班 级 1602051

学 号 16020510038

西安电子科技大学

本科毕业设计论文

题 目 基于单幅图像的面部表情生成算法研究

学 院 人工智能学院

专 业 智能科学与技术

学生姓名 邢博伟

导师姓名 毛莎莎

西安电子科技大学

毕业设计（论文）诚信声明书

本人声明：本人所提交的毕业论文《 基于单幅图像的面部表情生成算法研究 》是本人在指导教师指导下独立研究、写作的成果，论文中所引用他人的无论以何种方式发布的文字、研究成果，均在论文中加以说明；有关教师、同学和其他人员对本文的写作、修订提出过并为我在论文中加以采纳的意见、建议，均已在我的致谢辞中加以说明并深致谢意。

本论文和资料若有不实之处，本人承担一切相关责任。

论文作者：邢博伟（签字）时间： 2020 年 5 月 26 日

指导教师已阅：毛莎莎（签字）时间： 2020 年 5 月 26 日

西 安 电 子 科 技 大 学

毕业设计（论文）任务书

学生姓名 邢博伟 学号 16020510038 指导教师 毛莎莎 职称 讲师

学院 人工智能学院 专业 智能科学与技术

题目名称 基于单幅图像的面部表情生成算法研究

任务与要求

现有的表情数据库大部分是根据研究需要对人为表情的收集，并非真实环境下的自然表情，且都是一些单一夸张的表情，同时大部分收集表情数据具有数据量小的缺点。该课题将研究在基于单幅的人脸表情数据下，生成更多更自然的表情数据用于人脸表情识别算法研究。主要的任务和要求如下：

1. 了解 GANs 算法研究现状，查阅相关论文文献和资料；
2. 完成不少于三篇英文论文的阅读以及一篇文献综述，并完成不少于一万字的论文翻译；
3. 收集课题相关数据、学习 GANs 环境的搭建；
4. 学习 GANs 模型架构，掌握一种基于单幅图像数据生成算法，并完成程序编写与复现。
5. 设计合适的算法，实现基于单幅图像数据生成的表情识别算法；
6. 撰写毕业论文。

开始日期 2019.12.1 完成日期 2020.6.1

院长（签字）_____ 年 月 日

注：本任务书一式两份，一份交学院，一份学生自己保存。

西 安 电 子 科 技 大 学

毕业设计（论文）工作计划

学生姓名 邢博伟 学 号 16020510038

指导教师 毛莎莎 职 称 讲师

学 院 人工智能学院 专 业 智能科学与技术

题目名称 基于单幅图像的面部表情生成算法研究

一、毕业设计（论文）进度

起 止 时 间	工 作 内 容
2019.12.1 -- 2020.2.17	查阅 GANs 算法相关文献资料，学习 GANs 算法和表情识别算法的概念和理论。大量阅读相关的文献资料并翻译英文文献 1 万字以上。
2020.2.18 -- 2020.3.20	学习 GANs 模型及环境搭建方法，研究现有的 GANs 生成人脸表情算法，编程复现相应算法代码，并进行仿真实验。
2020.3.21 -- 2020.5.9	针对已有算法中存在的问题，设计合理的实验方案及改善方法，对算法进一步训练。
2020.5.10 -- 2020.6.1	撰写毕业论文，进行毕业论文答辩。

二、主要参考书目（资料）

[1] Shaham T R, Dekel T, Michaeli T. Singan: Learning a generative model from a single natural image[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 4570-4580.

[2] 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.

[3] 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.

[4] Pumarola A, Agudo A, Martinez A M, et al. Ganimation: Anatomically-aware facial animation from a single image[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 818-833.

三、主要仪器设备及材料

1. 硬件：计算机、高性能服务器（有 GPU）等；
2. 软件：Python、PyTorch、Anaconda、Word 等；

四、教师的指导安排情况（场地安排、指导方式等）

1. 每周集中汇报、指导一次；
2. 通过邮件、电话等方式随时进行讨论、交流和指导；

五、对计划的说明

说明：时间和内容可根据实际情况进行相应调整。

注：本计划一式两份，一份交学院，一份学生自己保存（计划书双面打印）

西安电子科技大学 人工智能 学院

本科生毕业论文（设计）开题报告

（ 2020 届）

学生姓名 邢博伟

专　　业 智能科学与技术

学　　号 16020510038

指导教师 毛莎莎

2019 年 12 月 26 日

（本表一式三份，学生、指导教师、学院各一份）

一、论文名称及项目来源

1、论文名称

基于单幅图像的面部表情生成算法研究

2、项目来源

导师课题

二、研究目的和意义

1、研究目的

现有的表情数据库大部分是根据研究需要对人为表情的收集，并非真实环境下的自然表情，且都是一些单一夸张的表情，同时大部分收集表情数据具有数据量小的缺点。该课题将研究在基于单幅的人脸表情数据下，生成更多更自然的表情数据用于人脸表情识别算法研究。

2、研究意义

人脸表情分析作为一种特殊的人类行为分析任务，在人与人的交流中，它能够有效地传达非语言信息及情感的交流，从而辅助听者推断说话人的意图，其中，面部表情是人类情绪的直观反应，它能够有效地表达个人的情绪、认知和主体状态，也是人际交往的重要渠道。人脸表情是传播人类情感信息与协调人际关系的重要方式，据心理学家 A.Mehrabia 的研究表明，在人类的日常交流中，通过语言传递的信息仅占信息总量的 7%，而通过人脸表情传递的信息却达到信息总量的 55%。尤其，当说话人在试图掩盖内在情绪时，面部表情的细微变化是无法隐藏和无法抑制的，它所传达的信息暗含了潜在的个体行为信息，是与人类情感、精神状态、健康状态等诸多因素相关的一种复杂的表达方式。

人脸识别技术正在经历前所未有的发展，关于人脸识别技术讨论从未停歇。目前，人脸识别精度已经超过人眼，同时大规模普及的软硬件基础条件也已具备，应用市场和领域需求很大，基于这项技术的市场发展和具体应用正呈现蓬勃发展态势。人脸表情识别作为人脸识别技术中的一个重要组成部分。人脸表情分析与识别对于公共监控、人机交互、安全驾驶、机器人、临床医学、测谎技术、精神病理分析等领域具有相当重要的研究意义。

三、国内外研究现状和发展趋势

1、基于 GAN 的人脸表情识别算法

国内的王小庆团队为了解决目标数据集样本不足的问题，他们在目标数据集上训练了一个生成式对抗网络(GAN)，利用生成的样本对原数据集上预训练的模型进行微调。

2、基于 WGAN 的人脸表情识别算法

国内的姚乃明等人提出了一种鲁棒的人脸表情识别方法, 能够以用户无关方式识别具有局部遮挡的人脸表情.基于 Wasserstein GAN (WGAN), 训练了一个稳定的人脸图像生成网络, 然后使用遮挡图像集优化网络的输入隐变量, 对遮挡区域进行保持上下文一致性的人脸图像补全.对无遮挡图像和遮挡补全图像, 在表情识别任务和身份识别任务之间建立了一种对抗关系, 通过在表情特征提取过程中抑制由身份信息导致的类内差异来提升表情识别的准确性和鲁棒性。

3、基于 GC-GAN 的生成人脸表情

乔凤春等人提出了一种几何对比生成对抗网络（GC-GAN），用于在不同主题之间传递连续的情感。给定带有某种情感的输入面部和来自另一受试者的目标面部表情，GC-GAN 可以生成具有目标表情的身份保留面部。将几何信息作为连续条件引入 CGAN，以指导面部表情的生成。为了处理跨不同主题或情感的错位，对比学习用于将几何学流形转换为面部表情的嵌入式语义流形。因此，将嵌入的几何体注入到 GAN 的潜在空间中，并有效地控制情绪的产生。

4、MaskGAN：趋于多样化和交互式的面部图像处理

香港中文大学的郑寒利提出了一个新颖的面向几何的面部操作框架 MaskGAN，其中包含两个经过精心设计的组件：密集映射网络和编辑行为模拟培训。语义掩膜可作为具有保真度的保留功能的灵活面部操作的合适中间表征。对语义掩膜赋予情感元素，并与 GAN 结合，可生成多样化的人脸表情。

5、StarGAN: 多领域的图像迁移学习

国外的 Yunjey Choi 等人提出了 StarGAN 多领域的图像迁移学习的模型。这种统一模型体系结构允许在单个网络中同时训练具有不同域的多个数据集。这使得 StarGAN 的翻译图像的质量优于现有的模型，以及灵活转换输入图像到任何理想的目标领域的新能力。通过实验验证了该方法在人脸属性转移和表情合成任务

上的有效性。

6、GANimation：生成连续变化的表情

入选于 ECCV 2018 的 GANimation 构建了一种人脸解剖结构（anatomically）上连续的面部表情合成方法，能够在连续区域中呈现图像，并能处理复杂背景和光照条件下的图像。GANimation 的目的是建立一种具有 FACS 表现水平的合成面部动画模型，并能在连续领域中无需获取任何人脸标志（facial landmark）而生成具有结构性（anatomically-aware）的表情。它与其他已有的 GAN 方法相比，无论是在结果的视觉质量还是生成的可行性上，都是具有优势的。

四、主要研究内容、要解决的问题及本文的初步方案

1、主要研究内容

基于生成对抗神经网络（GAN），将单幅的人脸表情图像数据以无监督的方式进行训练，进而可以生成更多样更自然的同一个人的人脸表情新样本。

2、要解决的问题

(1) 数据量少。一般情况下，训练 GAN 来生成图像，至少需要有成千上万的训练数据才可能有不错的效果。仅凭单张人脸表情图像，如何训练一个生成式模型？

(2) 如何生成新的表情图像？虽然之前有单张图像训练生成式模型，但这些工作大多是条件式生成模型，只能完成某种特定的图像到图像的转换任务（如超分辨率等），另外一些非条件式生成模型也只局限在纹理这样简单结构的图像生成上。怎样让训练后的模型可以接受一个输入，生成新的表情图像，是一个难点。

(3) 如何同时调节多个 AU？AU 为人脸的动作单元(Action Unit)，来源于面部动作编码系统(FACS)。AU 的不同组合，可构成不同的表情。如果生成更多样更自然更复杂的情绪，则需要同时调节多个 AU。相比于调节单个 AU，调节多个 AU 难度更大，是在挑战模型的极限。

(4) 要克服光照、遮挡物的影响，模型的鲁棒性要强。

3、初步方案

毕业设计尝试采用将 ICCV 2019 最佳论文 SinGAN 的单幅图像生成式算法和入选 ECCV 2018 的 Oral 的 GANimation 的连续表情生成式算法相结合，对单幅图像数据生成多种人脸表情。SinGAN 可以生成新的具有真实感的图像样本，在保留了原始的图像块分布的基础上，创造了新的物体外形和结构。SinGAN 是第一个非条件式的、使用单张自然图像训练的生成式模型。GANimation 可同时调节多个 AU，生成更复杂多样的表情，且克服光照、遮挡物的影响，训练结果非常好。将两种算法结合使用，便可解决上述问题，得到令人满意的结果。

五、工作的主要阶段、进度和完成时间

1、第一阶段（2019.12.1 -- 2020.2.17）

查阅 GANs 算法相关文献资料，学习 GANs 算法和表情识别算法的概念和理论。大量阅读相关的文献资料并翻译英文文献 1 万字以上。

2、第二阶段（2020.2.18 -- 2020.3.20）

学习 GANs 模型及环境搭建方法，研究现有的 GANS 生成人脸表情算法，编程复现相应算法代码，并进行仿真实验。

3、第三阶段（2020.3.21 -- 2020.5.9）

针对已有算法中存在的问题，设计合理的实验方案及改善方法，对算法进一步训练。

4、第四阶段（2020.5.10 -- 2020.6.1）

撰写毕业论文，进行毕业论文答辩。

六、已进行的前期准备工作

查阅了相关领域已有的部分研究成果，包括组 GANs 和人脸表情识别相关的论文，对两者的发展历史与基本原理有了初步了解。

七、指导教师意见

该生对课题内容了解详细，工作计划步骤合理，予以通过。

签名　毛莎莎

2019 年 12 月 27 日

八、学院审核意见

签名

年　　月　　日

学号 16020510038

西安电子科技大学
本科生毕业设计（论文）中期报告

题　　目	基于单幅图像的面部表情生成算法研究
学生姓名	邢博伟
专　　业	智能科学与技术
学　　号	**16020510038**
指导教师	毛莎莎
报告日期	**2020** 年 **3** 月 **23** 日

西安电子科技大学本科生院制

西安电子科技大学

本科生毕业设计（论文）中期报告要求

一、本科生在完成学位论文开题之后三个月内，必须进行学位论文中期考核，考核由各学院自行组织，具体要求参照毕业设计文件执行。

二、中期考核结论分为两种：1.通过，按专家意见修改后继续学位论文撰写工作；2.不通过，重新考核，正式答辩前达不到通过标准的，答辩延期进行。

三、中期报告由学生填写，填写完成后，需在限定时间内，在教务系统中上传最终版（如有更新，可重新上传覆盖）。

四、中期考核时需携带此表，本表一式三份，本人、指导教师、学院各一份，用 A3 纸张正反套印；承担毕业设计单位审核盖章后的表格最终胶装入存档论文中。

五、表格填写要求：正文字体宋体，字号小四，行间距固定值 20 磅，可续页，请勿更改表格样式。

1、毕业设计工作是否更换题目及是否按开题报告预定的内容及进度安排进行

有更换题目。

毕业设计按开题报告预定的内容及进度安排进行。

2．目前已完成的研究工作及结果（内容要详实充分）

（1）开题报告已完成。

（2）完成翻译英文文献共计一万字（《Facial Expression Database》，（《Shaham_SinGAN_Learning_a_Generative_Model_From_a_Single_Natural_Image_ICCV_2019_paper》）

（3）完成周记共十五篇。

（4）研究 SinGAN 模型，并掌握其理论依据为使用多个 GAN 结构分别学习了不同尺度下分辨率 11x11 的图像块的分布，并从粗糙到细致、从低分辨率到高分辨率逐步生成真实图像；其可应用于随机生成自然图像、超分辨率、绘画图像转换、图像编辑、图像和谐化等；其局限性为如果图像块差异较大，训练结果失真和生成图像的内容高度受限于训练图像提供的语义信息。

（5）选定论文编辑方式为 Latex，并已拟定初稿。

（6）编程语言选择 python，深度学习框架为 PyTorch 并已配置完毕。

（7）已复现 SinGAN 代码，生成自然图像和人脸图像。

3．后期拟完成的研究工作及进度安排（要有可行性）

（1）研究已有的 SinGAN 算法，进一步理解理论递推关系。

（2）根据人脸的 AU 单元，调节 SinGAN 深度学习神经网络结构，使得生成的人脸图像不失真。

（3）在已有的 SinGAN 算法的基础上，尝试将其与 StarGAN 算法和 GANimation 融合，解决 SinGAN 生成人脸图像失真且单一的难题。

（4）由于处理数据庞大，需借助强大的算力平台。之后将采用华为云深度学习平台，完成大规模数据的深度学习处理。

（5）根据已拟定的初稿，进一步构思并编辑论文框架，细化论文主题。进一步加快学习 Latex 排版技能，解决编辑不规范不便捷的问题。

4．存在的困难与问题

（1）由 SinGAN 生成的人脸图像有图像失真的问题。为解决这一问题，我将会将原图降采样后输入 N-1 层和 N-2，并根据人脸的 AU 单元，调节 SinGAN 深度学习神经网络结构，使得生成的人脸图像不失真。

（2）由 SinGAN 生成的人脸图像有语义单一的难题。为解决这一难题，我将尝试将其与 StarGAN 算法和 GANimation 融合，使得生成的人脸表情图像内容不受限于训练图像提供的语义，生成种类更多的人脸表情图像。

（3）由于自身算力的有限，且需要处理大量的人脸表情数据，需借助线上强大的算力平台。我将采用华为云深度学习平台，完成大规模数据的深度学习处理，解决这个问题。

5．如期完成全部论文工作的可能性

大概率可以如期完成全部论文工作

6、指导导师意见

学生基本完成该阶段的目标和任务。通过早期相关文献阅读以及论文翻译等工作，对课题内容和课题方向具有充分了解，能够准确理解毕业设计的研究方向和思路，目前已构建具体的算法思路和框架，并已初步实现设计算法的代码测试。

导师签名：毛莎莎

2020 年　3 月 23　日

7、中期报告检查组意见

（中期考核结论分为两种：1.通过，按专家意见修改后继续学位论文撰写工作；2.不通过，重新考核，正式答辩前达不到通过标准的，答辩延期进行。评语重点指出中期报告存在的问题并提出具体修改意见和建议。）

