。论文进行了人脸图片检索的实验来比较不同设置得到的生成模型的结果。对于数据集中已经存在的人脸图片合成，论文从 MS-Celeb-1M 的测试数据集中随机选择了 1 万个人进行测试，从每个人中随机选择 6 张图片。每个人放 1 张图片在人脸库（face gallery）中，其余 5 张图片作为查询图片。然后对于每个人的查询图片，身份保持的生成对抗网络框架将其作为身份图片和其他随机选择的 5 张图片作为属性图片输入到生成网络 G 中合成新的图片，对于每种训练设置可以得到 50,000 张图片进行图片检索实验，同时也将这些合成图片也作为查询图片。然后论文利用查询图片在人脸库中检索 top-1 准确率作为衡量的指标。

如表格5.2所示，为真实图片和五种不同的训练设置中生成模型合成图片在人脸库中的 top-1 检索准确率，该指标反映了合成图片过程中模型身份保持信息的能力。通过比较结果可以知道：损失函数 $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GC})$ 对人脸图片合成中保持身份特征有着非常重要的作用，同时其他的如损失函数 $\mathcal{L}_{\text{IP-GAN}}(\boldsymbol{GD})$，训练策略均对人脸图片合成中身份保持特征有着作用。

对于训练数据集中不存在的人脸图片合成，论文选用了 Multi-PIE[113] 数据集测试。和上面一样的做法，对于数据集中的每个身份随机选择了 6 张图片，选择其中的一张构建 Multi-PIE 人脸库，其余的五张作为查询人脸图片。对于各张查询图片，用上面一样的方法合成 5 张图片并作为查询图片。表格5.3为开放数据集中各种不同训练设置的身份保持的生成对抗网络合成图片保持身份信息能力的比较。从结果可以看到，身份保持的生成对抗网络中的损失函数 $\mathcal{L}_{\text{IP-GAN}}(GC)$、$\mathcal{L}_{\text{IP-GAN}}(GD)$ 和其训练策略均对人脸图片合成过程中的身份保持能力有帮助。

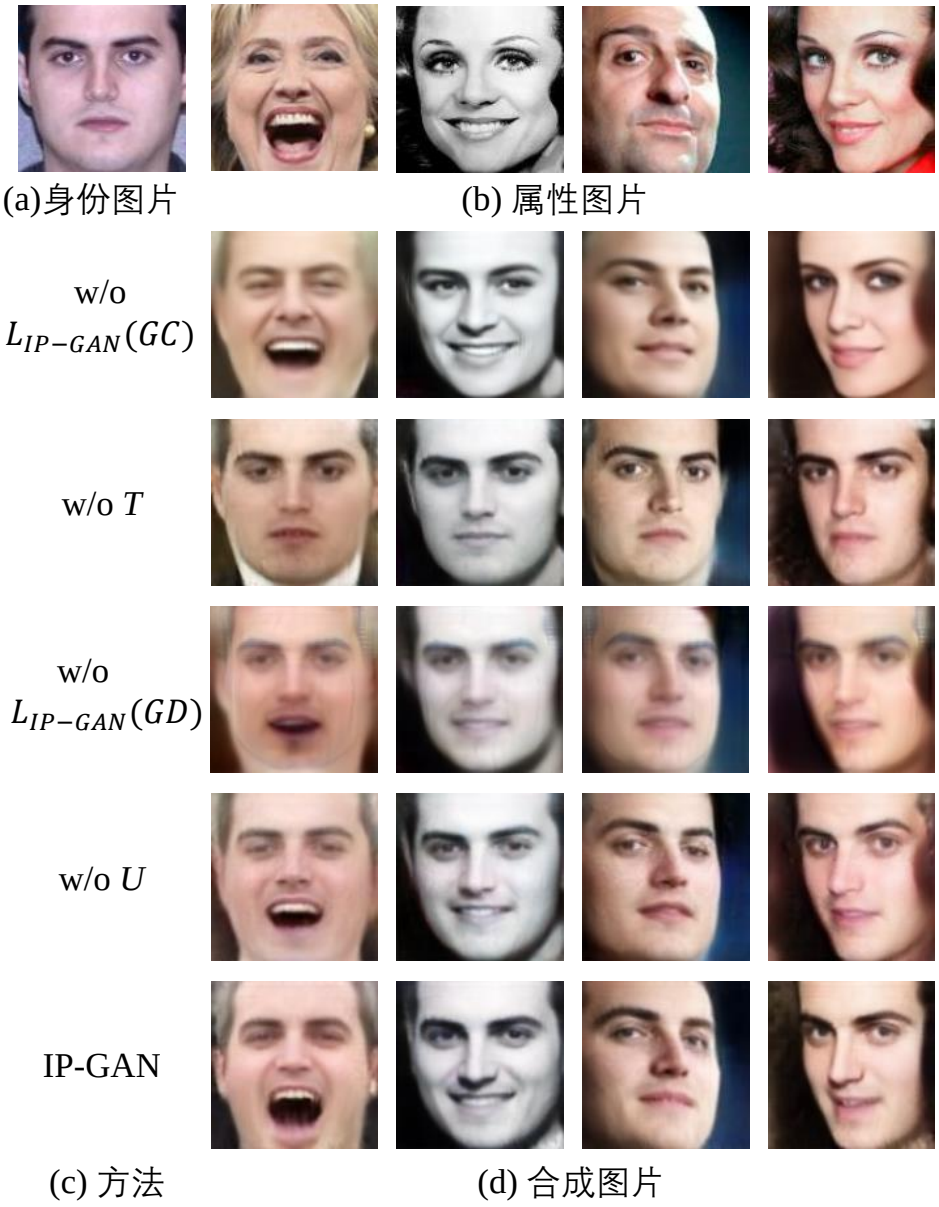


图 5.3　使用不同训练设置的身份保持的生成对抗网络框架合成人脸图片质量比较。(a) 为输入的身份图片；(b) 为输入的属性图片；(c) 不同训练设置的方法；(d) 为不同训练设置的身份保持的生成对抗网络框架合成图片的结果。

表 5.2　不同训练设置的身份保持的生成对抗网络框架在 MS-Celeb-1M 数据集上合成人脸图片 top-1 检索准确率比较。

图片来源	top-1 检索准确率
真实图片	87.39%
w/o $\mathcal{L}_{\text{IP-GAN}}(GC)$	5.71%
w/o T	79.19%
w/o $\mathcal{L}_{\text{IP-GAN}}(GD)$	80.52%
w/o U	80.24%
IP-GAN	**81.11%**

表 5.3　不同训练设置的身份保持的生成对抗网络框架在 Multi-PIE 数据集上合成人脸图片 top-1 检索准确率比较。

图片来源	top-1 检索准确率
真实图片	97.47%
w/o $\mathcal{L}_{\text{IP-GAN}}(GC)$	11.76%
w/o T	95.47%
w/o $\mathcal{L}_{\text{IP-GAN}}(GD)$	95.53%
w/o U	96.41%
IP-GAN	**96.80%**

5.3.3　重构损失函数的分析

在重构损失函数5.2中，对当身份图片和属性图片不同时给重构损失函数加了一个权重 λ。权重 λ 影响了生成图片保持输入图片中属性特征的能力，为了具体研究权重 λ 对合成结果的影响。论文进行实验验证。

在实验中，论文使用了完全一样的框架结构，只是在重构损失函数 $\mathcal{L}_{\text{IP-GAN}}(GR)$ 中使用不同的 λ，实验中具体设置了三组不同的 λ 值，分别是 0.01，0.1，和 1。然后分别训练得到了三个模型。如图5.4所示，为不同模型合成图片结果。(a) 为输入身份图片，(b) 为输入属性图片，(d) 为不同的 λ 值下的生成网络合成的图片结果。比较结果可以知道，当 $\lambda = 0.01$ 时，合成图片丢失了很多属性信息，例如嘴巴的张开与闭合，背景等。当 $\lambda = 1$ 时，合成的图片有很多不真实的细节（artifacts）。在另外一方面，当 $\lambda = 0.1$ 时，合成图片的结果更加真实同时也更好地保持了属性图片的属性。因此，如果 λ 设置的偏小的话，会导致合成图片丢失来自属性图片的属性特征。如果 λ 设置的偏大的话，会导致合成图片有很多不真实的细节。所以，论文在实验中选择了 $\lambda = 0.1$ 的情况，这样保证