通过

实现的算法效果目前还有一定的提升空间。建议：通过对面部的局部区域结构信息进行分析，加入结构信息约束来改善面部表情生成效果。

组长签名：王佳宁

成员签名：陈璞花
毛莎莎
古晶
李玲玲

2020 年 3 月 23 日

8、承担毕业设计单位审核（盖章）

（校内毕设学生由学院审核，校外毕设学生由承担毕设企业或单位审核）

审核意见：

盖章：

年 月 日

西安电子科技大学
毕业设计（论文）中期检查表

学生姓名	邢博伟	学　号	16020510038	班　级	1602051
学　院	人工智能学院		专　业	智能科学与技术	
导师姓名	毛莎莎	职　称	讲师	学　院	人工智能学院
题目名称	基于单幅图像的面部表情生成算法研究				

检　查　内　容	检　查　结　果
题目是否更换及更换原因	是 原题目与毕业设计主题稍有不妥，需进行适当修改
学生出勤情况	全勤
进度评价 （完成总工作量的百分比）	60%
质量评价	学生基本完成该阶段的目标和任务。通过早期相关文献阅读以及论文翻译等工作，对课题内容和课题方向具有充分了解，能够准确理解毕业设计的研究方向和思路，目前已构建具体的算法思路和框架，并已初步实现设计算法的代码测试。
总体评价 （按优、良、中、及格、不及格五档评价）	优
存在的问题与建议	实现的算法效果目前还有一定的提升空间。建议：通过对面部的局部区域结构信息进行分析，加入结构信息约束来改善面部表情生成效果。
学院审核（盖章）	

注：此表由指导教师填写，与学生填写的中期报告配套，填写完成交学院办公室，中期检查成绩将作为毕业设计总成绩的一部分；此表装订入毕业设计（论文）中。

西安电子科技大学
毕业设计（论文）指导教师评定意见表

<table>
<tr><td>学　　院</td><td colspan="3">人工智能学院</td><td>专　业</td><td colspan="2">智能科学与技术</td></tr>
<tr><td>姓　　名</td><td>邢博伟</td><td>学 号</td><td colspan="2">16020510038</td><td>成　绩</td><td>优</td></tr>
<tr><td>题目名称</td><td colspan="6">基于单幅图像的面部表情生成算法研究</td></tr>
<tr><td>指导教师</td><td colspan="3">毛莎莎</td><td>职　称</td><td colspan="2">讲师</td></tr>
<tr><td>指导教师评语及对成绩的评定意见</td><td colspan="6">（指导教师可从以下几方面对毕业设计（论文）进行评审：1.独立查阅文献及调查论证的能力；2.方案设计与实验技能；3.分析与解决问题的能力；4.工作量、工作态度；5.论文质量；6.创新能力。）

该生学习态度认真、积极主动，认真完成任务书中内容，顺利阅读外文资料并按要求完成外文翻译，译文准确，同时也能及时追踪与该选题相关的最新研究成果，查阅计算机视觉顶级会议中的相关论文。在算法设计和实现方面，该生能够熟练运用深度学习算法平台，通过实验运行和分析对研究问题进行深入探索，提出相应的改进和完善策略，具有较强独立分析和解决问题能力。毕业论文结构清晰，章节内容完整丰富，格式规范，算法描述准确，实验方案合理，并通过大量实验分析展现其设计算法性能，具有一定的创新能力。

建议成绩：☑优　□良　□中　□及格　□不及格

签名　毛莎莎　2020 年 5 月 12 日</td></tr>
</table>

注：学院、专业名均写全称。

西安电子科技大学
毕业设计（论文）评阅人评定意见表

<table>
<tr><td>学　　院</td><td colspan="3">人工智能学院</td><td>专　业</td><td colspan="2">智能科学与技术</td></tr>
<tr><td>姓　　名</td><td>邢博伟</td><td>学 号</td><td colspan="2">16020510038</td><td>成　绩</td><td>优</td></tr>
<tr><td>题目名称</td><td colspan="6">基于单幅图像的面部表情生成算法研究</td></tr>
<tr><td>指导教师</td><td colspan="3">毛莎莎</td><td>职　称</td><td colspan="2">讲师</td></tr>
<tr><td>评阅人评语及成绩评定意见</td><td colspan="6">（评阅人主要从以下几方面对毕业设计（论文）进行评价：1.综合运用理论知识和基本技能的能力；2.分析和解决实际问题的能力；3.论文写作能力；4.所解决的问题与论文质量评价。）

论文选题符合专业培养目标，具有较高的学术研究价值。论文层次清晰，内容完整，算法描述准确，较好掌握与课题相关的基础知识和专业知识，能熟练运用深度学习算法平台进行算法编程及实验，实验方案设计合理，将单幅图像数据生成问题有效应用到面部表情数据生成问题中，具有一定的创新性及研究价值，该论文实验充分且展示其提出算法的有效性，论文格式基本符合规范要求，论文整体质量较优。

建议成绩：☑优　☐良　☐中　☐及格　☐不及格

签名　焦昶哲　　2020 年 5 月 12 日</td></tr>
</table>

注：学院、专业名均写全称。

西安电子科技大学
毕业设计（论文）成绩登记表

<table>
<tr><td>学　　院</td><td colspan="3">人工智能学院</td><td>专　业</td><td colspan="2">智能科学与技术</td></tr>
<tr><td>姓　　名</td><td>邢博伟</td><td>学 号</td><td>16020510038</td><td>成　绩</td><td colspan="2">优秀</td></tr>
<tr><td>题目名称</td><td colspan="6">基于单幅图像的面部表情生成算法研究</td></tr>
<tr><td>指导教师</td><td colspan="3">毛莎莎</td><td>职　称</td><td colspan="2">讲师</td></tr>
<tr><td>答辩小组
意见</td><td colspan="6">该生答辩时语言表达流畅，重点突出地阐述论文选择背景、算法设计目标、算法内容、实验结果、当前实验中存在的不足、下一步改进工作、以及通过算法设计和实验总结的一些经验分享，表达准确，能流利回答答辩老师提出问题，根据盲审评审的意见也准确地对论文进行修改。毕业论文结构清晰、内容充实，算法描述准确，实验充足，格式符合规范要求。

建议成绩：☑优　☐良　☐中　☐及格　☐不及格

签名 王佳宁，毛莎莎，陈璞花，古晶，李玲玲
2020 年 5 月 31 日</td></tr>
<tr><td>学院答辩委
员会意见</td><td colspan="6">答辩委员会
主任签名 _____ （学院盖章） 年 月 日</td></tr>
</table>

注：学院、专业名均写全称。

摘 要

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

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

关键词：单幅图像　表情生成　GANimation　SinGAN

摘　要

ABSTRACT

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

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

Key words: single image facial expression generation GANimation SinGAN

ABSTRACT

目 录

第一章 绪 论

1.1 人脸表情概述

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

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

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

1.2 研究意义与目的

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

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

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

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

1.3 内容安排

第一章为绪论部分，主要介绍面部表情的相关知识和本文的研究意义与目的。

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

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

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

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

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

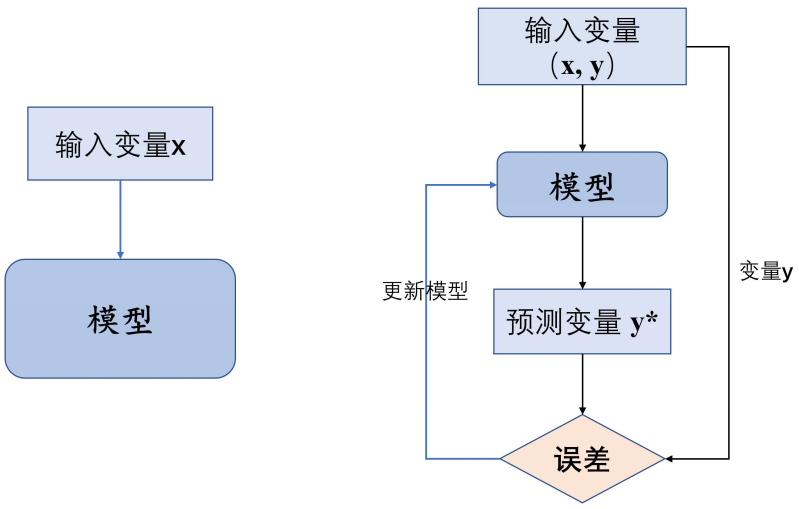
第二章　生成对抗网络（GAN）

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

2.1　生成对抗网络的理论基础

2.1.1　生成式模型

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



（a）无监督学习　　（b）有监督学习

图 2.1 无监督学习与有监督学习流程图

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

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

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

2.1.2 生成对抗网络模型

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

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

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

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

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

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

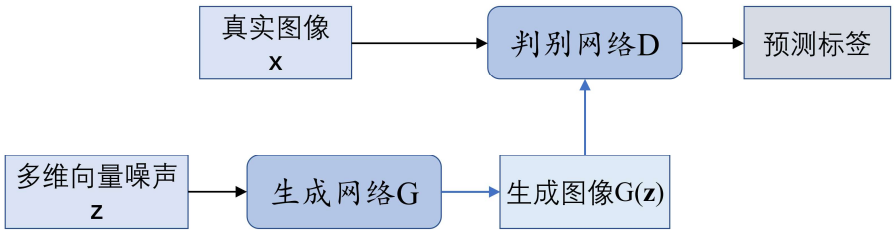


图 2.2 GAN 训练过程流程图

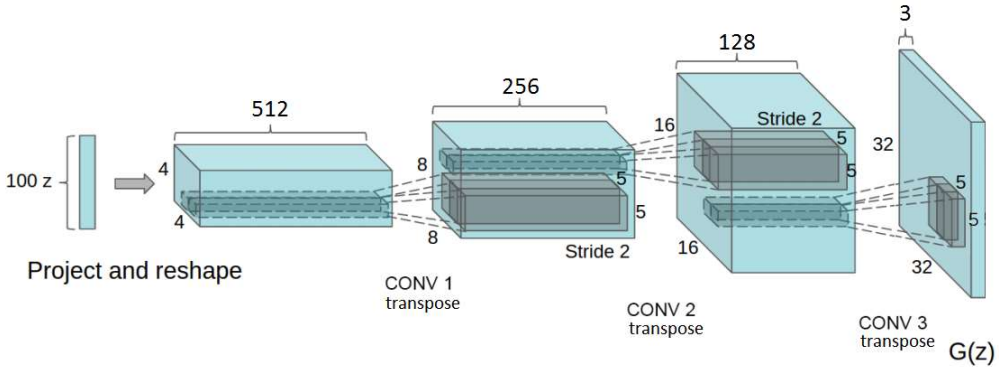


图 2.3 DCGAN 的生成网络结构图

2.2　生成对抗网络的常用模型

2.2.1　深度卷积生成对抗网络（DCGAN）

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

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

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

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

2.2.2 条件生成对抗网络（CGAN）

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

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

2.2.3 循环生成对抗网络（CycleGAN）

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

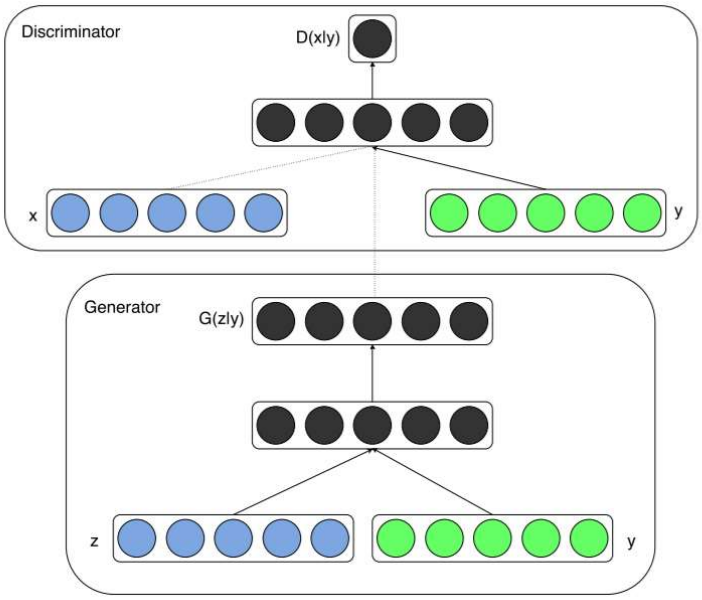


图 2.4 CGAN 网络结构图

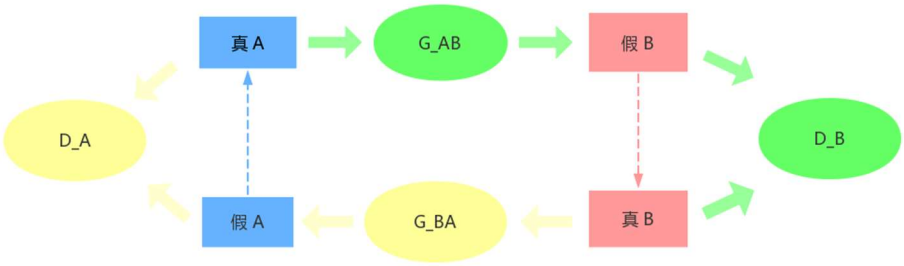


图 2.5 CycleGAN 网络结构图

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

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

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

对于判别网络 B：

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

对于生成网络 BA：

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

对于生成网络 AB：

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

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

第三章　SinGAN 模型及其应用

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

3.1　SinGAN 相关基础

3.1.1　单项深度模型

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

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

3.1.2　图像处理的生成模型

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

图 3.1 SinGAN 与单个图像纹理生成

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

3.2 SinGAN 模型的基本原理

3.2.1 概述

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

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

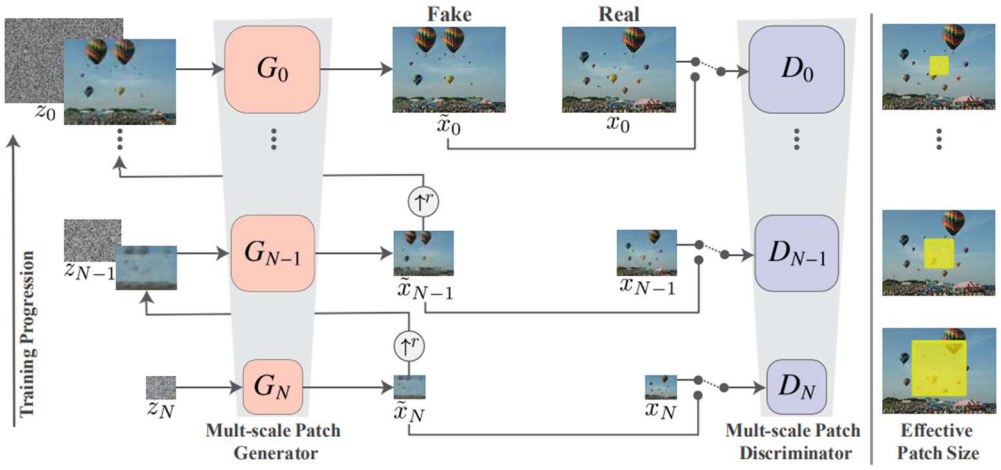


图 3.2 SinGAN 的多尺度传递途径

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

3.2.2　多尺度结构

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

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

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

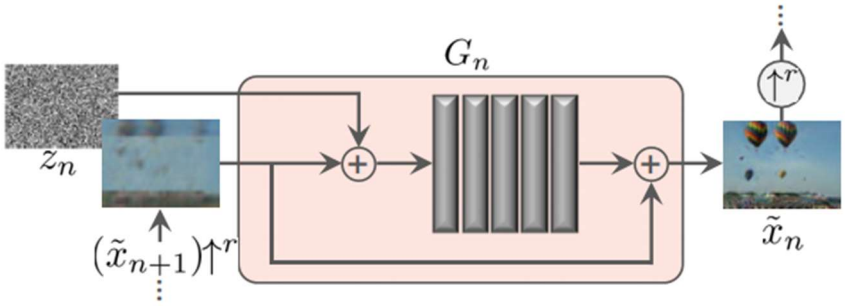


图 3.3 单规模迭代

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

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

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

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

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

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

3.3　SinGAN 模型的应用

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

图 3.4 SinGAN 模型的应用展示

3.3.1　超分辨率

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

表 3.1 超分辨率对比

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

图 3.5 超分辨率效果对比。

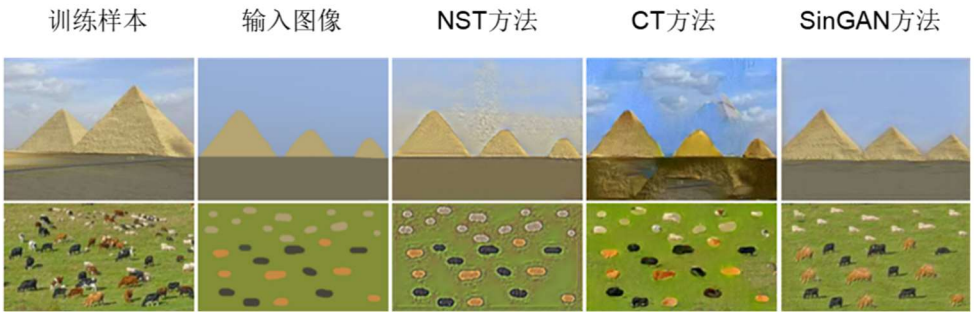


图 3.6 画风迁移效果对比

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

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

3.3.2 图画到图像的画风迁移

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

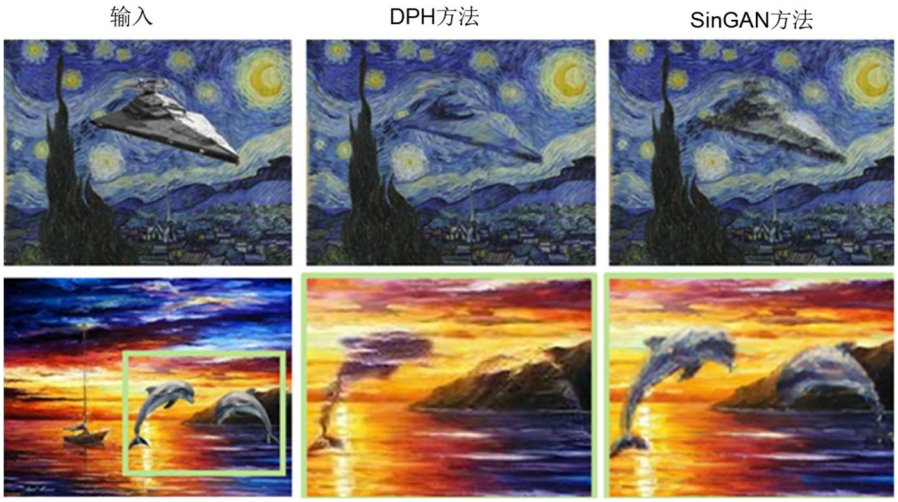


图 3.7 图像调和效果对比

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

3.3.3　图像调和

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

3.3.4　图像编辑

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

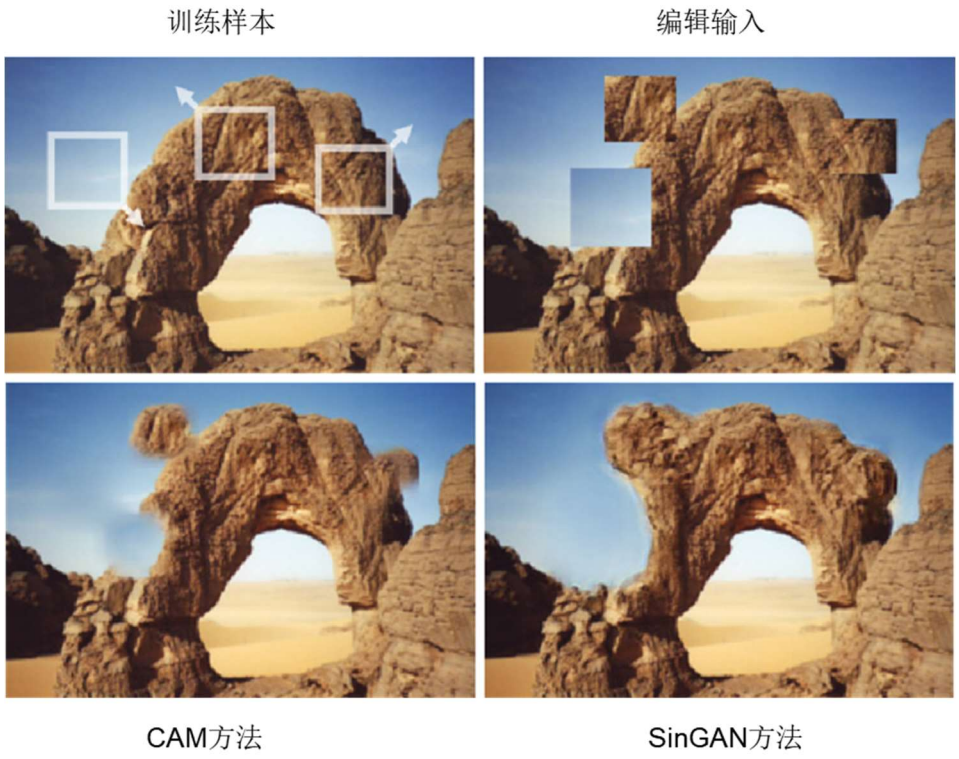


图 3.8 图像编辑效果对比

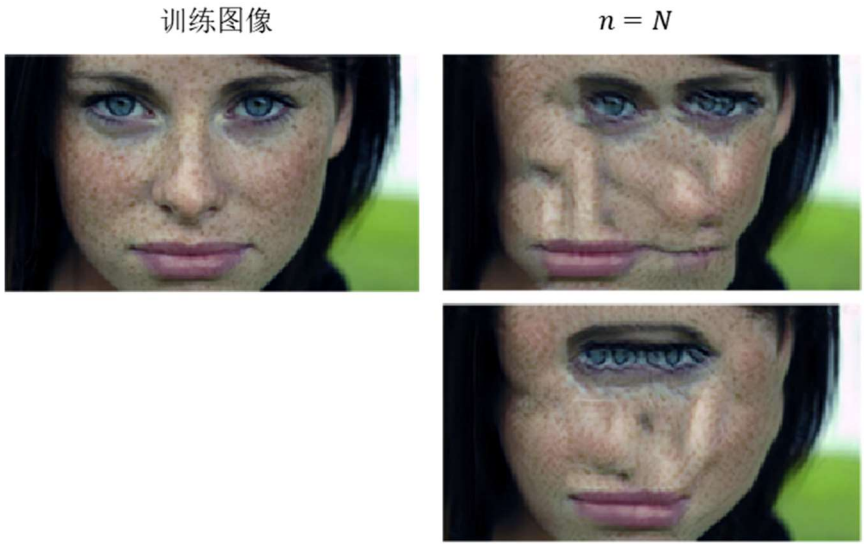


图 3.9 人脸图像训练 SinGAN 模型的生成效果

3.3.5 单一图像生成动画

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

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

3.4　本章小结

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

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

第四章　GANimation 模型及其应用

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

4.1　GANimation 模型的相关基础

4.1.1　非匹配的图像转换

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

4.1.2　面部图像处理

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

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

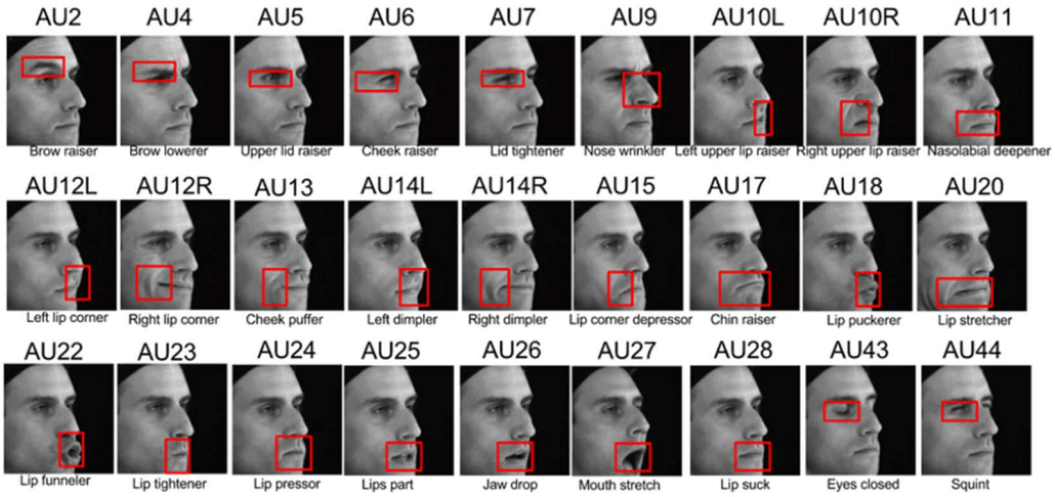


图 4.1 常见 AU 类别表示

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

4.1.3 人脸动作单元（AU）

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

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

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

4.2　GANimation 模型架构和方法

4.2.1　待解决的问题

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

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

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

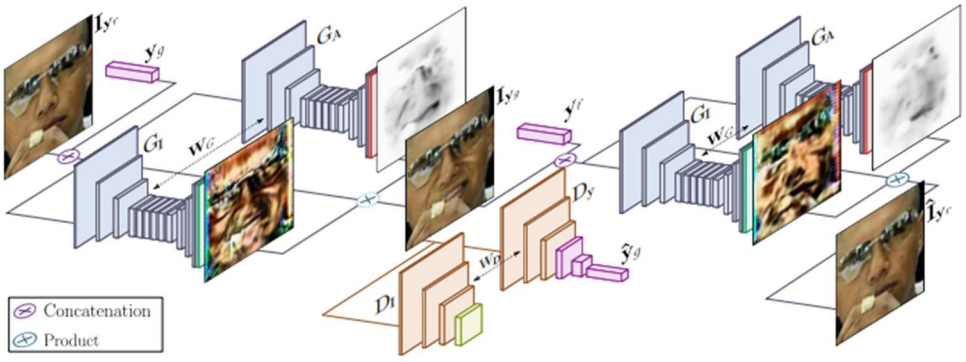


图 4.2 GANimation 模型的结构

4.2.2 网络结构

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

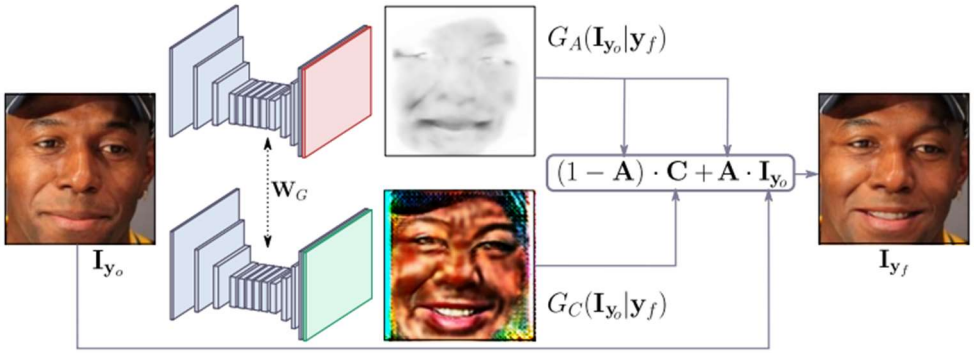


图 4.3 基于注意力机制的生成网络

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

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

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

条件化判别网络，即以生成图像真实性和期望表情完成度作为评价标准的判别网络。$D(\mathbf{I})$ 的结构类似于 PatchGan 网络[18]，从输入图像 I 映射到一个矩阵

$\mathbf{Y}_{\mathrm{I}} \in \mathbb{R}^{H/2^6 \times W/2^6}$，其中 $\mathbf{Y}_{\mathrm{I}}[i,j]$ 表示重叠图像块 ij 为真实的概率。此外，为了评估其条件作用，在网络顶端添加副回归项首部，来估计在图像中 AUs 的激活函数 $\hat{\mathbf{y}} = \left(\hat{y}_1, \ldots, \hat{y}_N\right)^{\top}$。

4.2.3　模型学习

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

① 图像对抗损失函数

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

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

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

其中，λ_{gp} 为惩罚系数。

② 注意力损失函数

在训练模型时，与颜色掩膜 C 类似，没有对注意力掩膜 A 进行 ground-truth 注释，而从判别模块的结果梯度和其他损失函数中学习的。然而，注意力掩膜很容易

饱和到 1，这使得$\mathbf{I}_{\mathbf{y}_o}=G\left(\mathbf{I}_{\mathbf{y}_o}\mid\mathbf{y}_f\right)$，也就是说，生成网络没有起效。为了防止这种情况，GANimation 用一个l_2权重惩罚系数来调整掩膜。同时，为了在将输入图像像素与颜色变换C相结合时，进行平滑的空间颜色变换，GANimation 对A进行全变差正则化。因此，注意力损失$\mathcal{L}_{\mathrm{A}}\left(G,\mathbf{I}_{\mathbf{y}_o},\mathbf{y}_f\right)$可以定义为：

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

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

③ 条件化表情损失函数

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

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

④ 一致性损失函数

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

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

为了生成逼真的图像，对低频信号和高频信号都进行建模。GANimation 的 PatchGan 基于判别网络D_{I}，通过限制对局部图像块结构的注意力来强化高频信号

图 4.4 GANimation 模型生成的失败结果

的准确性。为了捕获低频信号，使用 l_1 范数便已足够。在初步实验中，尽管没有性能的提升，还是尝试用更复杂的感知损失函数[22]来代替 l_1 范数。

⑤ 全损失函数

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

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

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

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

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

4.3 本章小结

GANimation 是一种基于 AU 标注的 GAN 条件化方法，该方法在连续的流行中描述了定义人类表情的面部解剖运动。GANimation 模型采用完全无监督策略训练，只需要激活 AU 标注图像，并利用注意力机制，便可对不断变化的背景和光照条件具有鲁棒性。相比于 StarGAN[15]只能由数据集决定生成离散的表情，其生成的图像连续自然，较为清晰。相比于其他条件生成模型，GANimation 在合成多类表情和处理自然图像的能力上均有超越。

GANimation 在某些情况下会出现失败结果，如图 4.4 所示。这可能是因为输入图像仅为一张，训练数据不足引起的。当输入极端表情时，颜色掩膜没有及时调整权重，会导致局部出现透明化。如果输入图像的对象是非人类，模型的效果也会很差。GANimation 生成图像作为人脸表情数据集的扩充，数量方面还远远不够，尤其不足以满足深度学习表情识别海量训练数据的需要，此方面待以改进。