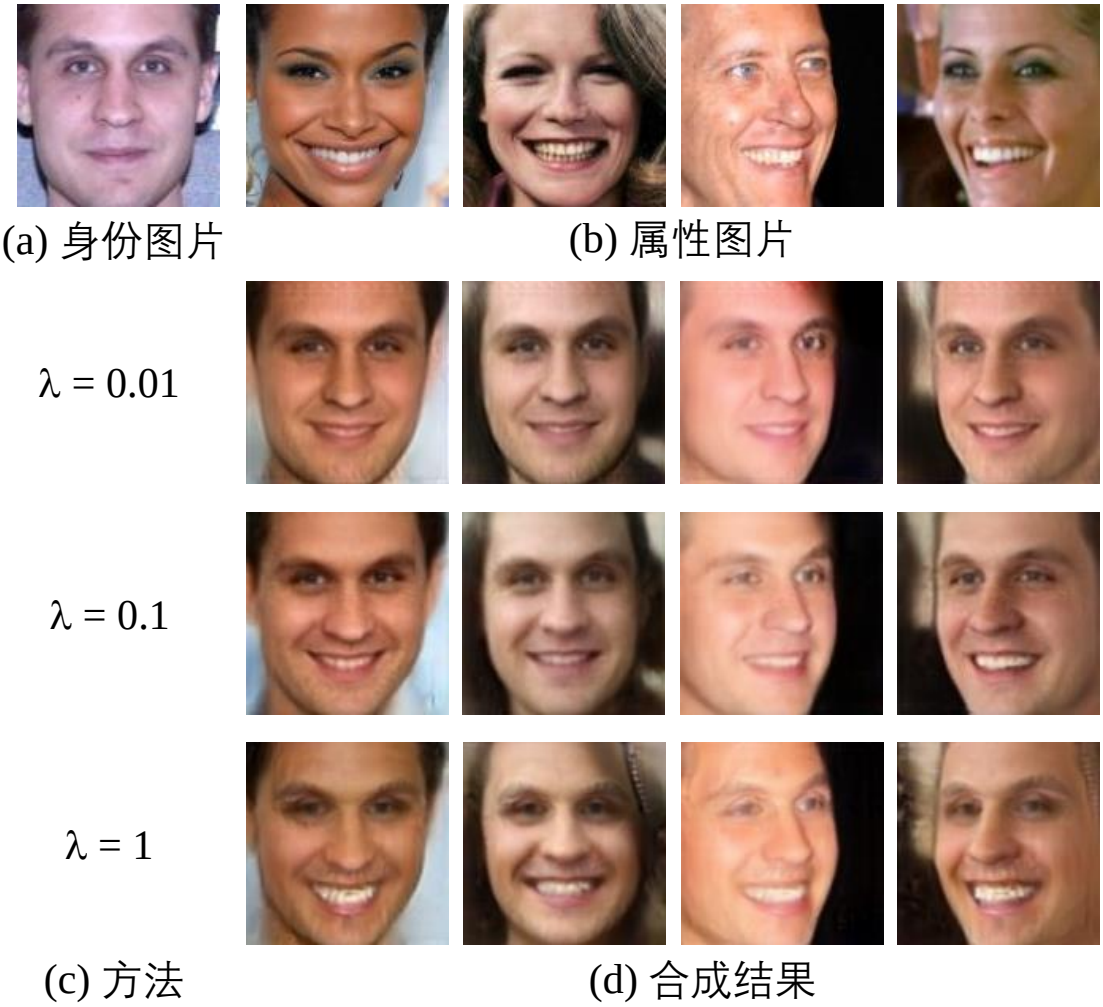


图 5.4　重构损失函数中使用不同的 λ 值的合成结果比较。(a) 为输入身份图片，(b) 为输入属性图片，(c) 为设置的不同的 λ 值，(d) 为使用不同的 λ 值训练出的生成网络合成的图片结果。

了输入属性图片的属性特征又尽量保证合成图片的真实性。

5.3.4　KL 损失函数的作用

在本节中，论文将验证 KL 散度损失函数是不是对属性提取网络 A 中消除身份特征有帮助。选用 MS-celeb-1M 数据集作为训练数据，论文训练了两个框架：一个框架使用了 KL 散度损失函数，一个框架未使用 KL 散度损失函数。然后用 Facescrub[99] 数据集作为测试数据集，论文随机地将数据集分成两个部分，一部分图片作为训练数据，一部分作为验证数据。在两个框架中，均使用属性提取网络 A 来提取 Facescrub 中所有数据的属性特征。然后用 Facescrub 中的训练数据得到的属性特征作为训练集，论文使用训练集数据的特征和训练集数据的特征的类别标签训练一个分类网络。该分类网络是一个简单的多层感知机，其中全连接层的特征维数分别为 1024，1024, 1024，和 530。然后论文使用 Facescrub

中的验证数据得到的属性特征作为验证集，论文使用 top-1 的准确率衡量结果。

如图5.5所示，(a) 为在训练集中训练的分类网络在训练集上的 top-1 准确率结果；(b) 为在训练集中训练的分类网络在验证集上的 top-1 准确率结果；可以看出，使用 KL 散度损失函数训练得到的属性特征有一个更低的验证集 top-1 准确率。这也表示该属性特征中含有更少的身份特征。它验证了 KL 散度损失函数对属性提取网络 A 中消除身份特征有帮助。

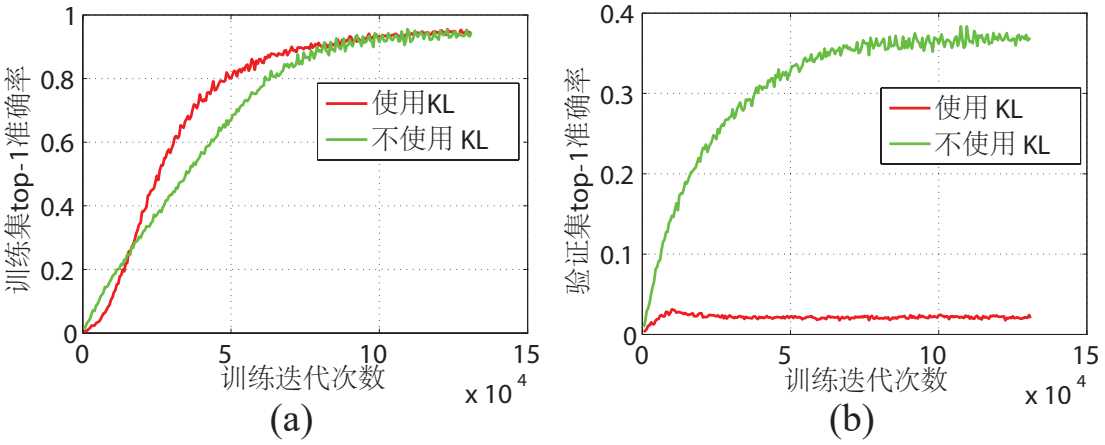


图 5.5　KL 散度损失函数的分析，论文使用 Facescrub 中的数据的属性特征进行分类训练。(a) 为训练的分类网络在训练集上的 top-1 准确率结果；(b) 训练的分类网络在验证集上的 top-1 准确率结果；可以看出使用 KL 散度损失函数得到特征向量在分类中的效果更差，也就是该特征向量中包含更少的身份特征。