第五章　SinGANimation 算法仿真实验与分析

5.1　SinGANimation 模型架构

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

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

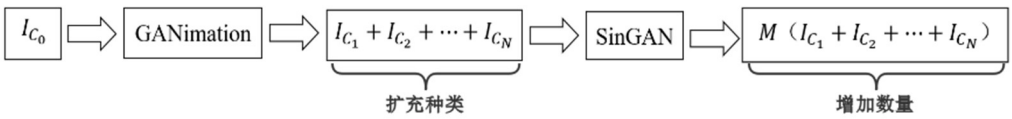


图 5.1 SinGANimation 模型架构

5.2　SinGANimation 模型训练

5.2.1　GANimation 模型训练

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

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

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

5.2.2 SinGAN 模型训练

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

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

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

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

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

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

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

5.3　实验数据集

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

5.3.1　CelebA 数据集

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

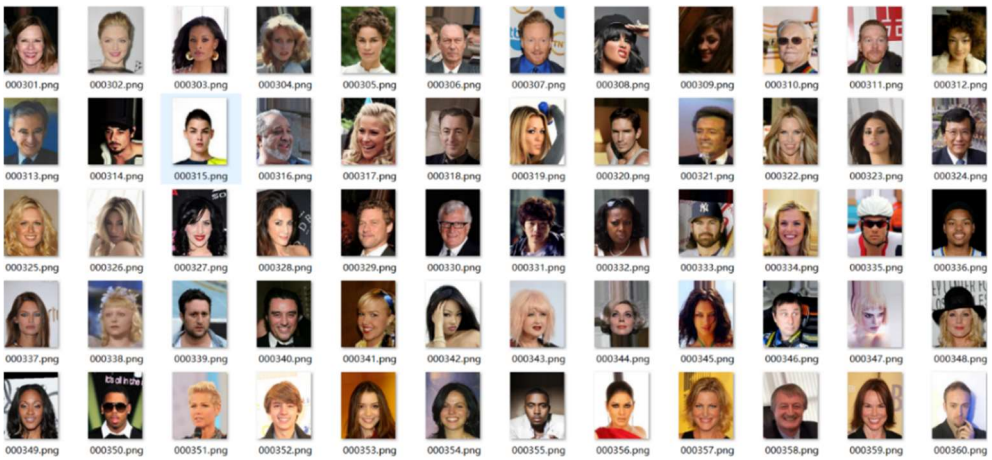


图 5.2　CelebA 数据集

5.3.2 RAF-DB 数据集

图 5.3 RAF-DB 数据集

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

5.3.3 数据预处理

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

5.4 实验结果及其定性分析

5.4.1 单个 AU 变换结果

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

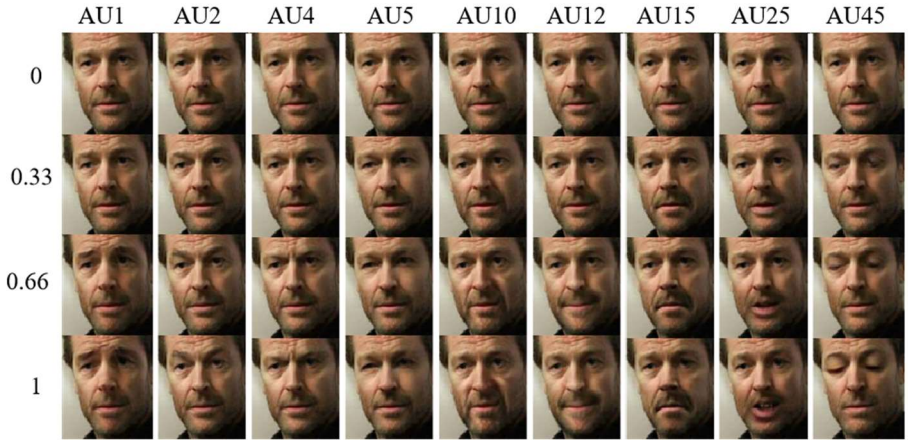


图 5.4 单个 AU 变换在 CelebA 数据集生成的结果

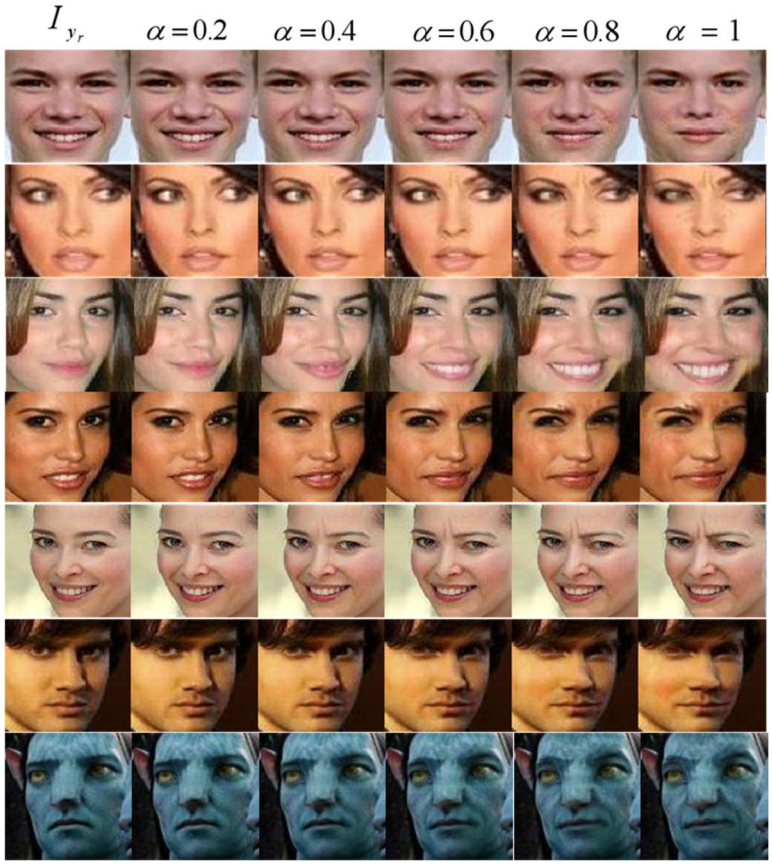


图 5.5 多个 AU 连续变换在 CelebA 数据集生成的结果

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

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

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

5.4.2 多个 AU 连续变换结果

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

5.4.3 多个 AU 离散变换结果

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

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

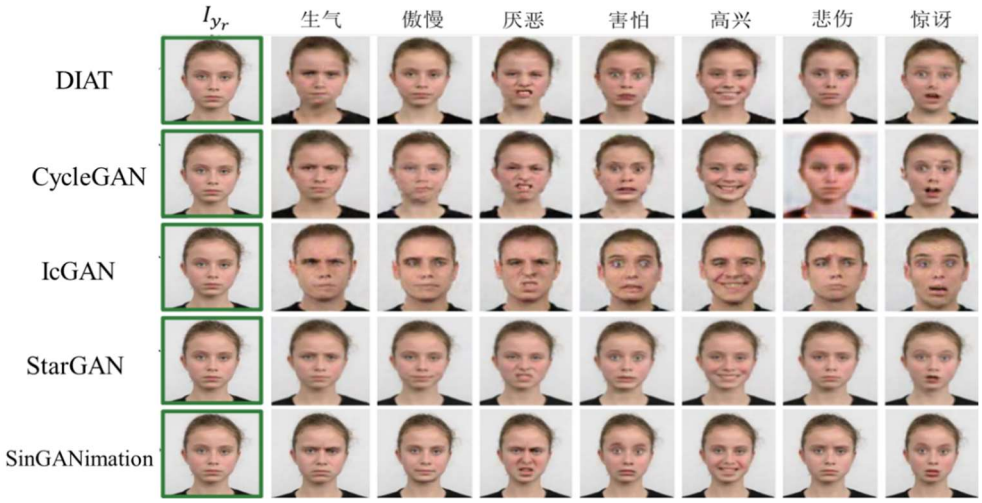


图 5.6 多个 AU 离散变换在 RAF-DB 数据集生成的结果

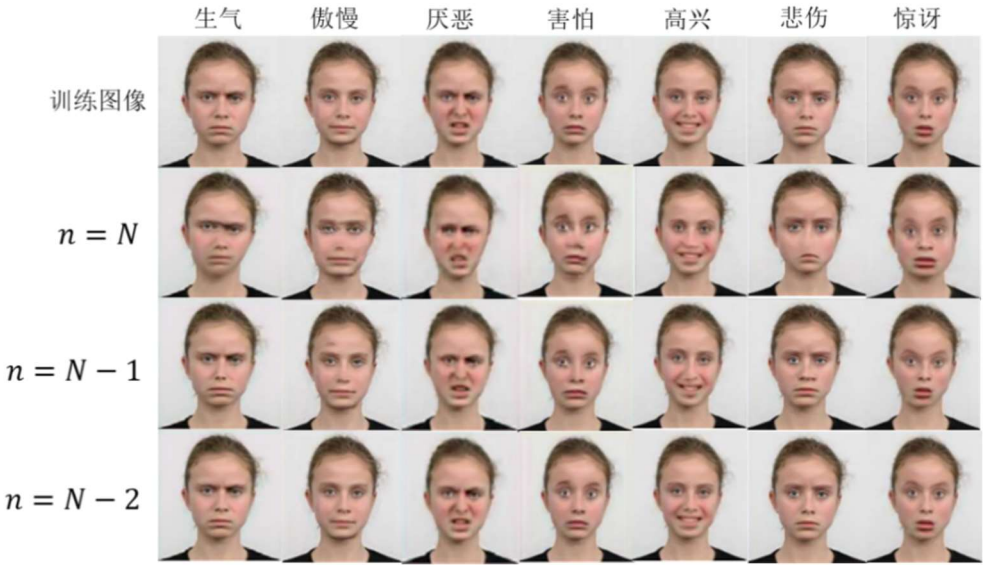


图 5.7 在 RAF-DB 数据集的再生成结果

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

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

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

5.4.4 SinGANimation 再生成结果

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

5.5 SinGANimation 实验结果定量分析

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

5.5.1 AMT 真伪用户测试

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

① 匹配(真伪对比)

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

② 非匹配(不区分真伪)

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

本次实验对两种情况都采用了两种生成方法：从较大尺度 $N-1$ 开始生成和从 $N-2$ 尺度开始生成，如图 5.7 所示。在这 4 个测试中，有 50 个不同的受试者。在所有测试中，前 10 个测试都是包含反馈的教程，具体的实验结果如表 5.1 所示。通过实验得出两个生成过程的混淆率：从最大尺度 $N-1$ 开始(生成具有较大多样性

表 5.1 真伪 AMT 测试结果

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

表 5.2 SIFID 结果

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

的样本)和从第二个最大尺度 $N-2$ 开始(保存原始图像的全局结构)。在每种情况下，都进行了匹配和非匹配测试，方差通过 Bootstrap[30]计算。从获得的实验结果可知，实验结果正如预期的那样，在非配对情况下，混淆率始终较大，即使改变了大型结构，这表明通过本文提出算法生成的图像也很难与真实图像区分开来(混淆率 50%意味着完全混淆真实图像和虚伪图像)。

5.5.2　单幅图像 FID 测量

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

实验中将 FID 指标应用于单个图像，并得出完全生成的 50 个图像的平均分。实验结果如表 5.2 所示。表 5.2 可以看出，AMT 结果的相关性表明，SIFID 与人为

测试结果一致。由 $N-2$ 尺度生成的 SIFID 平均值低于 $N-1$ 尺度生成的 SIFID 平均值，这与用户测试结果保持一致。该部分实验还说明了 SIFID 值与虚伪图像混淆率之间的相关性，两者之间存在明显的的反相关性，这意味着越小的 SIFID 通常对应较大的混淆率。匹配测试的相关性更强，因为 SIFID 是匹配的度量标准，其作用于匹配图像 x_n，$\tilde{x}_n$。

5.6 本章小结

本文提出的算法 SinGANimation 结合了 GANimation 和 SinGAN，在完全无监督的情况下，对单幅人脸图像进行学习训练，并且生成的表情图像种类多样，规模庞大，质量较高。在 SinGANimation 模型搭建过程中，通过将训练数据下采样输入，解决了 SinGAN 原有模型的人脸生成失真问题，进一步优化 SinGAN 模型，拓宽了 SinGAN 的适用范围。对于 SinGANimation 模型的性能评估使用了 CelebA 和 RAF-DB 两个数据集，前者为特定环境下拍摄的数据集，后者为在自然环境下拍摄的数据集，两者的光照条件，人脸特征等方面差距较大。在两个数据集上生成的图像均达到不错的效果，证明 SinGANimation 可以在不同数据集上应用，具有良好的鲁棒性。

由于现有的人脸表情数据库的图像种类和数量有限，而且例如深度学习等人脸表情识别需要大规模的训练数据，所以在人脸数据量较少的情况下，SinGANimation 模型可以对人脸表情图像进行种类和数量的扩充，满足实验的需求。SinGANimation 可生成连续的表情序列，所以在后续的工作中，可以应用于视频序列等工作中。由此可见，SinGANimation 具有非常重要的商用和科研实验价值。另外，SinGANimation 的再生成图像，与对应种类的图像差别较小。这可能是 SinGAN 内部学习训练的原因，在语义多样性方面具有一定的限制。训练 SinGANimation 模型耗时较长，算法时间复杂度较高，还待算法性能进一步优化。

第六章　总结与展望

6.1 工作总结

人脸表情生成算法与自然图像生成相比，在保证生成图像的画质清晰以外，还需要确保人脸的高度结构性。原有的 SinGAN 对自然图像的生成效果很好，但对人脸图像的效果差强人意。于是，本文对人脸图像的特性进行研究，提出了结合 SinGAN 和 GANimation 两种深度学习网络模型的 SinGANimation 模型，以改善生成效果。本文工作总结如下所示。

- 首先，创新设计出新型基于单幅图像生成人脸表情算法。本文将图像下采样后再输入，可以保证人脸结构的一致性，解决了 SinGAN 原有模型人脸生成失真的问题。即便这样，人脸表情的数量有所增加，但种类基本保持不变。因此，本文首次引入了 GANimation 到 SinGAN 模型中，构建出一种新的完全无监督表情生成算法 SinGANimation。在 SinGANimation 模型中，通过 GANimation 进行单个 AU 变换、多个 AU 连续变换、多个 AU 离散变换等操作，对图像的表情种类进行扩充，这样有效地实现了生成表情的数量和种类都有一定程度的增加。
- 此外，本文对算法进行了定性和定量分析。在实验中，通过与其他经典模型对比，发现本文提出的 SinGANimation 模型可以控制 AU 变换，既可以生成连续自然的表情，还可以生成离散情绪的图像。此外，本文进行了 AMT 真伪用户测试和单幅图像 FID 测量，得到的生成图像的深度特征分布与真实图像的深度特征分布之间的偏差分别为 0.09 和 0.05，混效率接近 50%，表明本文提出的 SinGANimation 模型生成的图像与真实图像高度相似。
- 同时，SinGANimation 算法的鲁棒性较强。算法在不同数据集上实验结果均达到不错的效果，由此得以验证 SinGANimation 算法具有良好的鲁棒性。

6.2　工作展望

虽然本文提出的人脸表情生成算法与原有算法相比较，在图像画质，表情种类和数量方面取得一定程度的进步，但是由于个人时间和能力有限，算法还存在多处不足，之后的学习研究会进一步改进。

- SinGANimation 算法的再生成图像，与对应种类的图像差别较小。这可能是 SinGAN 内部学习训练的原因，在语义多样性方面具有一定的限制。之后对 SinGAN 的内部网络结构做进一步的调整，以解决此问题。
- 训练 SinGANimation 模型耗时较长，算法时间复杂度较高，还待算法性能进一步优化。

致 谢

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

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

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

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

参考文献

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

西安电子科技大学本科生毕业设计（论文）
盲审意见书

专业名称	智能科学与技术
题　　目	基于单幅图像的面部表情生成算法研究

评审专家审阅意见

参照以下几个方面提出意见：

（1）论文选题：理论意义、学术价值、实用价值和经济效益，是否接触学科前沿；

（2）文献综述：对本领域文献资料是否有比较深入的了解；

（3）研究内容和成果：是否有新见解、新思想，或应用新技术解决实际问题；工作量和工作难度如何；

（4）能力水平：基础理论、专业知识、实际技能、设计能力和独立工作能力等方面的水平；

（5）论文规范和撰写质量：论文写作是否科学规范，文字通顺，图表齐全，字数符合规定要求，提炼及文字组织表达能力如何；

（6）外文翻译：表达正确，语句通顺，字符数达到要求。

评阅意见(注：评阅意见里无需显示盲审结果，结果在系统里选择)：

人脸表情识别算法一般需要大量的训练数据，而现有的数据库表情种类和数量较少。论文针对此问题展开研究，论文选题触及学科前沿理论，具有较高的学术研究价值。

论文围绕生成对抗网络（GAN）和 SinGAN 模型展开研究，将 GANimation SinGAN 两种生成模型相结合，设计了一种新的无监督表情生成算法 SinGANimation，并进行了仿真实验验证。结果表明本文提出的 SinGANimation 算法，可在完全无监督的情况下对单幅人脸图像进行学习训练，且生成的表情图像种类多样、规模庞大、质量较高。

毕业论文撰写格式规范，实验方案合理，并通过大量实验分析展现其设计算法性能，具有一定的创新性，但文献综述能力略欠缺。论文工作反映出作者掌握了较扎实的基础理论和专业知识，具有较强的算法设计能力、综合分析问题求解问题的能力。

综上，论文达到了申请学士学位的要求，推荐进行论文答辩。

问题与建议：

（1）各个表格中的论文起止时间请按学校要求统一。

（2）论文结构安排缺乏国内外研究现状。

（3）图表中标题及文字内容按五号字体排版。

（4）建议修改标题：第三章 SinGAN—》第三章 SinGAN 模型及其应用；第四章 GANimation---》

第四章 GANimation 模型及其应用；第五章 SinGANimation 表情生成算法实验---》第五章 SinGANimation 算法仿真实验与分析

西安电子科技大学本科生毕业设计（论文）

盲审意见书

专业名称	智能科学与技术
题　　目	基于单幅图像的面部表情生成算法研究

评审专家审阅意见

参照以下几个方面提出意见：

（1）论文选题：理论意义、学术价值、实用价值和经济效益，是否接触学科前沿；

（2）文献综述：对本领域文献资料是否有比较深入的了解；

（3）研究内容和成果：是否有新见解、新思想，或应用新技术解决实际问题；工作量和工作难度如何；

（4）能力水平：基础理论、专业知识、实际技能、设计能力和独立工作能力等方面的水平；

（5）论文规范和撰写质量：论文写作是否科学规范，文字通顺，图表齐全，字数符合规定要求，提炼及文字组织表达能力如何；

（6）外文翻译：表达正确，语句通顺，字符数达到要求。

评阅意见(注：评阅意见里无需显示盲审结果，结果在系统里选择)：

论文基于单幅的人脸表情数据下，生成了更多更自然的表情数据，并将其用于人脸表情识别算法。选题属学科前沿，主要工作包括理解 GANimation 和 SinGAN 两种深度网络，并结合二者设计新的无监督表情生成算法 SinGANimation，实现了基于单幅表情图像的面部表情数据生成过程，并与相关算法做了性能比较,工作量饱满，论文撰写认真。

目前存在个别问题:

1.文中对文献作者名的应用需要规范。

2.参考文献个别页码缺失,会议论文也需要页码.期刊文献格式不统一。

文本复制检测报告单

№:ADBD2020R_20190429134059202006011212243041200864 43

检测文献： ff898918fe30412fa1d9a7121b99c585.pdf

作者：

检测范围： 中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库

时间范围： 1900-01-01至2020-06-01

检测时间： 2020-06-01 12:12:24

总文字复制比：4%

去除引用：4% 去除本人：4% 重合字数：1133 文献总字数：28229

总段落数： [3] 疑似段落数： [2] 疑似段落最大重合字数： [916]

前部重合字数： [365] 后部重合字数： [768] 疑似段落最小重合字数： [217]

8.6% ff898918fe30412fa1d9a7121b99c585.pdf 第1部分(总10592字)

0% ff898918fe30412fa1d9a7121b99c585.pdf 第2部分(总9940字)

2.8% ff898918fe30412fa1d9a7121b99c585.pdf 第3部分(总7697字)

ff898918fe30412fa1d9a7121b99c585.pdf_第1部分

总文字复制比：8.6%（916） 总字数：10592

1	基于深度学习的图像语义分割研究 肖旭(导师：储珺) - 《南昌航空大学硕士论文》- 2017-06-01	4.4% 是否引用：否
2	多源遥感数据测绘应用关键技术研究 芮杰(导师：王超;张红) - 《中国科学院大学(中国科学院遥感与数字地球研究所)博士论文》- 2017-05-01	2.5% 是否引用：否
3	人脸表情识别的研究进展 蒋斌;贾克斌;杨国胜; - 《计算机科学》- 2011-04-15	1.0% 是否引用：否
4	学科交叉视角下的情感识别研究进展 程静;刘光远; - 《计算机科学》- 2012-05-15	1.0% 是否引用：否
5	基于HOG特征和SVM的人脸表情识别 王阳;穆国旺;睢佰龙; - 《河北工业大学学报》- 2013-12-15	1.0% 是否引用：否
6	多媒体对联辅助生成系统的设计与实现 郭燕华(导师：潘志庚) - 《浙江大学硕士论文》- 2012-04-01	0.8% 是否引用：否
7	3D打印跑车、人脸测谎“表情大作战”……这些“黑科技”亮瞎你的眼 - 《重庆商报》- 2019-08-29	0.8% 是否引用：否
8	生成对抗网络GAN综述 程显毅;谢璐;朱建新;胡彬;施佺; - 《计算机科学》- 2019-03-15	0.7% 是否引用：否
9	基于卷积神经网络的目标跟踪算法研究 赵银妹(导师：胡硕;丁栋) - 《燕山大学硕士论文》- 2018-05-01	0.6% 是否引用：否
10	工厂物品自动分拣视觉算法研究	0.5%

	张超(导师：赵林辉) - 《哈尔滨工业大学硕士论文》- 2019-06-01	是否引用：否
11	生成对抗网络理论框架、衍生模型与应用最新进展 赵增顺;高寒旭;孙骞;滕升华;常发亮;Dapeng Oliver Wu; - 《小型微型计算机系统》- 2018-12-11	0.5% 是否引用：否
12	基于GAN的数据生成模型研究 肖睿(导师：陈亦欧) - 《电子科技大学硕士论文》- 2019-04-11	0.5% 是否引用：否
13	运动捕获技术在三维动画角色面部表情设计的运用 张漫宇; - 《天津美术学院学报》- 2013-12-25	0.5% 是否引用：否
14	双相障碍患者面部表情识别的脑功能研究现状 黄世伟;刘铁榜;荣晗; - 《临床精神医学杂志》- 2016-02-20	0.5% 是否引用：否
15	基于条件生成对抗网络的室内场景布局估计 曹丹丹(导师：刘天亮) - 《南京邮电大学硕士论文》- 2019-12-09	0.4% 是否引用：否
16	基于特征转换学习的卡口车牌检测识别方法 陈曙(导师：谢雪梅;李青) - 《西安电子科技大学硕士论文》- 2019-06-01	0.3% 是否引用：否
17	行人重识别统一评测框架研究与实现 吴婉银(导师：王丽珍) - 《云南大学硕士论文》- 2019-03-01	0.3% 是否引用：否

摘要

摘要

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

受启发于不同生成模型的结构和特性，本文将 GANimation 和 SinGAN 两种生成模型相结合，设计了一种新的无监督表情生成算法 SinGANimation，实现了基于

单幅表情图像的面部表情数据生成过程。该算法首先通过 GANimation 进行单个AU 变换、多个 AU 连续变换、多个 AU 离散变换等操作，对图像的表情种类进行扩充，得到初步结果再输入 SinGAN，进行再生成操作，增加图像的数量。其中，

本文提出的算法将图像下采样后再输入，解决了 SinGAN 原有模型人脸生成失真的问题，保证了人脸的高度结构性。随后，本文对提出算法的生成结果进行了定性和定量分析。通过与其他经典模型对比，发现本文提出的算法既可以生成连续自然的表情也可以生成离散情绪的表情，其画质更加真实清晰。在多个数据集上训练，

均达到不错的效果，证明算法的鲁棒性良好。此外，测试实验中也进行了 AMT 真伪用户测试和单幅图像 FID 测量，得到的混淆率接近 50%，生成图像与真实图像的深度特征分布之间的偏差接近 0.1，表明两种图像高度相似。最后，本文对提出的算法模型的优缺点进行分析，计划将算法应用于扩充人脸表情数据库，视频序列等商业和科研工作中。

关键词：单幅图像表情生成 GANimation SinGAN

摘要

ABSTRACT

ABSTRACT

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

Inspired by the structure and characteristic of various generation models, this paper combines th e GANimation and SinGAN models together to develop a new fully unsupervised generation algorithm expr ession, called SinGANimation, which achieves the generation of facial expression images via utilizing only one expression image. The proposed method operates the single AU transformation, multiple AU con tinuous transformation, multiple AU discrete transformation of GANimation, et al, and it expands the expression types of the image and inputs the results to SinGAN for regeneration operation to increase the number of images. In order to solve the problem that generated face may be distorted by SinGAN, t he proposed method adds the downsampling strategy for input images, which effectively improves the pr oblem of face distortion in SinGAN and ensures the high facial structure information. Then, this pape

r makes qualitative and quantitative analyses for the generated results obtainted by the proposed method.

Comparing with other classical generation models, it is found that the proposed method can generate both continuous, natural expressions and discrete emotional expressions, and the quality of pictures are more real and clear. The training on multiple data sets has achieved good results and proved the robustness of the algorithm. Additionally, this paper conducts the AMT true and false user test and FID measurement of a single image. The confusion rate are close to 50%. The deviation between the depth feature distribution of the generated image and the real image are close to 0.1, which indicates that the two images are highly similar. Finally, this paper analyzes the advantages and disadvantages of the algorithm, and plan to apply the algorithm to expand the facial expression database, video sequence and other commercial and scientific works.

Key words: single image facial expression generation GANimation SinGAN

ABSTRACT

目录 i

目录

第一章绪论 1

第一章绪论

1.1 人脸表情概述人的面部表情在社交中极为重要。通常，交流涉及言语和非言语。非语言交流是指人与动物之间通过眼神交流，手势，面部表情，肢体语言和非语言进行的交流。

非语言交流是通过面部表情表达的。面部表情是更大范围交流的微妙信号，它能够有效地传达非语言信息及情感的交流，从而辅助听者推断说话人的意图。

人脸表情是传播人类情感信息与协调人际关系的重要方式，据心理学家A.Mehrabia 的研究表明[

1]，在人类的日常交流中，通过语言传递的信息仅占信息总

量的 7%，而通过人脸表情传递的信息却达到信息总量的 55%。尤其，当说话人在试图掩盖内在情绪时，面部表情的细微变化是无法隐藏和无法抑制的，它所传达的信息暗含了潜在的个体行为信息，是与人类情感、精神状态、健康状态等诸多因素相关的一种复杂的表达方式。

在 20 世纪，Ekman 和 Friesen 对人脸表情进行研究[

2]，得出人类的七个基本情

感:幸福，惊讶，愤怒，悲伤，恐惧，厌恶和中立。他们建立了不同种类表情的人类面部表情数据库，详细记录每种表情的面部变化，比如皱鼻、嘴角拉伸等动作变化，

这便是 1976 年创造的“面部运动编码系统”（FACS，Facial Action Coding System）。

FACS 包含 44 个面部动作单元（AU，Action Unit），例如抬高眉毛、眼睛变化等用作描述人面部局部表情的变化。AU 可以精确细致的描述人的面部表情，但其标注的成本高，耗时长，例如：标注一个人眼部 AU，需要标注员长达 30 分钟的时间。所以，现在的人脸表情数据库的采集对象和面部表情都相对有限。

1.2 研究意义与目的如今，人脸识别技术发展迅猛，应用市场和用户需求大，而人脸面部表情识别作为人脸识别技术中的一个重要组成部分，对公共安全、安全驾驶、智能医疗、测谎技术、智慧课堂等领域具有非常重要的商业贡献，对学术界也有很大的研究意义。

比如，在安全驾驶中，通过识别司机的眼部表情，判断司机是否为疲倦状态，若是便发出安全警告，减少安全隐患；在智能医疗场景中，根据患者的微表情，评估患者的精神健康状态； 在智慧课堂上，老师可以根据智能摄像头采集学生们的面部

2 基于单幅图像的面部表情生成算法研究

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

如上文所述，现有的人脸表情数据库有以下两个不足：第一，因表情均为人为收集，并非真实环境下的自然表情，而且面部表情受许多因素的限制，例如年龄，

性别，肤色等，故致表情种类单一，幅度夸张。第二，表情数据量小，难以满足人脸表情识别算法的的数据需求。如今的深度学习发展迅猛，许多人脸识别算法也纷纷采用深度学习框架。深度学习训练模型需要大规模的数据，而现在的表情数据库容量有限，不足以支撑基于深度学习框架的人脸表情识别算法。

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

1.3 内容安排

第一章为绪论部分，主要介绍面部表情的相关知识和本文的研究意义与目的。

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

型。

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

优缺点。

第四章从 GANimation 相关基础，模型架构和方法以及优缺点分析等几个部分

进行描述，为下章的实验奠定理论基础。

第五章主要讲解基于 SinGAN 和 GANimation 的 SinGANimation 面部表情生

成算法实验，得出了多种实验结果，并将此模型与其他模型对比，进行定性和定量分析。

第六章是本文的总结与展望。对本文已做的工作进行梳理总结，并提出该算法

之后需要改进优化之处。

第二章生成对抗网络（GAN） 3

第二章生成对抗网络（GAN）

生成对抗网络，简称 GAN，是一种使用深度学习进行生成式建模的方法。所以，本章在介绍生成对抗网络模型的同时，需提及其前身生成建模，便于读者理解，

深入浅出地介绍生成对抗网络的理论基础及其延伸拓展模型。

2.1 生成对抗网络的理论基础

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

（a）无监督学习 （b）有监督学习图 2.1 无监督学习与有监督学习流程图

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

4 基于单幅图像的面部表情生成算法研究 (|) ( ) argmax (|) arg max ( ) argmax (|) ( ) y y y p x y p y p y x p x p x y p y （2-1）

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

2.1.2 生成对抗网络模型生成对抗网络（GAN）是基于深度学习的生成模型。一般而言，GAN 是用于

训练生成模型的模型架构，最常见的是在该架构中使用深度学习模型。2014 年,Ian

Goodfellow 等人首次给出了 GAN 架构[

3]。GAN 模型结构包括两个子模型：生成网

络G （Generator）和判别网络 D （Discriminator），它们的功能如下所示：

生成网络G 负责接收一个随机的多维向量噪声 z ，由其生成的图像为 z G ()。

判别网络 D 负责判别输入图像的真伪。假设m 为输入图像，则输出 D)m(是输入图像m 为真的概率。若 x D)(为 1，则m 为真实图像；若 D)m(为 0，则m

为不真实图像。

在 GAN 的训练过程中，生成网络G 试图生成真实的图像来欺骗判别网络 D ，

判别网络 D 试图将生成网络G 生成的图像和真实的图像区分开来。如此，两者便形成了动态的零和博弈。当 GAN 训练到最好状态时，生成图像G（z）足以骗过 D ，

而 D 难以分辨生成图像 z G () 的真伪，所以，生成图像为真实图像的概率D（G（z）） 0.5。GAN 的数学原理为

)))]((1[log()]([log),(maxmin) (~)(~

DVEx DGz GE Dz pzxpx

GDz data （2-2）

式中 x 表示真实图像，z 表示多维向量噪声，而 z G ()表示生成网络G 的生成图像。

判别函数 x D () 表示判别网络 D 鉴别真实图像的概率，其值接近或等于 1。而Dz G） （)(为判别网络 D 鉴别生成图像 z G ()为真实的概率。

生成网络G 的目标为 Dz G） （)(最大化，此时 GDV),(会变小，所以对于G 求最

小(

G min)。判别网络 D 的目标为 x D ()最大化， Dz G） （)(最小化，此时 GDV),(会

变大，所以对于 D 求最大（

D max ）。具体 GAN 的训练过程，如图 2.2 所示。

图 2.2 GAN 训练过程流程图图 2.3 DCGAN 的生成网络结构图

2.2 生成对抗网络的常用模型

2.2.1 深度卷积生成对抗网络（DCGAN）

深度卷积生成对抗网络[4]

（DCGAN）是由 Alec Radford 等人提出的 GAN 模

型，该网络能很有效地将监督学习中的 CNN 和无监督学习中的 GAN 结合在一起。

DCGAN 可跨一系列数据集进行稳定的训练，并允许训练更高分辨率和更深入的生成模型。DCGAN 的生成网络结构如图 2.3 所示：

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

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

此模型对 GAN 的体系结构进行约束，可以通过稳定地训练使其更趋于收

敛。

DCGAN 训练大量没有标签的图像时，特征提取非常有效，这既来自于生成网络也有判别网络（主要是判别网络）。由于 DCGAN 出色的特征提取，

它可用于更高级别的监督任务，例如图像分类。

对 GAN 学习到的 filter 进行了定性的分析。

DCGAN 的生成网络具有很好的矢量计算特性，可以轻松操纵生成样本的许多语义质量。

2.2.2 条件生成对抗网络（CGAN）

条件生成对抗网络[

5]（CGAN）是由 Goodfellow Ian 等人提出的一种带有条件

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

))]|((1[log()]|([log),(maxmin) (~)(~

DVEyx DGyz GE Dz pzxpx

GDz data （2-3）

2.2.3 循环生成对抗网络（CycleGAN）

一般的 GAN 面向一个域的数据，而循环生成对抗网络 CycleGAN[

6]实现的是

两个域的数据迁移。CycleGAN 是一个 A→B 单向 GAN 加上一个 B→A 单向 GAN。

两个 GAN 共享两个生成网络，然后各自带一个判别网络，所以加起来总共有两个判别网络和两个生成网络。CycleGAN 本质上是两个镜像对称的 GAN，构成了一个环形网络。

图 2.4 CGAN 网络结构图图 2.5 CycleGAN 网络结构图

如图 2.5 所示，真图像 A 经过生成网络A

B

G 表示为假图像 B，把假图像 B 视作

为真图像 B；同理可得，真图像 B 经过生成网络B

A

G 表示为假图像 A，把假图像 A

视作为真图像 A。

一个单向 GAN 有两个损失函数，而 CycleGAN 加起来总共有四个损失函数。

对于判别网络 A：

)1log()(log 2

DELEx Dx）（D

PAPx AxA

BAA （2-4）

对于判别网络 B：

)1log()(log 2

DELEx Dx）（D

PBPx BxB

ABB （2-5）

对于生成网络 BA：

))(()(log 2

EEx DLGx GxG

PABBAPx AxA

ABBA （2-6）

对于生成网络 AB：

))(()(log 2

EEx DLGx GxG

PBAABPx BxB

BAAB （2-7）

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

第三章 SinGAN 模型及其应用 9

第三章 SinGAN 模型及其应用

SinGAN：Learning a Generative Model from a Single Natural Image[

7]，即从单张

自然图像中学习的生成模型。此模型通过使用一种专门的多尺度对抗训练方案，对多个尺度上学习子图像块数据。然后，它可以用来生成新的逼真的图像样本，在创建新的对象配置和结构时，保持原始的子图像块的分布。本章将主要介绍 SinGAN