5.4　身份保持的生成对抗网络的应用

在本章节中，论文将展示身份保持的生成对抗网络框架可以被用在各种应用中：(1) 人脸属性转换。该技术可以将属性图片中的属性转换到身份图片中。(2) 任意人脸的随机合成。给定一张人脸图片，身份保持的生成对抗网络框架利用身份提取网络 I 得到其身份特征，然后在属性特征随机采样属性特征。然后一起输入进生成网络 G 可以得到各种随机合成的图片。(3) 人脸图片的渐变。选择一张人脸图片作为身份图片。对于任意指定两张人脸图片作为属性图片，身份保持的生成对抗网络框架可以提取到两个属性特征。然后可以对它们得到的属性特征进行线性插值，将身份图片的身份特征和属性特征插值的结果输入到生成模型即可得到该身份图片的在指定属性图片上的渐变结果。(4) 侧脸图片转正脸图片，对于任意一张侧脸图片，身份保持的生成对抗网络框架将其设为身份图片，然后给定一个正脸图片作为属性图片。这样便生成了这张侧脸图片的正脸图片。(5) 人脸识别中对抗样本的检测，对抗样本常常通过在图片中加一个

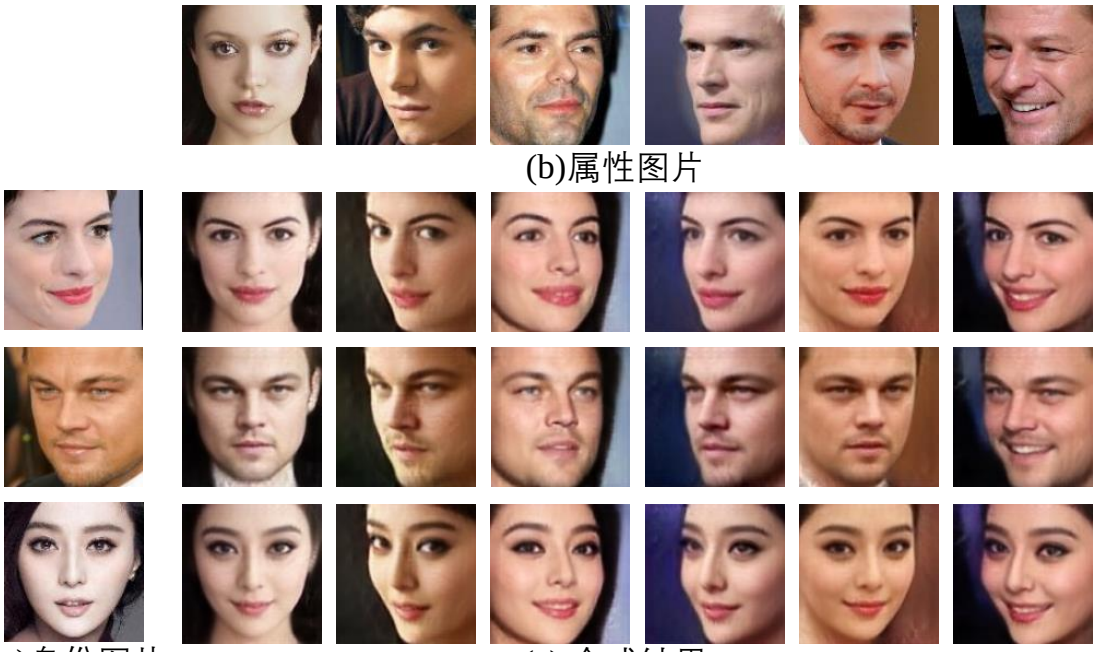


图 5.6　使用训练集中已经存在的身份的人脸图片作为身份图片的属性转换结果。(a) 中为提供身份特征的身份图片，(b) 为提供属性特征的属性图片。(c) 为使用 (a) 中身份特征和 (b) 中属性特征合成的人脸图片。

“扰动”以使得识别网络将其误分类为其他类，身份保持的生成对抗网络框架中的生成网络可以通过对对抗样本的身份特征进行合成，这样对比合成图片和对抗样本可以知道该图片是否为对抗样本。如果该图片为对抗样本，也可以知道该对抗样本的攻击对象是谁。在下面的章节中将作具体的介绍。

5.4.1　人脸属性转换

在本节中，论文展示人脸属性转换的结果。人脸属性转换的目标是使用身份图片 x^s 的身份特征和属性图片 x^a 属性特征合成图片 x'。使得属性图片 x^a 的属性特征转换到身份图片 x^s 上。论文分为两种情况进行试验，一种是测试图片的身份已经存在于训练数据集中，一种是开放数据集中人脸合成，即使用训练数据集中不存在的身份的人脸图片进行人脸图片合成。

图5.8中展示了测试图片的身份已经存在于训练数据集中的人脸属性转换的结果，(a) 中为提供身份特征的身份图片，(b) 为提供属性特征的属性图片。(c) 为使用 (a) 中身份特征和 (b) 中属性特征合成的人脸图片。从结果可以看出，生成的人脸图片非常清晰，同时即保持了输入身份图片的身份特征又保持了输入属性图片的属性特征。

图5.9中展示了测试图片的身份不存在于训练数据集中的人脸属性转换的结果，(a) 中为提供身份特征的身份图片，(b) 为提供属性特征的属性图片。(c)

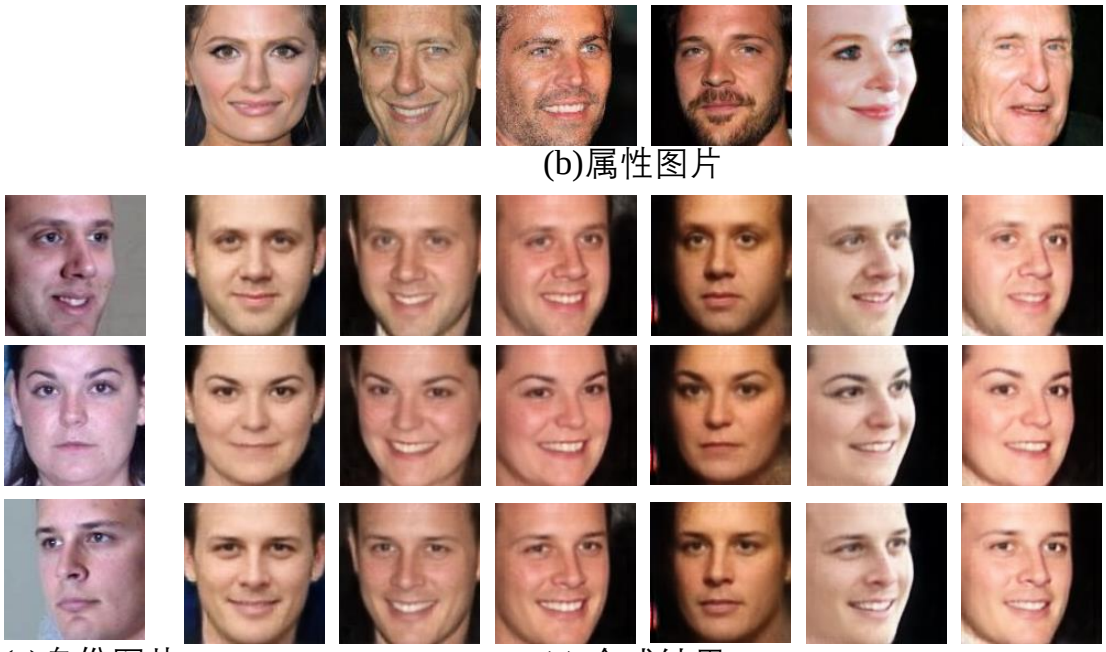


图 5.7　使用训练集中不存在的身份的人脸图片作为身份图片的属性转换结果。(a) 中为提供身份特征的身份图片，(b) 为提供属性特征的属性图片。(c) 为使用 (a) 中身份特征和 (b) 中属性特征合成的人脸图片。

为使用（a）中身份特征和（b）中属性特征合成的人脸图片。从结果可以看出，即使在身份不存在于训练数据集中的人脸合成中，身份保持的生成对抗网络框架合成的人脸图片还是非常清晰，同时即保持了输入身份图片的身份特征又保持了输入属性图片的属性特征。

5.4.2　任意人脸的随机合成

在身份保持的生成对抗网络框架中，属性特征的分布被 KL 散度损失函数约束到一个先验分布 $P(z) \sim N(0, I)$ 中，所以也可以从先验分布 $P(z)$ 中采样属性特征 z_A 作为输入到生成网络 G 的属性特征。

图5.8中展示了测试图片的身份已经存在于训练数据集中的人脸图片和随机采样的属性特征合成人脸图片的结果。（a）中为提供身份特征的身份图片，（b）为（a）中人脸和和随机采样的属性特征合成人脸图片的结果。可以看到生成的人脸图片非常清晰，同时保持了输入身份图片的身份特征.