模型的基本原理，模型细节及该模型的优缺点和部分实验结果。

3.1 SinGAN 相关基础

3.1.1 单项深度模型现有的几项研究提出将深度模型过度拟合到单个训练实例中，然而，这些方法是为特定的任务而设计的(如：超分辨率，纹理扩展等)。Shocher 等人[

8]首先为单个

自然图像引入了基于内部GAN的模型，并在重新定向的背景下进行了说明。然而，

它们的生成取决于输入图像，即将图像映射到图像，而不是用来绘制随机样本。相比之下，SinGAN 的框架是纯生成的，即将噪声映射到图像样本，因此适合许多不同的图像处理任务。

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

3.1.2 图像处理的生成模型在许多不同的图像处理任务，基于 GAN 模型的研究已经证明了对抗性学习的能力，例如：交互式的图像编辑和其他图与图之间的翻译任务。然而，已有的方法大部分都是在具体的数据集上训练，将生成条件设置为另一个输入信号，SinGAN

也是如此。SinGAN 并不着重于提取一般的同类图像特征，而是通过不同来源的训练数据——单幅自然图像的多尺度的全部重叠图像子块。SinGAN 展示了一个强大

10 基于单幅图像的面部表情生成算法研究

图 3.1 SinGAN 与单个图像纹理生成的生成模型是可以从上述训练数据中学习，并用于多种图像处理任务，下面将详细介绍 SinGAN 模型的基本原理及其应用。

3.2 SinGAN 模型的基本原理

3.2.1 概述SinGAN 模型的主要目的是学习一个无条件生成模型，它可以捕获单个训练图像的内部统计信息。这个任务在概念上与传统的 GAN 设置类似，只是这里的训练

样本是单个图像的子块，而不是来自数据库的整个图像样本。

SinGAN 选择超越纹理生成，并处理更综合的自然图像。这需要捕捉在很多不同尺度下的复杂图像结构分布。例如，SinGAN 想要捕获全局属性，比如图像中大型物体的排列和形状(顶部的天空，底部的地面)，以及图像细节和纹理信息。为了

实现这一目标，SinGAN 的生成式结构，如下图 3.2 所示，包含一个多层次的子块-

GANs（马尔可夫链的判别网络)[

9],其中每个负责捕捉不同规模 x 的子块分布。GANs

的感受野比较小，而且容量有限，这些特点阻止它们记忆单一的图像. 同时，相似的多尺度的架构一直在探索传统 GAN 设置，SinGAN 是第一个从单一图像内部学习探索它的网络模型。

第三章 SinGAN 模型及其应用 11

图 3.2 SinGAN 的多尺度传递途径SinGAN 模型是由一个 GANs 的金字塔结构组成，其中训练和传递都是尺度由

大到小的方式完成。在每个尺度，生成网络nG 学习生成图像样本，判别网络 n

D 无法将生成图像的所有重叠图像块与降采样训练图像中的图像块nx 区分开来。当沿着金字塔向上移动时，有效的图像块尺寸不断减小(在原始图像中用黄色标记以供

说明)。输入到生成网络nG 是随机噪声图像nz 。对之前尺寸nx~生成的图像进行上采

样至当前的分辨率（除了纯生成的最大尺度）。当前尺度的生成过程涉及到所有生

成网络 }G,,{Gn

N 和噪声图谱 }z,,{zn N 的参与。

3.2.2 多尺度结构SinGAN 的模型由一个金字塔状生成网络 }G,,{GN

0 组成，对 x 的图像金字塔

},,{xO

N x 进行训练，其中当r 1，nx 是一个因子nr 的 x 的下采样版本。每个生成网络负责生成真实的图像样本，即关于对应图像中的子块分布。通过对抗训练，实

现nG 学习，欺骗相关的判别网络 n

D ，判别网络nD 试图将生成样本中的子块与 n x

中的子块区分开来。

通常，图像样本的生成从最大的尺度开始，依次通过所有生成网络，直到最小的尺度，并在每个尺度都输入噪声。所有的生成网络和判别网络都有相同的感受野，

因此在生成过程中捕获的结构尺寸都在减小。在最大尺度上，生成结果是纯生成的，

即NG 将空间高斯白噪声 n z 映射到图像样本Nx ，即

)(~N

NN x z G （3-1）

图 3.3 单规模迭代一般而言，这一层的有效感受野一般是图像高度的 1/2，因此NG 可以生成图像的总体布局和对象的全局结构。每个生成网络在更小的尺度（ Nn ）上添加之前

尺度没有生成的细节。因此，除了空间噪声nz 外，每个生成网络 n

G 还接受较大尺度图像的上采样版本，即 1

(,( )rn n n n x G z x n N)， （3-2）

实际上，所有的生成网络都具有类似的架构，如图 3.3 所示。在被输入到一系列卷积层之前，噪声nz 要被加在图像rnx)~(1。这确保了 GAN 不会忽略噪声，如同随机条件方案中经常发生的情况。

卷积层的作用是生成缺失的细节rnx)~(1（残差学习）[

10]，即nG 执行如下操作：))~(()~(~1 1 r nnn r nn xzxx （3-3）

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

在每个尺度n 上，对之前尺度的图像1~ n x 向上采样并输入噪声图谱nz 中。其结果输入至 5 个卷积层，输出是一个补充到 rnx 1~的残差图像，即生成网络nG 的输出nx~。

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

不同的输入规模导致不同的效果。

图 3.4 SinGAN 模型的应用展示

3.3.1 超分辨率SinGAN 将输入图像的分辨率提高了 s 。SinGAN 在低分辨率(LR)图像上训练模型,得出重构损失权重为 100和生成网络金字塔的比例因子r k s，k N 。在

自然场景不同尺度中，小型结构往往反复出现，因此在测试时，SinGAN 通过一个 r 因子对 LR 图像进行上采样，并将其连同噪声输入最后一个生成器OG 。SinGAN

重复k 次以获得最终的高分辨率输出，示例结果如图 3.5 所示。从对比结果可以看出，SinGAN 重建的视觉质量超过了目前最先进的内部生成方法，也超过了以最大信噪比为目标的外部生成方法。SinGAN 尽管只需要一张图像，但结果可以与外部训练的 SRGAN[

11]方法相媲美。在 BSD100 数据集

[12]上，基于失真程度(RMSE)和感知质量(NIQE[

13])两个指标比较了 5 种方法的性能，结果展示在表 3.1 中，注：这

表 3.1 超分辨率对比方法性能指标

外部训练方法内部训练方法

SRGAN EDSR DIP ZSSR SinGAN

RMSE 16.34 12.29 13.82 13.08 16.22

NIQE 3.41 6.50 6.35 7.13 3.71

图 3.5 超分辨率效果对比。

图 3.6 画风迁移效果对比两个指标本质上是相互冲突的。从表 3.1 中所示的结果可以看出：SinGAN 擅长感知，其 NIQE 数值仅略低于 SRGAN，但 RMSE 数值要高于 SRGAN。

另外，当 SinGAN 被训练在一个低分辨率的图像上时，可以进行超分辨率操作。这是通过不断迭代对图像进行采样，并将其输入到 SinGAN 的最小尺度的生成

网络来实现的。可见，SinGAN 的视觉质量优于 SOTA 的内部训练方法 ZSSR 和DIP，也与在大规模集合上进行外部训练的 SRGAN 方法的训练结果相近。括号中显示了相应的 PSNR 和 NIQE 的数值。

3.3.2 图画到图像的画风迁移图画到图像的画风迁移即将剪贴画转换成逼真的图像。这是通过对剪贴画图像向下采样，并将其输入至一个较大尺度(例如 N 1或 N 2)的生成网络来实现的。

从图 3.6 可以看出，SinGAN 保留了画面的整体结构，真实地生成了与原图匹配的

图 3.7 图像调和效果对比

纹理和高频信息。SinGAN 的画风迁移结果在视觉质量上要优于风格迁移（Syle

Transfer）方法。在目标图像上训练 SinGAN，并在测试时将下采样的图画输入到较大生成网络中，生成图像保留了剪贴画的布局和一般结构，同时生成与训练图像匹配的真实纹理和精密细节。

3.3.3 图像调和

图像调和为将粘贴对象与背景图像真实地混合在一起。在背景图像上训练SinGAN，并在测试时输入原始粘贴合成的下采样样本。SinGAN 将生成图像与原始背景相结合。从图 3.7 可以看出，SinGAN 模型对粘贴对象的纹理进行了裁剪以匹配背景，并且与其他图像调和方法相比，更好地保留了对象的结构。在 2、3、4

尺度下，粘贴对象的结构和转移背景纹理之间可以取得很好的平衡。SinGAN 模型能够保持粘贴对象的结构，同时调整其外观和纹理，而其他的协调方法过度混合对象与背景。

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

图 3.8 图像编辑效果对比图 3.9 人脸图像训练 SinGAN 模型的生成效果

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

形前进，从而将单一图像合成运动视频。对于许多类型的图像，真实效果是通过 z

空间中的随机漫步实现的，即在所有生成尺度中，第一帧画面由r ecz 开始生成。

3.4 本章小结SinGAN 模型是首次使用单张自然图像训练、非条件的生成式模型。SinGAN

模型生成的效果目前已经可以做到以假乱真，它可以生成新的具有真实感的图像样本，在保留了原始的图像块分布的基础上，创造了新的物体外形和结构。SinGAN

模型具有超越纹理和生成自然复杂图像的各种真实样本的能力，为广泛的图像处理任务提供非常强大的工具。

然而，SinGAN 模型也存在一定的局限性，这可能源于该模型是“单张图像训练”的设定，具体表现为：第一，当图像块差异较大时，容易产生不真实的现象，

无法学到很好的分布。如图 3.9 所示，如果直接使用人脸图像训练 SinGAN，生成的图像失真严重。这个问题也是本文针对面部图像生成对 SinGAN 进行改进的出

发点，目的是生成无失真更真实的人脸表情。第二，与外部训练生成方法相比，

SinGAN 经过内部学习生成图像的内容语义多样性受到了限制，例如：如果训练图像是一只猫，模型不会生成不同猫品种的样本。

第四章 GANimation 模型及其应用

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

4.1 GANimation 模型的相关基础

4.1.1 非匹配的图像转换在 GANimation 框架中，一些工作解决了使用非匹配训练数据的问题。在图像个别领域的边缘分布中，首次尝试应用依赖马尔科夫随机场先验的贝叶斯生成模型。其他模型则探索了利用变分自动编码器策略来增强 GANs。后来，一些模型应

用了驱动系统生成变换样式映射的思想，而且没有改变原始输入图像内容。

GANimation 方法更接近于那些利用循环一致性来保存输入和映射图像之间的关键特征的模型，比如 CycleGAN[

6]、DiscoGAN

[14]和 StarGAN [15]。

4.1.2 面部图像处理人脸生成与编辑是计算机视觉和生成模型研究的热点。大多数的工作都是处理属性编辑的任务，试图修改诸如添加眼镜、改变头发颜色、性别交换和老化等属性类别。这些工作与 GANimation 最相关的是面部表情的合成。早期的方法是使用质量-弹簧模型来模拟皮肤和肌肉运动[

16]。这种方法的问题是很难产生自然的面部

表情，因为有许多细微的皮肤运动是很难用简单的弹簧模型渲染的。另一种思路是依赖于 2D 和 3D 的形态[

17]，但在区域边界周围产生了强大的伪影，无法模拟光照

变化。

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

图 4.1 常见 AU 类别表示模型恢复了皮肤和肌肉建模的想法，但将其整合到现代深度学习机制中。更具体地说，GANimation 学习了一个基于肌肉运动的连续嵌入 GAN 模型，允许在视频序列中生成大量基于人脸结构的面部表情以及平滑的面部运动转换。

4.1.3 人脸动作单元（AU）

人脸动作单元（AU）源于面部动作编码系统（FACS），是一种基于面部表情对人类面部运动进行分类的系统，而 AU 是人脸单个肌肉或一组肌肉的基本动作。

人脸做出表情时，面部区域会有不同程度的变化，即多种 AU 会有一定的强度变化。常见的 AU 包括内侧眉头上扬，上眼睑上扬，嘴唇提起等，具体 AU 类别表示如图 4.1 所示。GANimation 数据集的标签便是基于 AU，使用表情向量来表示面部各区域不同程度的变化。通过调节表情向量使得 GANimation 模型输出不同程度的表情。表情向量如下所示： 1 2

, , ,Tr

N y y y y （4-1）

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

4.2 GANimation 模型架构和方法

4.2.1 待解决的问题

定义一个在任意面部表情下捕获的输入 RGB 图像为3 r

H W yI 。每个表情表

达式都由 N 个动作单元 1, , r N y y y 决定，其中每个ny 表示第n 个动作单元大小的归一化值，其值范围为从 0 到 1。由于这种连续的表现，自然插值可以在不同

的表情之间，渲染的范围更加真实，面部表情更加光滑。

GANimation 的目标是学习一个映射 ，将ryI 转换成输出图像gyI 条件下的动作单元目标gy ，例如：映射为：: , r gg y y I y I 。为此，GANimation 对进行

无监督训练，并借训练三元向量组 1, ,rMm m mr g m y

I y y ，其中目标向量mgy 随机生成。

GANimation 既不需要同一个人在不同表情下的成对图像，也不需要期望的目标图

像 gy

I 。

如图 4.2 所示，该网络结构由两个主要部分组成:一个用于回归注意力的生成网络G 和颜色掩膜;判别网络要对生成图像的真实性ID 和表情条件完成度 gy 进

行评估。需要说明的是，GANimation 是无监督的，即同一个人不同表情的图像对

和目标图像 gy

I 都假设是未知的。

图 4.2 GANimation 模型的结构D

4.2.2 网络结构设G 为生成网络块，因为它是双向应用的(例如，将任一输入图像映射到所需表情，反之亦然)，在下文中，将使用下标 o 和 f 来表示起点和终点。给定图像 3 o

H W yI 和编码所需的表达式 N 维向量fy ，将生成器的输入作为一组串联

，oo yI yH W (N 3) ，其中oy 表示为大小为的 N 个数组。

图 4.3 基于注意力机制的生成网络GANimation 系统的一个关键组成部分是使G 只专注于图像中那些负责合成新表情的区域，并且保持图像中其他元素，如头发、眼镜、帽子或珠宝等非表情元

素不受影响。为此，GANimation 在生成网络中嵌入了注意力机制。具体来说，

GANimation 生成网络输出两个掩膜，一个颜色掩膜C 和一个注意力掩膜 A ，而不是回归一个完整的图像。最终图像的获得方式如下： O (1) f y y

I A C A I （4-2）

其中， O| {0, ,1}H WA fG yA I y 和 3|oH WC y fG C I y 。掩膜A表示扩展C

的每个像素并对输出图像有贡献。通过这种方式，生成网络不需要渲染静态元素，只关注定义面部运动的像素，从而生成更清晰、更真实的合成图像。此过程如图 4.3 所示。在整个图像上，给定了输入图像，目标表情，生成网络回归表达式和注意力掩膜 A 和 RGB 颜色转换掩膜C 。注意力掩膜定义了每个像素的强度，确定了将原始图像每个像素的扩展程度，并将在最终呈现的图像中起作用。

条件化判别网络，即以生成图像真实性和期望表情完成度作为评价标准的判别网络。 D(I) 的结构类似于 PatchGan 网络[

18]，从输入图像 I 映射到一个矩阵

H W fy

I

I

H WY ，其中IY[i, j]表示重叠图像块ij 为真实的概率。此外，为了评估其条件作用，在网络顶端添加副回归项首部，来估计在图像中 AUs 的激活函数 1 ， ，

N y y y 。

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

下面将给出上述损失函数的详细信息：

① 图像对抗损失函数为了了解生成网络G 的参数，GANimation 使用了 WGAN-GP 提出的标准 GAN

算法的修正版本。具体来说，原始的 GAN 公式是基于 Jensen-Shannon (JS)散度损失函数，其目的是最大化真实图像的分类正确概率和当生成网络欺骗判别网络时，

对图像进行渲染。这种损失可能不是连续的生成网络参数，而且局部饱和会导致判别网络中的梯度消失。通过WGAN[

20]替换连续地球移动距离的 JS 函数，可以解决

此类问题。为了保持 Lipschitz 约束，WGAN-GP 为判别网络添加一个梯度惩罚作为判别网络输入的梯度范数。

令 oy

I 作为初始条件oy 的输入图像，fy 为期望的最终条件，o 为输入图像的数据分布，I 为随机插值分布。然后，判别损失 I I, , ,ofG Dy I y 为：

2~ I ~ I gp ~I2| ( ) lo o o oo o If I ID G D D I y yI y I yI y I （4-3）

其中，g p

为惩罚系数。

② 注意力损失函数在训练模型时，与颜色掩膜C 类似，没有对注意力掩膜 A 进行 ground-truth 注释，而从判别模块的结果梯度和其他损失函数中学习的。然而，注意力掩膜很容易

饱和到 1，这使得 | o of Gy yI I y ，也就是说，生成网络没有起效。为了防止这种情况，GANimation 用一个2l 权重惩罚系数来调整掩膜。同时，为了在将输入图像像素与颜色变换C 相结合时，进行平滑的空间颜色变换，GANimation 对 A 进行全变差正则化。因此，注意力损失可以定义为：

,2 2TV ~ 1, , , 1 , ~ 2 , o oo o

H W i j i j i j i j i j y yI IA A A A ‖A‖ （4-4）

其中， | oA f GyA I y ，i, jA 是 A的第i, j 个入口。TV 是惩罚系数。

③ 条件化表情损失函数在减少图像对抗损失的同时，生成网络还必须减少 D 上 AUs 回归产生的误差。

这样，G 不仅学会了渲染真实的样本，还学会了满足fy 编码的目标面部表情。这个损失由两个部分定义:一个是用于优化G 的伪图像的 AUs 回归损失，另一个是用于学习 D 上回归的真图像的 AUs 回归损失。这个损失 y y, , , ,oo fG Dy I y y 如下所示：

22~ 2 ~2|o o o oo of f oD G D y yI y y I y y‖I y y ‖I y （4-5）

④ 一致性损失函数

由上文所述的损失函数，生成网络进行生成逼真的面部转换。但是，如果没有 ground-truth 监督，就无法保证输入和输出图像中的人脸源于同一个人。通过使用循环一致性损失函数[

21]，惩罚原始图像 oy

I 和其重建之间的差异使得生成网络保持

每个个体的一致性。具体公式如下： idt ~1, , , | |o o o ooo f f oG G G yy I y y

I y y I y y I （4-6）

为了生成逼真的图像，对低频信号和高频信号都进行建模。GANimation 的

PatchGan 基于判别网络ID ，通过限制对局部图像块结构的注意力来强化高频信号A, ,ofGy I y

图 4.4 GANimation 模型生成的失败结果的准确性。为了捕获低频信号，使用l1 范数便已足够。在初步实验中，尽管没有性能的提升，还是尝试用更复杂的感知损失函数[

22]来代替l1 范数。

⑤ 全损失函数

为了生成目标图像 gy

I ，通过线性组合上文所述的部分损失函数，来建立全损

失函数 ：

I I y y y, , , , , , ,r rg r g G D G Dy y I y I y y

A A A idt idt, , , , , , ,g r rr g r g G G Gy y y I y I y I y y （4 -7）

其中，A ,y 和 idt

控制每个部分损失函数相对重要性的超参数。最后，定义极大极

小问题，如下所示： argmin maxG

D

G （4-8）

其中，G 从数据分布中抽取样本。另外，将判别网络 D 约束在中， 表示 1-

Lipschitz 函数的集合。

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

15]只能由数据集决定生成离散的表情，其生成

的图像连续自然，较为清晰。相比于其他条件生成模型，GANimation 在合成多类

表情和处理自然图像的能力上均有超越。

GANimation 在某些情况下会出现失败结果，如图 4.4 所示。这可能是因为输入图像仅为一张，训练数据不足引起的。当输入极端表情时，颜色掩膜没有及时调整权重，会导致局部出现透明化。如果输入图像的对象是非人类，模型的效果也会很差。GANimation 生成图像作为人脸表情数据集的扩充，数量方面还远远不够，

尤其不足以满足深度学习表情识别海量训练数据的需要，此方面待以改进。

第五章 SinGANimation 算法仿真实验与分析

5.1 SinGANimation 模型架构本文意图构建基于单幅图像的面部表情生成模型，即输入单幅人脸表情图像，

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

该算法的基本原理为：输入一种表情类别为OC 的单幅图像OCI ，首先通过GANimation，进行单个 AU 变换、多个 AU 连续变换，多个 AU 离散变换等操作，

对图像的表情种类扩充至 N 个，值得注意的是此时每种表情类别还是单幅图像。

然后，将多种表情的单幅图像输入 SinGAN 中，进行再生成操作，对每种表情图像增加至 M 个。因为 SinGAN 再生成的图像与训练图像的差别较小，但又与完全复制不同，所以 SinGAN 再生成只改变每种图像的数量，并不会改变图像种类的多少，即最终结果为

1 2 NC C C

M（I +I + +I ）。这也是本文最大的亮点所在。

图 5.1 SinGANimation 模型架构

5.2 SinGANimation 模型训练

5.2.1 GANimation 模型训练GANimation 生成网络建立在 Johnson 等人[

23]提出的网络变化基础上，其被证

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

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

在参数优化方面，使用 Adam 优化算法[

24]，其参数为学习率 0.0001, 1 0.5, 2

0.999，批量大小为 25。训练 30 个周期，学习率在最后的 10 个周期内线性衰减到 0。每 5 次判别网络的优化，对应执行一次生成网络的单一优化。损失函数的权

重系数设为g p A

10, 0.1, T

V y idt

0.0001, 4000, 10。为了提高稳定性，

尝试在不同的生成网络更新中，使用带有生成图像的缓冲区来更新判别网络，但是没有明显性能的改进。该模型需要在 GTX 1080Ti GPU 上训练两天。

5.2.2 SinGAN 模型训练根据顺序训练多尺度体系结构，从最大的尺度到最小的尺度。一旦每个 GAN

被训练，它就会被固定下来。对第n 个 GAN 的训练损失包括对抗阶段和重建阶段，

即

)(),(maxminn recnnadv

GD

GLDGL nn （5-1）

对抗损失a dv

L 是为nx 的子块距离分布和生成样本 n x~的子块距离分布构造的惩

罚函数。重建损失r ec

L 确保可以产生nx 的特定噪声图谱。

对抗损失每个生成网络nG 都与一个马尔可夫链的判别网络 n

D 相结合，该nD

将其输入的每个重叠的子块分类为真或假。本文使用 WGAN-GP 损失函数来增加训练的稳定性，其中最终的判别得分是图像块判别得分的平均值。相对于纹理的单一图像 GANs，本文定义了整个图像的损失，而不是随机的切割图像。

这允许网络

学习边界条件，这是 SinGAN 设置的一个重要特性。nD 的架构与 n

G 中的n 网络一样,所以它的块大小(网络的感受野)是11 11。

为了确保特定的输入噪声图谱，可以生成原始图像 x 。本文特别选择了 * 1 0

{z ,z ,z } {z ,0, ,0}r ec rec rec

N N ，其中*z 是一些固定的噪音图谱(只绘制一次，在训

练时保持固定)。在使用图像的噪声图谱时，由r ec n x 表示第n 个尺度的生成图像。则对于 n < N，即

第五章 SinGANimation 算法仿真实验与分析 29 2

lrec

))~(,0(nrrec nn

L xx G （5-2）

对于 Nn ， 2 * rec

)(n

N

L xz G 。重建的图像recnx 在训练中还负责确定每个尺度中

噪声nz 的标准差 n

。具体来说，把n 与1(x)rec r n 和 xn 之间的均方误差(RMSE) 成

正比，这提供说明了此尺度内需要添加的细节数量。

5.3 实验数据集

为了验证本文提出算法的性能，两个面部表情数据集被采用进行实验分析：

CelebA 和 RAF-DB 数据集，其中 CelebA 数据集的图像光照均匀，而 RAF-DB 数据集的图像因在一般环境下拍摄，光照分布相对不规则。下面将分别对这两个数据集的细节以及数据处理进行介绍。

5.3.1 CelebA 数据集CelebA 数据集[

25]（CelebFaces Attribute）数据集包含 10177 个名人身份的

202599 张人脸图片，每张图片都有特征标记，包含 40 个二进制属性标注，5 个人

脸特征点坐标等。图 5.2 展示了一些 CelebA 数据集的面部图像。

图 5.2 CelebA 数据集

5.3.2 RAF-DB 数据集

图 5.3 RAF-DB 数据集RAF-DB 数据集[

26]即真实世界的情感面孔数据库，是一个大规模的面部表情

数据库，包括从网上下载的大约 3 万张多样的面部图像。该数据库中的图像在受试者的年龄，性别和种族，头部姿势，光照条件等方面变化很大。图 5.3 展示了一些 RAF-DB 数据集的面部表情图像。

5.3.3 数据预处理在实验中，随机选取数据集的 80%作为训练集，余下的 20%作为测试集。为了加快训练时间，实验中使用 OpenCV，将 CelebA 数据集的图像尺寸从178 218

调整为128 128。另外，由于 RAF-DB 数据集原生图像格式不一，因此对其进行统一裁剪提取图像的人脸区域，本文采用 OpenFace 提取每个图像动作单元，并将每个输出存储在与图像同名的 csv 文件中，以供后续训练模型使用。对 GANimation

模型生成的初步结果进行下采样，便于 SinGAN 模型生成结构更完整、图像更清晰的结果。

5.4 实验结果及其定性分析

5.4.1 单个 AU 变换结果首先，对模型在不同强度激活 AUs 的能力进行评估。该部分实验使用 CelebA

数据集进行测试，在该数据集上进行单个 AU 变换，9 个 AU 子集分别转换为 4 个强度级别(0、0.33、0.66、1)，实验结果如图 5.4 所示。从图 5.4 所示的结果可发现：

当强度为 0 时，不改变相应的 AU；当强度非 0 时，可以观察到每个 AU 是如何逐

图 5.4 单个 AU 变换在 CelebA 数据集生成的结果图 5.5 多个 AU 连续变换在 CelebA 数据集生成的结果步变化。在不同强度下，本文提出的算法模型都能很好地生成与输入图像相对应的结果。为了避免引入不需要的面部运动，恒等变换是至关重要的。

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

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

5.4.2 多个 AU 连续变换结果为了充分展现本文提出算法的性能，该部分实验尝试在 CelebA 数据集上进行多个 AU 连续变换，评估其插入多种表情的能力。实验结果如图 5.5 所示，其中，

第一列为表情ry 的原始图像，最后一列是以gy 为目标表情的综合生成图像，其余几列的结果根据生成网络的条件生成，其线性插值的初始和目标表达式为

(1)g r y y 。由图 5.5 所示的实验结果可知，本文提出的算法其跨帧转换的平滑一致性非常显著。特别地，在该实验中特意选择了具有挑战性的样本，来验证提出算法对光照条件的鲁棒性，甚至是对非现实世界数据分布的鲁棒性，而这些现有的模型中是看不到的。实际上，对于多个 AU 连续变化的实验结果对于之后该模型扩展到视频生成领域具有一定的指导意义。

5.4.3 多个 AU 离散变换结果接下来，本文将 GANimation 与 DIAT[

27]、CycleGAN

[6]、IcGAN

[28]和 StarGAN [15]

模型进行比较。为了公平比较，本文采用了这些模型的结果，即是由 StarGAN 在

RAF-DB 数据集中生成的离散情绪结果(例如，快乐、悲伤和恐惧)。因为 DIAT 和CycleGAN 是非条件生成，所以对于每一对可能的原始和目标情绪，它们被独立来

训练。下面将首先对几种对比算法模型进行简单介绍：

DIAT：给定输入图像 x X 和参考图像 y Y , DIAT 学习 GAN 模型在图像 x 上渲染参考域Y 的属性，同时保持人物的不变性。通过经典的对抗性损失和循环

损失 1( )Y X X Yx G G x 进行训练，以保持人物的一致性。

CycleGAN：如第二章所述，与 DIAT 类似，CycleGAN 也学习两个域之间的映射 X Y 和Y X 。为了训练域之间的变换，使用正则项来表示两个周期的周期一致性损失： 1( )Y X X Yx G G x 和 1

( )X

Y Y X y G G y 。

IcGAN：对于给定的输入图像，IcGAN 使用预训练的编码-解码器将图像编码为潜在表示，并与表情向量 y 连接，然后重构原始图像。在通过解码器之前，

用目标表情替换 y 来修正表情。

图 5.6 多个 AU 离散变换在 RAF-DB 数据集生成的结果

图 5.7 在 RAF-DB 数据集的再生成结果StarGAN：它使用一个掩码向量来忽略未指定的标签，并且只对已知的真值标签进行优化。当同时使用多个数据集进行训练时，它会产生更实际的结果。

相比以上对比算法模型，本文提出的算法与它们主要在两个方面不同。首先，

本文提出的算法不以离散的情感类别为模型设定条件，而是学习了一种在解剖学上可行的变形理论，它允许产生连续的表情。其次，在使用掩膜时，只允许在裁剪后的面部进行变形，并将其放回到原始图像上，而不会产生任何人为修改痕迹。图

5.6 展示了本文算法和 4 种对比算法在多个 AU 离散变换下对 RAF-DB 数据集的面

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

5.4.4 SinGANimation 再生成结果如在第三章的优缺点分析所述，当图像的整体结构非常重要时，SinGAN 可能会产生不真实的结果。为此，本文对 SinGAN 原始模型进行了改进，通过在更小的

范围内启动生成过程来避免人脸图像失真的情况。图 5.7 展示了本文算法在 RAF-

DB 数据集上的生成结果。从图 5.7 所示的结果可知，在原始模型下，以最大尺度

（n N ）开始生成会导致图像严重失真。然而，通过对图像进行下采样，在更小的尺度上输入 N 1层和 N 2 层，可以保持面部的整体结构，同时只改变更微小的细节，如眼睛、鼻子和嘴唇的形状、皮肤和眉毛的纹理等。

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

29]测量来进行定量分析。

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

① 匹配(真伪对比)

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

② 非匹配(不区分真伪)

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

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

通过实验得出两个生成过程的混淆率：从最大尺度 N 1开始(生成具有较大多样性

第五章 SinGANimation 算法仿真实验与分析 35

表 5.1 真伪 AMT 测试结果

输入尺度调查类型混淆率N 1

匹配 21.45% 1.5%

非匹配 42.9% 0.9%

N 2

匹配 30.45% 0.9%

非匹配 47.04% 0.8%

表 5.2 SIFID结果

输入尺度 SIFID 调查类型 SIFID/AMT 混淆率N 1 0.09

匹配 -0.55

未匹配 -0.22

N 2 0.05

匹配 -0.56

未匹配 -0.34

的样本)和从第二个最大尺度 N 2 开始(保存原始图像的全局结构)。在每种情况下，

都进行了匹配和非匹配测试，方差通过 Bootstrap[

30]计算。从获得的实验结果可知，

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

意味着完全混淆真实图像和虚伪图像)。

5.5.2 单幅图像 FID 测量接下来量化了 SinGANimation 对输入图像 x 的内部数据的捕捉程度。GAN 评估的一个常用度量是 FID，它测量生成图像的深度特征分布与真实图像的深度特征分布之间的偏差。然而，在本文实验的设置中，仅仅只有一张真实图像，并且对它的内部图像块数据非常感兴趣。因此，本文提出了单幅图像 FID ，简称 SIFID 测量，它没有使用 Inception 网络[

31]中最后一个池化层之后的激活向量(每个图像一个

向量)，而是使用在第二个池化层之前使用卷积层输出深层特征的内部分布(地图中每个位置一个向量)。本文提出的 SIFID 是真实图像和生成图像中特征统计数据之间的 FID。

实验中将 FID 指标应用于单个图像，并得出完全生成的 50 个图像的平均分。

实验结果如表 5.2 所示。表 5.2 可以看出，AMT 结果的相关性表明，SIFID 与人为

测试结果一致。由 N 2 尺度生成的 SIFID 平均值低于 N 1尺度生成的 SIFID 平

均值，这与用户测试结果保持一致。该部分实验还说明了 SIFID 值与虚伪图像混

淆率之间的相关性，两者之间存在明显的的反相关性，这意味着越小的 SIFID 通常对应较大的混淆率。匹配测试的相关性更强，因为 SIFID 是匹配的度量标准，其

作用于匹配图像nx ， n x 。

5.6 本章小结本文提出的算法 SinGANimation 结合了 GANimation 和 SinGAN，在完全无监督的情况下，对单幅人脸图像进行学习训练，并且生成的表情图像种类多样，规模庞大，质量较高。在 SinGANimation 模型搭建过程中，通过将训练数据下采样输入，解决了 SinGAN 原有模型的人脸生成失真问题，进一步优化 SinGAN 模型，拓