图5.9中展示了测试图片的身份不存在于训练数据集中的人脸图片和随机采样的属性特征合成人脸图片的结果。(a) 中为提供身份特征的身份图片，(b) 为 (a) 中人脸和和随机采样的属性特征合成人脸图片的结果。可以看到生成的人脸图片非常清晰，同时保持了输入身份图片的身份特征.

(a)身份图片　　(b) 合成结果

图 5.8　使用训练集中已经存在的身份的人脸图片和随机采样的属性特征合成人脸图片的结果。(a) 中为提供身份特征的身份图片，(b) 为使用 (a) 中身份特征和随机采样的属性特征合成的人脸图片。

5.4.3　人脸图片的渐变

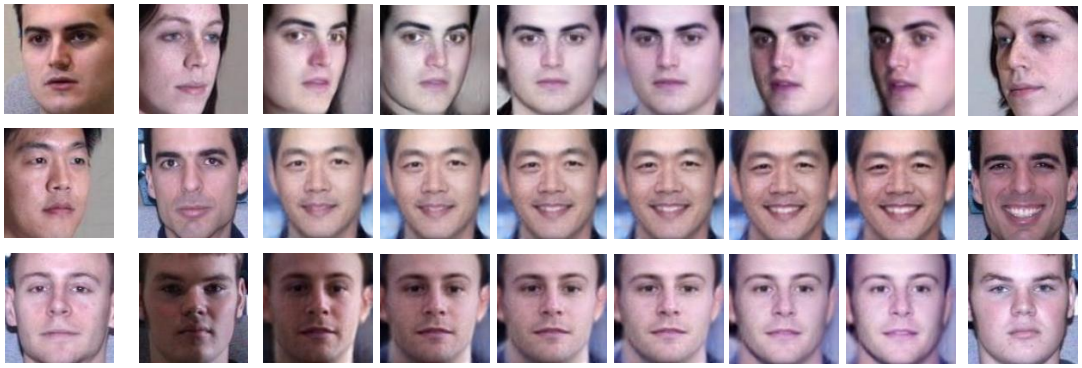
在本节中，论文展示身份保持的生成对抗网络模型可以被用在图片渐变这个任务中，论文在 Multi-PIE 数据集上进行图片渐变的实验。首先选择一对图片 x_1^a 和 x_2^a 作为属性图片，然后身份保持的生成对抗网络框架使用属性提取网络 A 提取器隐空间的表达 z_1 和 z_2。这样之后可以使用线性插值的方式得到一系列的隐空间变量 z。线性插值的表达式为：

$$z = \alpha z_1 + (1-\alpha) z_2, \alpha \in [0,1]. \tag{5.8}$$

最后将得到的线性插值的隐空间变量 z 和其他身份图片提取的身份特征输入生成网络 G 中得到其图片渐变的结果。如图5.10所示为图片渐变的结果。(a) 为输入身份图片，(b) 和 (d) 为输入的属性图片，(c) 是在两张属性图片中的渐变结果。其中第一行为人脸图片的角度的渐变，第二行为表情的渐变，第三行为光照的渐变，可以看到，渐变的过程中身份信息被保持住了。同时渐变的人脸图片

(a)身份图片　　(b) 合成结果

图 5.9　使用训练集中不存在的身份的人脸图片和随机采样的属性特征合成人脸图片的结果。(a) 中为提供身份特征的身份图片，(b) 为使用 (a) 中身份特征和随机采样的属性特征合成的人脸图片。



(a) 身份图片 (b)属性图片x_1^a　　(c)渐变结果　　(d)属性图片x_2^a

图 5.10　身份保持的生成对抗网络应用在人脸图片渐变的结果。(a) 为输入身份图片，(b) 和 (d) 为输入的属性图片，(c) 是 (a) 中身份图片在两张属性图片中的渐变结果。

结果真实且清晰。

5.4.4　侧脸转正脸

尽管身份保持的生成对抗网络框架不是专门设计用来做侧脸转正脸的任务的，但是身份保持的生成对抗网络框架也可以被用来完成这一任务。图5.11中展示身份保持的生成对抗网络框架与别人方法的对比，其他的方法的结果均来自最新的论文 TP-GAN[114] 中。先使用一张正脸图片作为属性图片输入到属性提取网络 A 提取属性特征，然后将侧脸图片输入到身份提取网络 I 提取身份特征，然后可以将身份特征和属性特征输入到生成网络 G 得到最后正脸的结果。比较 IP—GAN 的结果与 TP-GAN[114] 的结果，身份保持的生成对抗网络框架更好滴保证了光照和肤色等属性信息。需要注意的是，身份保持的生成对抗网络在训练过程中没有使用任何带有人脸角度标注的数据，或者说对于侧脸图片转为正脸图片这个任务而言，身份保持的生成对抗网络是一个无监督的方法。而其他的方法在训练中使用了带人脸角度标注的数据。

5.4.5　人脸识别中对抗样本的检测

基于深度卷积神经网络的人脸识别系统已经被广泛用在监控和访问控制中。但是，对抗样本（adversarial examples）[117-120] 的存在使得这些系统的安全性存在风险。在这节中，身份保持的生成对抗网络框架也可以被用在对抗样本的检测中。

对抗样本是由 Szegedy C. 等人在 2013 年发现。他们发现机器学习模型，特别是深度学习模型在正确分类的图片中加一些微小的 “扰动” 即让模型出现分类错误的情况。如图5.12所示。左侧为被正确分类的为熊猫的图片，中间是一个微小的 “扰动”，右侧为在熊猫原图片加上 “扰动” 的图片，该图片被分类为长臂猿了，但是这张照片的内容在人眼看来几乎没有变化。这张加了 “扰动” 的图片被称为对抗样本。

其实在人脸识别模型中同样存在这样的对抗样本，所以论文的目标是在人脸模型中检测出这些对抗样本。论文致力于在人脸校验这个任务中解决该问题。人脸校验这个任务是给定两张人脸图片然后判断两张脸是不是来自同一个人。现在通用的做法是用深度人脸识别网络提取特征。然后计算两个特征的距离，再将该距离与一个设定好的阈值比较。如果距离小于阈值，则判别为来自同一个人，如果大于的话则判别为来自不同人。

假设有两张人脸图片 x_1 和 x_2 有不同的身份，可以发现一个微小 “扰动” r，然后使得 $x_1 + r$ 在人脸识别模型中被误认为是 x_2。这里 $x_1 + r$ 就是这个对抗样本。为了发现这个对抗样本，可以优化下面这个公式：

(a) LFW　(b) IP-GAN　(c) [114]　(d) [115]　(e) [116]

图 5.11　侧脸图片转正脸图片的结果比较。(a) 为输入侧脸图片，(b)，(c)，(d) 和 (e) 是不同方法将输入侧脸转为正脸的结果。

$$\begin{aligned} &\min \|r\|_2^2 \\ &s.t. \|f_C(x_1 + r) - f_C(x_2)\|_2^2 < \theta, \end{aligned} \tag{5.9}$$

其中 f_C 是从人脸识别模型中提取的特征，θ 是预设的阈值。

如图5.13中所示，(a) 和 (c) 是 x_1 和 x_2，(b) 是试图使人脸识别系统误识别为 (c) 的对抗样本 $x_1 + r$，因为对抗样本 $x_1 + r$ 和 x_2 有着相近的身份特征，所以当用身份保持的生成对抗网络重构原图时，对抗样本 $x_1 + 1$ 重构的人脸图片的身份和 x_2 相似。基于这个观察，可以使用身份保持的生成对抗网络模型对人脸图片生成重构人脸图片，然后比较重构人脸图片和原输入图片的身份即可以识别对抗样本。

论文使用 LFW 数据集执行该实验。对于 LFW 中的 3000 个不同人的成对图

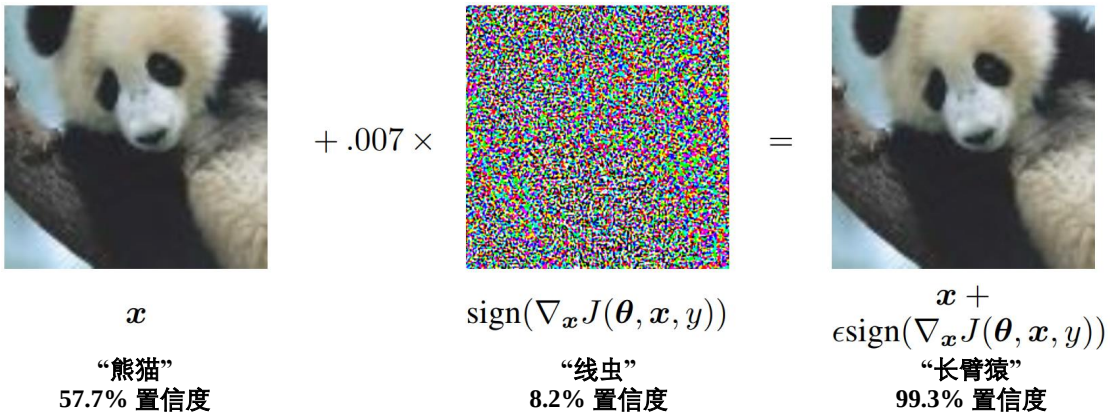


图 5.12　深度学习中的对抗样本实例，左侧为被正确分类的为熊猫的图片，中间是一个微小的“扰动”，右侧为在熊猫原图片加上“扰动”的图片，该图片被以 99.3% 的置信概率分类为长臂猿了，但是这张照片的内容在人眼看来和左侧图片几乎相同。这张加了“扰动”的图片被称为对抗样本。

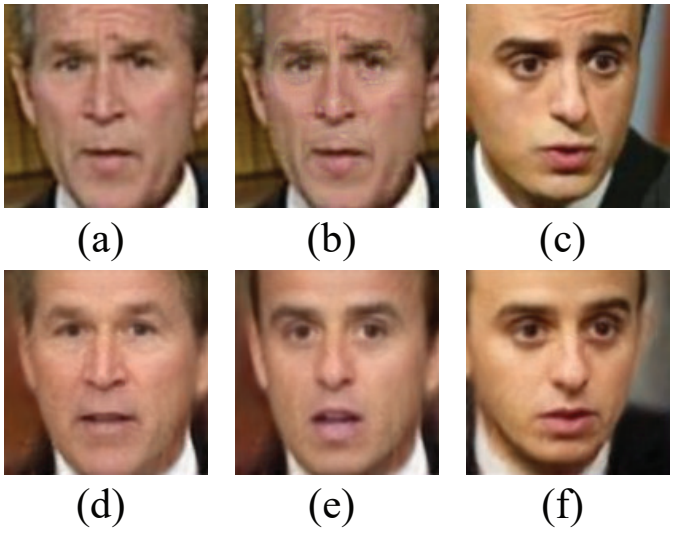


图 5.13　人脸识别系统中的对抗样本检测，(a) 是用作为对抗样本的原图，(b) 是试图使人脸识别系统误识别为 (c) 的对抗样本，(d)，(e) 和 (f) 是经过身份保持的生成对抗网络后重构出的图片。可以观察到尽管对抗样本 (b) 和 (a) 有着差不多的内容。但是它们经过身份保持的生成对抗网络重构出的图片缺有着完全不同的身份，该过程即是使用身份保持的生成对抗网络检测对抗样本的方法。

片。论文利用上面生成对抗样本的方法相互“攻击”一次得到两个对抗样本。对于每一个预设的阈值可以得到 6000 个对抗样本。实验中选择了不同的阈值进行试验：$[0.4, 0.6, 0.8, 1]$。然后针对 6000 张原图片和 6000 个对抗样本，可以利用训练好的身份保持的生成对抗网络框架得到他们的重构图片。然后对图片和其重构图片的图片对进行一个分类的任务。尽管也可以训练一个分类网络去检测。但是这个过程也是容易被攻击的，所以本文使用了传统的方法做分类。论文使用了传统的特征 LBP [121]. 对于原图，对抗样本和它们的重构均使用了 LBP 提取

表 5.4　不同特征距离阈值下对抗样本的检测准确率。

特征距离阈值	准确率
1.0	76.73%
0.8	82.58%
0.6	87.18%
0.4	92.41%

特征，然后使用 SVM[122] 训练一个分类器。论文做了一个 10 子集的交叉验证，结果如表格 5.4所示，在特征距离小于阈值 0.4 的情况下，对抗样本检测准确率能达到 92.41%，这反映了身份保持的生成对抗网络框架在合成人脸过程中优秀的身份保持能力。

5.5　小结与讨论

在本章中，论文提出了一个面向开放数据集的身份保持人脸图片合成框架。以实现指定身份和属性的人脸图片合成。该框架可以分离人脸图片中的身份特征和属性特征，然后重组该身份特征和从另外一张人脸图片提取的属性特征得到一张新的人脸图片。实验结果表明，该框架实现了开放集中的保持身份的人脸图片合成。同时该框架可以被应用在各种任务中，例如人脸属性转换，侧脸转正脸，人脸识别系统中的对抗样本检测等等。在未来工作工作，论文希望可以将身份保持的生成对抗网络框架用在更多的数据集中。

在第 3 章和第 4 章中，论文分别提出特征匹配条件生成对抗网络框架和条件变分生成对抗网络框架完成标签作为输入的图片合成。但是在这两个框架中，标签被表示成一位有效向量（one-hot vector），所以框架仅能完成训练集中已有标签的图片合成。本章为了解决训练集中不存在标签的图片合成，进一步提出了身份保持的生成对抗网络框架。

借助于深度卷积神经网络中人脸识别网络的发展，论文将身份的表达由原来的一位有效向量升级为人脸识别网络中的特征表达。这样针对任意指定人脸，只需用人脸识别网络提取其特征即可，然后完成其人脸的图片合成。如何打破人脸图片的限制，进一步完成指定物体的图片的合成是值得研究的。

第 6 章　总结与展望

本章首先对于提出的三种基于生成对抗网络的图像合成算法进行总结，然后讨论本文工作的不足，最后基于这些不足展望基于生成对抗网络的图像合成技术的未来发展方向。

6.1　全文总结

图像合成是计算机视觉、计算机图形学中的一个广泛研究的问题。它具有广泛的应用前景，例如：现代电影中的特效处理、游戏画面渲染、动画设计、图片处理等。目前图像合成的主要挑战在于难以保证合成图像的真实性、多样性和与输入条件一致性。因此如何在图像合成中满足这些性质，一直是图像合成研究的重点。针对现有基于生成对抗网络的图像合成算法存在的挑战与局限性，本文围绕这一方向做出了以下贡献：

1. 为了解决原始生成对抗网络训练不稳定的问题，本文提出了特征匹配损失函数。在该损失函数中，判别网络 D 依然使用原始的二元交叉熵损失函数，生成网络 G 则使用判别网络 D 中特征匹配损失函数。该损失函数解决了生成对抗网络中原始的梯度消失问题，使得生成对抗网络的训练更加稳定，从而帮助合成模型合成质量更高的图像。同时，该损失函数可以用在条件生成框架中，帮助合成模型在条件生成中合成更加符合条件的图片。

2. 提出新的基于生成对抗网络的框架条件变分生成对抗网络（CVAE-GAN），在该框架中加入的编码网络可有效解决生成对抗网络的训练中出现的模式坍缩的问题。编码网络将图片空间映射到隐空间，再使用生成网络将隐空间映射回图片空间，因为原图片空间的分布中的图片是多样的，所以生成网络生成的图片也是多样的。这样解决了生成对抗网络中的模式坍缩问题。同时在该框架中，本文加入分类网络约束生成照片从而保证合成图片能够保持给定的条件信息。同时实验结果表明提出的条件变分生成对抗网络框架可以完成很多应用，比如（1）图片的修复：将图像中破损的区域修复好；（2）图片的渐变，可以得到两张图片的渐变图片；（3）相同属性的图片的检索，在人脸库中检索出属性（表情、角度、光照等）和查询图片相同的图片；（4）数据增强：将生成数据作为数据增强的方法加入训练集中以提升分类模型的精度。

3. 为了解决条件变分生成对抗网络只能合成训练集中已有标签的图像的限制，本文提出了身份保持的生成对抗网络（IP-GAN）。该框架可以实现指定身份和属性的人脸图片合成。当输入为任意一张人脸图片时，身份保持的生成对抗网络框架可以解耦人脸图片中的身份特征和属性特征（角度、表情、光照等等），然后重组该身份特征和从另外一张脸中提取的属性特征得到一张新的人脸图片。身份保持的生成对抗网络框架同样可以完成很多应用，比如（1）人脸属性转换。该技术可以将一个人脸图片中的属性特征转换到另外一个人脸图片中。（2）侧脸图片转换为正脸图片。对于任意一张侧脸图片，可以利用框架得到其正脸图片。（3）人脸识别系统中对抗样本的检测。身份保持的生成对抗网络框架可以利用重构的方法检测出试图攻击人脸识别系统中的对抗样本。

6.2　不足与展望

图像合成是计算机视觉领域、计算机图形学中的研究热门方向，尽管经过几十年的发展，现在图像合成的水平与人类的水平依然有着非常大的差距。本文虽然在图像合成中提出了一些新的算法与框架，也取得了许多新的进展。但是由于作者时间、精力及研究水平有限，本文的工作也存在一些不足之处：

1. 虽然加入编码网络进入生成对抗网络中可以在一定程度上解决生成对抗网络中的模式坍缩问题，但是关于生成对抗网络中出现模式坍缩问题的数学原理仍不清楚。同时编码网络只能在真实图像分布较为简单时解决模式坍缩问题，对于复杂的真实图像分布，变分自编码器不能实现从真实图像分布到隐空间分布再回到真实图像分布的隐射，重构损失函数不能使生成器生成图像分布和真实图像分布一致，所以模式坍缩问题仍然存在，未来关于生成对抗网络的模式坍缩问题还需进一步研究。

2. 虽然身份保持的生成对抗网络可以完成指定身份的人脸合成，但是其合成的人脸图片的分辨率比较低，无法完成高分辨率的人脸图像合成。这是因为当合成图像的分辨率提高时，高分辨率图像对生成对抗网络训练的稳定性要求更高，另一方面是在高分辨率图像中保持输入的身份特征更加困难。

所以未来工作可以从以下几个方向着手：

1. 研究在生成对抗网络中出现模式坍缩的数学原理，目前有一些工作 [12] 从模型中参数的奇异值出发研究这个问题，并且发现模式坍缩出现的时候的一些共性，比如在发生模式坍缩时，模型参数的奇异值会发生跳变。但是为什么出现参数的奇异值的跳变原因还不清楚。研究其中的数学原理，并从数学原理出发提出解决方案使生成对抗网络训练中免于模式坍缩问题是一个值得研究的方向。

2. 研究高分辨率图像的合成，改进现在已有的生成对抗网络的损失函数、网路结构、或者训练方法使其满足高分辨率图像的合成是一个非常值得研究的方向。同时研究满足条件的高分辨率图像的合成。目前有一些工作 [123] 尝试使用不同的将条件信息输入到生成网络中的方式，并且发现新设计的加入条件信息的方式可以使生成网络生成更加满足条件信息的图片。

3. 探究图像合成的新应用，比如更困难的图像合成。完成从一个故事到一段动画的生成，一个电影剧本到电影的生成，未来有更多的图像合成应用值得研究。

参 考 文 献

[1] GÜÇLÜTÜRK Y, GÜÇLÜ U, VAN LIER R, et al. Convolutional sketch inversion[C]// European Conference on Computer Vision. Springer, 2016: 810-824.

[2] ZHANG H, XU T, LI H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[J]. arXiv preprint arXiv:1612.03242, 2016.

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

[4] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[5] JOHNSON M K, DALE K, AVIDAN S, et al. Cg2real: Improving the realism of computer generated images using a large collection of photographs[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17:1273-1285.

[6] LICKLIDER J C, TAYLOR R W. The computer as a communication device[J]. Science and Technology, 1968, 76(2):1-3.

[7] EFROS A A, LEUNG T K. Texture synthesis by non-parametric sampling[C]//Proceedings of the seventh IEEE International Conference on Computer Vision: volume 2. IEEE, 1999: 1033-1038.

[8] TURK M A, PENTLAND A P. Face recognition using eigenfaces[C]//Proceedings CVPR'91. 586-591.

[9] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database [C]//2009 IEEE conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.

[10] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.

[11] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[J]. arXiv preprint arXiv:1805.08318, 2018.

[12] BROCK A, DONAHUE J, SIMONYAN K. Large scale gan training for high fidelity natural image synthesis[J]. arXiv preprint arXiv:1809.11096, 2018.

[13] SMITH A R, BLINN J F. Blue screen matting[C]//SIGGRAPH: volume 96. 1996: 259-268.

[14] LEE H, BATTLE A, RAINA R, et al. Efficient sparse coding algorithms[C]//Advances in Neural Information Processing Systems. 2007: 801-808.

[15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014: 2672-2680.

[16] VAN DEN OORD A, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[J]. arXiv preprint arXiv:1601.06759, 2016.

[17] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[J]. arXiv preprint arXiv:1606.03498, 2016.

[18] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[19] HYVÄRINEN A, KARHUNEN J, OJA E. Independent component analysis: volume 46[M]. John Wiley & Sons, 2004.

[20] LAROCHELLE H, MURRAY I. The neural autoregressive distribution estimator.[C]// AISTATS: volume 1. 2011: 2.

[21] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[J]. arXiv preprint arXiv:1401.4082, 2014.

[22] SOHN K, LEE H, YAN X. Learning structured output representation using deep conditional generative models[C]//Advances in Neural Information Processing Systems. 2015: 3483-3491.

[23] LARSEN A B L, SØNDERBY S K, WINTHER O. Autoencoding beyond pixels using a learned similarity metric[J]. arXiv preprint arXiv:1512.09300, 2015.

[24] DENTON E L, CHINTALA S, FERGUS R, et al. Deep generative image models using a laplacian pyramid of adversarial networks[C]//Advances in Neural Information Processing Systems. 2015: 1486-1494.

[25] DOSOVITSKIY A, SPRINGENBERG J, TATARCHENKO M, et al. Learning to generate chairs, tables and cars with convolutional networks[J]. 2016.

[26] YAN X, YANG J, SOHN K, et al. Attribute2image: Conditional image generation from visual attributes[J]. arXiv preprint arXiv:1512.00570, 2015.

[27] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[28] XU L, JORDAN M I. On convergence properties of the em algorithm for gaussian mixtures [J]. Neural Computation, 1996, 8(1):129-151.

[29] PERMUTER H, FRANCOS J, JERMYN I H. Gaussian mixture models of texture and colour for image database retrieval[C]//Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on: volume 3. IEEE, 2003: III-569.

[30] THEIS L, HOSSEINI R, BETHGE M. Mixtures of conditional gaussian scale mixtures applied to multiscale image representations[J]. PloS One, 2012, 7(7):e39857.

[31] STARNER T, PENTLAND A. Real-time american sign language recognition from video using hidden markov models[M]//Motion-Based Recognition. Springer, 1997: 227-243.

[32] MNIH V, HINTON G E, et al. Generating more realistic images using gated mrf's[C]// Advances in Neural Information Processing Systems. 2010: 2002-2010.

[33] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):504-507.

[34] SALAKHUTDINOV R, HINTON G E. Deep boltzmann machines.[C]//AISTATS: volume 1. 2009: 3.

[35] TU Z. Learning generative models via discriminative approaches[C]//2007 IEEE conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-8.

[36] GREGOR K, DANIHELKA I, GRAVES A, et al. Draw: A recurrent neural network for image generation[J]. arXiv preprint arXiv:1502.04623, 2015.

[37] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[J]. arXiv preprint arXiv:1511.05644, 2015.

[38] SAK H, SENIOR A, BEAUFAYS F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition[J]. arXiv preprint arXiv:1402.1128, 2014.

[39] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. Lstm: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10):2222-2232.

[40] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier gans [J]. arXiv preprint arXiv:1610.09585, 2016.

[41] MIYATO T, KOYAMA M. cgans with projection discriminator[J]. arXiv preprint arXiv:1802.05637, 2018.

[42] OORD A V D, KALCHBRENNER N, VINYALS O, et al. Conditional image generation with pixelcnn decoders[J]. arXiv preprint arXiv:1606.05328, 2016.

[43] ZHAO J, MATHIEU M, LECUN Y. Energy-based generative adversarial network[J]. arXiv preprint arXiv:1609.03126, 2016.

[44] BERTHELOT D, SCHUMM T, METZ L. Began: Boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717, 2017.

[45] QI G J. Loss-sensitive generative adversarial networks on lipschitz densities[J]. arXiv preprint arXiv:1701.06264, 2017.

[46] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[J]. arXiv preprint arXiv:1606.03657, 2016.

[47] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein gan[J]. arXiv preprint arXiv:1701.07875, 2017.

[48] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]//Proceedings

of the IEEE International Conference on Computer Vision. 2017: 2794-2802.

[49] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[J]. arXiv preprint arXiv:1704.00028, 2017.

[50] NOWOZIN S, CSEKE B, TOMIOKA R. f-gan: Training generative neural samplers using variational divergence minimization[C]//Advances in Neural Information Processing Systems. 2016: 271-279.

[51] WANG R, CULLY A, CHANG H J, et al. Magan: Margin adaptation for generative adversarial networks[J]. arXiv preprint arXiv:1704.03817, 2017.

[52] ZHU J Y, KRÄHENBÜHL P, SHECHTMAN E, et al. Generative visual manipulation on the natural image manifold[C]//European Conference on Computer Vision. Springer, 2016: 597-613.

[53] MROUEH Y, SERCU T, GOEL V. Mcgan: Mean and covariance feature matching gan[J]. arXiv preprint arXiv:1702.08398, 2017.

[54] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[J]. arXiv preprint arXiv:1812.04948, 2018.

[55] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

[56] XU B, WANG N, CHEN T, et al. Empirical evaluation of rectified activations in convolutional network[J]. arXiv preprint arXiv:1505.00853, 2015.

[57] GOLUB G H, VAN DER VORST H A. Eigenvalue computation in the 20th century[M]// Numerical Analysis: Historical Developments in the 20th Century. Elsevier, 2001: 209-239.

[58] YOSHIDA Y, MIYATO T. Spectral norm regularization for improving the generalizability of deep learning[J]. arXiv preprint arXiv:1705.10941, 2017.

[59] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[60] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[61] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1501-1510.

[62] WANG T C, LIU M Y, ZHU J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 8798-8807.

[63] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[J]. arXiv preprint arXiv:1605.05396, 2016.

[64] REED S E, AKATA Z, MOHAN S, et al. Learning what and where to draw[C]//Advances in Neural Information Processing Systems. 2016: 217-225.

[65] ZHANG H, XU T, LI H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. arXiv preprint arXiv:1710.10916, 2017.

[66] XU T, ZHANG P, HUANG Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 1316-1324.

[67] MA S, FU J, WEN CHEN C, et al. Da-gan: Instance-level image translation by deep attention generative adversarial networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 5657-5666.

[68] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 1125-1134.

[69] CHEN Q, KOLTUN V. Photographic image synthesis with cascaded refinement networks [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1511-1520.

[70] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2223-2232.

[71] LIU M Y, BREUEL T, KAUTZ J. Unsupervised image-to-image translation networks[J]. arXiv preprint arXiv:1703.00848, 2017.

[72] YI Z, ZHANG H, TAN P, et al. Dualgan: Unsupervised dual learning for image-to-image translation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2849-2857.

[73] CHOI Y, CHOI M, KIM M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 8789-8797.

[74] ZHU J Y, ZHANG R, PATHAK D, et al. Toward multimodal image-to-image translation[C]// Advances in Neural Information Processing Systems. 2017.

[75] HUANG X, LIU M Y, BELONGIE S, et al. Multimodal unsupervised image-to-image translation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 172-189.

[76] YU J, LIN Z, YANG J, et al. Generative image inpainting with contextual attention[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 5505-5514.

[77] CHEN J, CHEN J, CHAO H, et al. Image blind denoising with generative adversarial network based noise modeling[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 3155-3164.

[78] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 4681-4690.

[79] BERTALMIO M, SAPIRO G, CASELLES V, et al. Image inpainting[C]//Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co., 2000: 417-424.

[80] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4):107.

[81] YEH R A, CHEN C, YIAN LIM T, et al. Semantic image inpainting with deep generative models[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 5485-5493.

[82] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. Patchmatch: A randomized correspondence algorithm for structural image editing[C]//ACM Transactions on Graphics (ToG): volume 28. ACM, 2009: 24.

[83] PERARNAU G, VAN DE WEIJER J, RADUCANU B, et al. Invertible conditional gans for image editing[J]. arXiv preprint arXiv:1611.06355, 2016.

[84] PORTENIER T, HU Q, SZABÓ A, et al. Faceshop: Deep sketch-based face image editing [J]. ACM Transactions on Graphics (TOG), 2018, 37(4):99.

[85] NIMISHA T M, KUMAR SINGH A, RAJAGOPALAN A N. Blur-invariant deep learning for blind-deblurring[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4752-4760.

[86] KUPYN O, BUDZAN V, MYKHAILYCH M, et al. Deblurgan: Blind motion deblurring using conditional adversarial networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 8183-8192.

[87] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2):295-307.

[88] HUYNH-THU Q, GHANBARI M. Scope of validity of psnr in image/video quality assessment[J]. Electronics Letters, 2008, 44(13):800-801.

[89] EPHRAIM Y, MALAH D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(6):1109-1121.

[90] HORE A, ZIOU D. Image quality metrics: Psnr vs. ssim[C]//2010 20th International Con-

ference on Pattern Recognition. IEEE, 2010: 2366-2369.

[91] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[92] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]//Advances in Neural Information Processing Systems. 2017: 6626-6637.

[93] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[94] ARJOVSKY M, BOTTOU L. Towards principled methods for training generative adversarial networks[C]//NIPS 2016 Workshop on Adversarial Training. In review for ICLR: volume 2016. 2017.

[95] NGUYEN A, DOSOVITSKIY A, YOSINSKI J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[C]//Advances in Neural Information Processing Systems. 2016: 3387-3395.

[96] SHARIF RAZAVIAN A, AZIZPOUR H, SULLIVAN J, et al. Cnn features off-the-shelf: an astounding baseline for recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops. 2014: 806-813.

[97] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[98] MCMAHAN B, STREETER M. Delay-tolerant algorithms for asynchronous distributed online learning[C]//Advances in Neural Information Processing Systems. 2014: 2915-2923.

[99] NG H W, WINKLER S. A data-driven approach to cleaning large face datasets[C]//2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014: 343-347.

[100] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]//Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on. IEEE, 2008: 722-729.

[101] WELINDER P, BRANSON S, MITA T, et al. Caltech-UCSD Birds 200: number CNS-TR-2010-001[R]. California Institute of Technology, 2010.

[102] CHEN D, REN S, WEI Y, et al. Joint cascade face detection and alignment[C]//European Conference on Computer Vision. Springer, 2014: 109-122.

[103] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[104] YI D, LEI Z, LIAO S, et al. Learning face representation from scratch[J]. arXiv preprint arXiv:1411.7923, 2014.

[105] NGUYEN A, YOSINSKI J, BENGIO Y, et al. Plug & play generative networks: Conditional iterative generation of images in latent space[J]. arXiv preprint arXiv:1612.00005, 2016.

[106] JOLLIFFE I. Principal component analysis[M]. Springer, 2011.

[107] LEARNED-MILLER E, HUANG G B, ROYCHOWDHURY A, et al. Labeled faces in the wild: A survey[M]//Advances in Face Detection and Facial Image Analysis. Springer, 2016: 189-248.

[108] TRAN L, YIN X, LIU X. Disentangled representation learning gan for pose-invariant face recognition[C]//CVPR: volume 4. 2017: 7.

[109] YIN X, YU X, SOHN K, et al. Towards large-pose face frontalization in the wild[C]//Proc. ICCV. 2017: 1-10.

[110] GUO Y, ZHANG L, HU Y, et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition[C]//European Conference on Computer Vision. Springer, 2016: 87-102.

[111] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep face recognition[C]//British Machine Vision Conference: volume 1. 2015: 6.

[112] WU X, HE R, SUN Z, et al. A light cnn for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11):2884-2896.

[113] GROSS R, MATTHEWS I, COHN J, et al. Multi-pie[J]. Image and Vision Computing, 2010, 28(5):807-813.

[114] HUANG R, ZHANG S, LI T, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis[J]. arXiv preprint arXiv:1704.04086, 2017.

[115] ZHU X, LEI Z, YAN J, et al. High-fidelity pose and expression normalization for face recognition in the wild[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015: 787-796.

[116] HASSNER T, HAREL S, PAZ E, et al. Effective face frontalization in unconstrained images [C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015: 4295-4304.

[117] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [J]. arXiv preprint arXiv:1312.6199, 2013.

[118] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[119] CARLINI N, WAGNER D. Adversarial examples are not easily detected: Bypassing ten detection methods[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017: 3-14.

[120] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world

[J]. arXiv preprint arXiv:1607.02533, 2016.

[121] RAHIM M A, AZAM M S, HOSSAIN N, et al. Face recognition using local binary patterns (lbp)[J]. Global Journal of Computer Science and Technology, 2013.

[122] JOACHIMS T. Making large-scale svm learning practical[R]. Technical report, SFB 475: Komplexitätsreduktion in Multivariaten ⋯, 1998.

[123] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[J]. arXiv preprint arXiv:1903.07291, 2019.

致　谢

光阴如梭，一晃博士毕业的时间也到了，值此论文完成之际，谨对多年来对我予以关心与帮助的老师，同学，朋友和家人们表达我最诚挚的感谢。

感谢我的导师李厚强老师。自我研一进入实验室以来，李老师就在生活、学习、科研上给与我无微不至的关心与照顾，并在研二为我提供了微软亚洲研究院的实习机会。在学术上，李老师严谨而又谦虚的科研态度深深感染了我。在生活中，不管是在学校还是校外实习，李老师在繁重的工作之余始终关心着我的生活状态。依然清晰的记得，有一次在学校急性肠胃炎发作，李老师督促我赶紧去医院并且嘱咐我多休息。感谢李老师在科研、生活以及人生规划中所有的悉心指导与帮助，让我可以幸运地走在科研的道路上。

感谢我的辅导师罗杰波老师。罗杰波老师虽然远在国外，但却时常关心我的科研状态，时常在微信群中鼓励我们，给我们分享最新最热的科研方向。罗老师每次去微软亚洲研究院总是与我细心的讨论科研状态。他严谨的科研态度和开阔的科研视野深深影响着我。

感谢在学校实验室期间的帮助过我的周文罡老师，周老师严谨谦虚的科研风格深深感染着我。同时感谢在实验室一起玩耍一起组织活动的同学们：王敏、李跃、蒲俊胡、刘一丁、刘家俊、张鹏、王家喻等等。

感谢我在微软亚洲研究院实习期间亦师亦友的陈栋师兄。陈栋师兄是我开始计算机视觉研究的引路人，在我没有文章发表并对写论文感到惶恐时，陈栋师兄给我很多帮助与鼓励。在微软研究院实习期间，陈栋师兄手把手带着我思考每一个科研问题，仔细讨论每一个实验细节，探究实验结果背后的数学理论。正是陈栋师兄的悉心指导，让我体会到了科研的魅力并使我有了在科研上继续走下去的决心。依然清晰的记得在每一个 deadline 之前，陈栋师兄陪我们一起通宵，一起做实验，一起写论文。这些一起奋斗的时光在生活中深深激励着我。

感谢在微软亚洲研究院实习期间帮助过我的华刚老师、闻芳老师、袁路老师、张婷、杨昊、杨蛟龙。他们在我科研上也给予我很多帮助。华刚老师在科研上时常让我们探究科研工作的数学原理和本质，并且也为我提供很多新的科研方向。闻芳老师总是鼓励我不仅要做好科研，也要想办法将科研成果转化为实际的产品。感谢张婷，杨昊，杨蛟龙，袁路老师对我写论文上的指导，并让我在每一次书写论文过程中受益匪浅。

感谢在微软亚洲研究院实习期间的小伙伴们：陈冬冬、贾伟、李博杰、翟耀、张志锐、郑书新、李潇、付登攀、曾兆阳、张攀、古纾㊣、李凌志等等，我会始终铭记与你们一起奋斗一起玩耍的快乐时光。

感谢在大学本硕博期间认识的同学们：段文哲、刘中流、张睿智、何威、刘震、马晓鹏、冯译、尹沛然、陶辛锴、张义飞等等，那些一起上课一起玩耍的时光记忆犹新。

最后感谢我的父母家人们，他们一路关心我，鼓励我，是我永远最坚强的后盾。

在读期间发表的学术论文与取得的研究成果

已发表论文

1. **Jianmin Bao**, Dong Chen, Fang Wen, Houqiang Li, Gang Hua; CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training[C]. The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2745-2754
2. **Jianmin Bao**, Dong Chen, Fang Wen, Houqiang Li, Gang Hua; Towards Open-Set Identity Preserving Face Synthesis[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6713-6722
3. Shuyang Gu, **Jianmin Bao**, Dong Chen, Fang Wen, Lu Yuan. Mask-Guided Portrait Editing with Conditional GANs[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

已申请专利

1. Dong Chen, Fang Wen, Gang Hua, **Jianmin Bao**. Face Synthesis. 专利申请号: 201810082732.8.