宽了 SinGAN 的适用范围。对于 SinGANimation 模型的性能评估使用了 CelebA 和RAF-DB 两个数据集，前者为特定环境下拍摄的数据集，后者为在自然环境下拍摄的数据集，两者的光照条件，人脸特征等方面差距较大。在两个数据集上生成的图像均达到不错的效果，证明 SinGANimation 可以在不同数据集上应用，具有良好的鲁棒性。

由于现有的人脸表情数据库的图像种类和数量有限，而且例如深度学习等人

脸表情识别需要大规模的训练数据 ， 所以在人脸数据量较少的情况下 ，

SinGANimation 模型可以对人脸表情图像进行种类和数量的扩充，满足实验的需求。

SinGANimation 可生成连续的表情序列，所以在后续的工作中，可以应用于视频序列等工作中。由此可见，SinGANimation 具有非常重要的商用和科研实验价值。另外，SinGANimation 的再生成图像，与对应种类的图像差别较小。这可能是 SinGAN

内部学习训练的原因，在语义多样性方面具有一定的限制。训练 SinGANimation 模型耗时较长，算法时间复杂度较高，还待算法性能进一步优化。

第六章总结与展望

6.1 工作总结人脸表情生成算法与自然图像生成相比，在保证生成图像的画质清晰以外，还需要确保人脸的高度结构性。原有的 SinGAN 对自然图像的生成效果很好，但对人

脸图像的效果差强人意。于是，本文对人脸图像的特性进行研究，提出了结合SinGAN 和 GANimation 两种深度学习网络模型的 SinGANimation 模型，以改善生成效果。本文工作总结如下所示。

首先，创新设计出新型基于单幅图像生成人脸表情算法。本文将图像下采样后再输入，可以保证人脸结构的一致性，解决了 SinGAN 原有模型人脸生成失真的问题。即便这样，人脸表情的数量有所增加，但种类基本保持不变。因此，

本文首次引入了 GANimation 到 SinGAN 模型中，构建出一种新的完全无监督表情生成算法 SinGANimation。在 SinGANimation 模型中，通过 GANimation

进行单个 AU 变换、多个 AU 连续变换、多个 AU 离散变换等操作，对图像的表情种类进行扩充，这样有效地实现了生成表情的数量和种类都有一定程度的增加。

此外，本文对算法进行了定性和定量分析。在实验中，通过与其他经典模型对比，发现本文提出的 SinGANimation 模型可以控制 AU 变换，既可以生成连续自然的表情，还可以生成离散情绪的图像。此外，本文进行了 AMT 真

伪用户测试和单幅图像 FID 测量，得到的生成图像的深度特征分布与真实图像的深度特征分布之间的偏差分别为 0.09 和 0.05，混效率接近 50%，表明本文提出的 SinGANimation 模型生成的图像与真实图像高度相似。

同时，SinGANimation 算法的鲁棒性较强。算法在不同数据集上实验结果均达到不错的效果，由此得以验证 SinGANimation 算法具有良好的鲁棒性。

6.2 工作展望虽然本文提出的人脸表情生成算法与原有算法相比较，在图像画质，表情种类和数量方面取得一定程度的进步，但是由于个人时间和能力有限，算法还存在多处不足，之后的学习研究会进一步改进。

SinGANimation 算法的再生成图像，与对应种类的图像差别较小。这可能是

SinGAN 内部学习训练的原因，在语义多样性方面具有一定的限制。之后对SinGAN 的内部网络结构做进一步的调整，以解决此问题。

训练 SinGANimation 模型耗时较长，算法时间复杂度较高，还待算法性能进一步优化。

致谢

2020 年是特殊的一年，尤其对于我们这些毕业生意义非凡。从最初的恐慌，

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

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

在此，向毛老师致以最诚挚的谢意。

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

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

参考文献

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

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

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

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

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

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

[7] Udwary D W, Zeigler L, Asolkar R N, et al. Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica[J]. Proceedings of the National

Academy of Sciences, 2007, 104(25): 10376-10381.

[8] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani.Ingan: Capturing and remapping the

“DNA” of a natural image. arXiv preprint arXiv: arXiv:1812.00231, 2018.

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

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

[11] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro

Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo realistic single image super-resolution using a generative ad versarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681 - 4690,2017.

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

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

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

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

[16] Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE

Transactions on Computers 22(1), 67 - 92,1973.

[17] Yu, H., Garrod, O.G., Schyns, P.G.: Perception-driven facial expression synthesis.Computers &

Graphics 36(3),2012.

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

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

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

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

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

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

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

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

[26] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.:

Presentation and validation of the radboud faces database. Cognition and emotion 24(8), 1377- 1388,2010.

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

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

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner,Bernhard Nessler, and Sepp Hochreiter.

GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, pages 6626－6637, 2017.

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

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet,Scott Reed, Dragomir Anguelov,

Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions.

In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1－9, 2015

班级 1602051

学号 16020510038

本科毕业设计（论文）

外文资料翻译

毕业设计题目	基于单幅图像的面部表情生成算法研究
外文资料题目	SinGAN: Learning a Generative Model from a Single Natural Image
学院	人工智能学院
专业	智能科学与技术
学生姓名	邢博伟
指导教师姓名	毛莎莎

从单一的自然图像中学习生成模型

单张训练图像　　单张图像的随机采样

图 1:**从单个训练图像中学习的图像生成**。我们提出了一种基于单一自然图像的无条件生成模型。我们的模型通过使用一种专门的多尺度对抗训练方案，在多个尺度上学习子图像块数据;然后，它可以用来生成新的逼真的图像样本，在创建新的对象配置和结构时，保持原始的子图像块的分布。

摘要

我们介绍了 SinGAN，一个可以从单一自然图像中学习的无条件生成模型。我们的模型经过训练，能够捕捉图像内部子块的分布，然后能够生成与图像具有相同视觉内容的高质量、多样化的样本。SinGAN 包含一个完全卷积的 GANs 金字塔，每个 GANs 负责学习图像中不同尺度的子图像块分布。这允许生成具有显著可变性的任意大小和高宽比的新样本，同时保持训练图像的全局结构和精细纹理。与之前的单图像 GAN 方案相比，我们的方法不仅限于纹理图像，而且没有条件(即从噪声中生成样本)。用户研究证实，生成的样本通常被混淆为真实图像。我们说明了 SinGAN 在图像处理任务中的广泛应用。

一、简介

生成式对抗网(GANs)[19]已经做了一个对高维分布建模的巨大飞跃视觉数据。特别是无条件的 GANs 在特定类数据集(例如，脸[33],卧室[47])。然而，捕捉分布高度多样化的数据集的多个对象类(例如，ImageNet[12])，仍然被认为是一个主要的挑战，通常需要调节生成另一个输入信号[6]或为特定的任务训练模型(例如，超分辨率[30]，图像修复[41]，重定向[45])。

在这里，我们将 GANs 的使用带入一个新的领域，从一个单一的自然图像中无条件地生成学习。具体来说，我们证明了单个自然图像中子块的内部数据信息通常包含了足够的信息，可以用来学习一个强大的生成模型。我们新的单一图像生成模型 SinGAN 允许我们处理包含复杂结构和纹理的一般自然图像，而不需要依赖于来自同一类图像的数据库。这是通过一个完全卷积的轻量级 GANs 金字塔来实现的，每个 GANs 负责捕获不同规模的子块分布。一旦经过训练，

SinGAN 可以生成各种高质量的图像样本(任意维度)，这些样本在语义上与训练图像相似，但包含新的对象配置和结构(图 1)。

对单个自然图像中子块的内部分布进行建模，长期以来一直被认为是许多计算机视觉任务的重要前提[64]。经典的例子包括去噪[65]，去模糊[39]，超分辨率[18]，去雾[2,15]，图像编辑[37,21,9,11,50]。在这方面最相关的工作是[48]，其中定义了一个双向的子块相似性度量，并对其进行了优化，以保证处理后的图像子块与原始图像的子块是相同的。在这些工作的启发下，我们在这里学习如何在简单的统一学习中使用 SinGAN 解决多种图像处理任务的框架，包括从单个图像到绘图、编辑、协调、超分辨率和动画在所有这些情况下，我们的模型产生了高质量的结果，保持了训练图像的内部子块统计信息(见图 2)和我们的项目网页)。所有的任务是基于同一个生成式网络，没有任何额外的信息或进一步超出最初训练的图像。

图 2:**图像处理**。SinGAN 可用于各种图像处理任务，包括:转换一个油画(剪贴画)成一个现实的照片，重新安排和编辑图像中的对象，协调一个新对象成一个图像，图像的超分辨率和创建一个动画从一个单一的输入。在所有这些情况下，我们的模型只观察训练图像(第一行)，并以相同的方式对所有应用程序进行训练，不进行架构更改或进一步调优(参见第 4 节)。

1、相关工作

（1）单像深度模型

最近有几项研究提出将深度模型过度拟合到单个训练实例中[51,60,46,7,1]。然而，这些

方法是为特定的任务而设计的(例如，超分辨率[46]，纹理扩展[60])。Shocher 等人[44,45]首先为单个自然图像引入了基于内部 GAN 的模型，并在重定向的背景下进行了说明。然而，它们的生成取决于输入图像(即，将图像映射到图像)，而不是用来绘制随机样本。相比之下，我们的框架是纯生成的(即将噪声映射到图像样本)，因此适合许多不同的图像处理任务。无条件的单图像 GANs 仅在纹理生成的环境中被探索过[3,27,31]。这些模型在对非纹理图像进行训练时，并不能生成有意义的样本(图 3)。而我们的方法并不局限于纹理，可以处理一般的自然图像(如图 1)。

图 3:**SinGAN 与单个图像纹理生成**。用于纹理生成的单一图像模型[3,16]并不用于处理自然图像。我们的模型可以生成包含复杂纹理和非重复全局结构的真实图像样本。

生成模型对图像的操纵能力最近的研究证明了对抗性学习的重要性基于 GAN 的方法，在许多不同的图像处理任务[61,10,62,8,53,56,42,53]。案例包括交互式的图像编辑 [61,10]，sketch2image [8,43]，和其他图像到图像的翻译任务[62,52,54]。然而，所有这些方法在具体的数据集上训练，这里也一样，通常将生成条件设置为另一个输入信号。我们对捕获普通不感兴趣同类图像的特征，而是考虑不同来源的训练数据——全部重叠单个自然图像的多个尺度上的图像子块。我们展示了一个强大的生成模型是可以学习的此数据，可用于多种图像处理任务。

二、方法

我们的目标是学习一个无条件生成模型，它可以捕获单个训练图像 x 的内部统计信息。这

个任务在概念上与传统的 GAN 设置类似，只是这里的训练样本是单个图像的子块，而不是来自数据库的整个图像样本。

我们选择超越纹理生成，并处理与更一般的自然图像。这需要捕捉复杂图像结构的统计在很多不同的尺度。例如，我们希望捕获全局属性。例如大型物体的排列和形状图像(例如，顶部的天空，底部的地面)，以及作为精细的细节和纹理信息。为了实现这一目标，我们生成 式结构，如图 4 所示，包含一个多层次的子块-GANs（马尔可夫链的鉴别器）[31, 26],其中每个负责捕捉子块分布在不同规模的 x。The GAN 有小的接受域和有限容量，阻止它们记忆单一图像. 同时，相似的多尺度的架构一直在探索传统 GAN 设置(例如[28,52,29,52,13,24])，我们是第一个从单一图像内部学习探索它的人。

1、多尺度结构

我们的模型由一个生成器金字塔组成，$\{G_0,\cdots,G_N\}$，针对 x 的图像金字塔进行训练: $\{x_0,\cdots,x_N\}$，其中 x_n 是一个因子 r_n 的 x 的下采样版本，对于某个 r > 1。每个生成器 G_n 负责生成真实的图像样本 w.r.t，即对应图像 x_n 中的子块分布。这是通过对抗训练实现的，G_n 学习欺骗一个相关的鉴别器 D_n， D_n 试图将生成的样本中的子块与 x_n 中的子块区分开来。

图像样本的生成从最粗的尺度开始，依次通过所有生成器，直到最细的尺度，在每个尺度注入噪声。所有的发生器和鉴频器都有相同的接收域，因此在生成过程中捕获的结构尺寸都在减小。在最大尺度上，生成是纯生成的，即 G_N 将空间高斯白噪声 z_n 映射到图像样本 $\tilde{x}_N$，

$$\tilde{x}_N = G_N(z_N). \qquad (1)$$

这一层的有效接受域通常是图像高度的 1/2，因此 G_N 生成图像的总体布局和对象的全局结构。每个生成器在更细的尺度上(n < n)添加以前的尺度没有生成的细节。因此，除了空间噪声 z_n 外，每个生成器 G_n 还接受较粗尺度图像的上采样版本，即

$$\tilde{x}_n = G_n\left(z_n, (\tilde{x}_{n+1})\uparrow^r\right), \qquad n < N. \qquad (2)$$

所有的生成器都具有类似的架构，如图 5 所示。特别地，在被送入一系列卷积层之前，噪声 z_n 被加在图像 $\left(\tilde{x}_{n+1}\right)\uparrow^r$。这确保了 GAN 不会忽略噪声，就像随机条件方案中经常发生的那样[62,36,63]。对流层的作用是生成缺失的细节 $\left(\tilde{x}_{n+1}\right)\uparrow^r$（残余学习）[22,57]。

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

即 G_n 执行操作 ψn 在一个有五个 Conv(3 × 3)-BatchNorm-LeakyReLU 形式的卷积块的完全卷积形式。我们在最大的尺度上从每个块 32 个内核开始，然后每 4 个尺度增加 2 倍。因为生成器是全卷积的，所以我们可以在测试时生成任意大小和宽高比的图像(通过改变噪声图的尺寸)

2、训练

我们按顺序训练我们的多尺度体系结构，从最大的尺度到最小的尺度。一旦每个 GAN 被训练，它就会被固定下来。我们对 nth GAN 的训练损失包括一个对抗性术语和一个重建性术语，

$$\min_{G_n} \max_{D_n} \mathcal{L}_{\text{adv}}(G_n, D_n) + \alpha \mathcal{L}_{\text{rec}}(G_n). \qquad (4)$$

对抗性的的损失 L_{adv} 惩罚子块之间的距离分布的 x_n 和子块的分布生成样本 $\tilde{x}_n$。重建损失 L_{rec} 确保存在一组特定的噪声地图可以产生 x_n，图像处理的一个重要特性(章节 4)。我们接下来描述 L_{adv}， L_{rec} 细节。详见补充材料(SM)。

对抗性的损失

对抗损失每个生成器 G_n 都与一个马尔可夫过程的鉴别器 D_n 相结合，该 D_n 将其输入的每个重叠的子块分类为真或假[31,26]。我们使用 WGAN-GP 损失[20]来增加训练的稳定性，其中最终的识别分数是图像块鉴别匹配的平均值。相对于纹理的单图像 GANs(例如，[31,27,3])，这里我们定义了整个图像的损失，而不是随机的作物(批量大小为 1)。这允许网络学习边界条件(参见 SM)，这是我们设置的一个重要特性。 D_n 的架构与 G_n 中的 ψ_n 网络一样，所以它的块大小(网络接受域)是 11×11。

重建的损失

我们要确保存在一组特定的输入噪声映射，它生成原始图像 x。我们特别选择了

$\{z_N^{rec}, z_{N-1}^{rec} \cdots, z_0^{rec}\} = \{z^*, 0, \cdots, 0\}$，其中 z^* 是一些固定的噪音图(只绘制一次，在训练时保持固定)。表示由 $\tilde{x}_n^{rec}$ 量表时使用这些生成的图像噪声地图。对于 n < N，

$$\mathcal{L}_{\text{rec}} = \|G_n(0, (\tilde{x}_{n+1}^{\text{rec}}) \uparrow^r) - x_n\|^2, \qquad (5)$$

而且对于 n=N,我们使用 $L_{rec} = \left\|G_N(z^*) - x_N\right\|^2$。

重建的图像 $\tilde{x}_n^{rec}$ 在训练中有另一个角色，在每个规模中这是确定噪声 z_n 的标准差 σ_n。具体来说,我们把 σ_n 作为比例的均方误差(RMSE) 在 $(\tilde{x}_{n+1}^{rec}) \uparrow^r$ 和 xn 之间，这说明了在这个范围内需要添加的细节的数量。

三、结果

我们对我们的方法进行了定性和定量的测试,包括城市和自然风景,以及艺术和纹理图像。我们使用的图像取自伯克利分割数据库(BSD)[35]、地点[59]和网站。我们总是在最大的刻度上设置最小的尺寸为 25px，并选择 N 刻度数，因此比例因子 r 尽可能接近 4/3。对于所有的结果，(除非另有说明)，我们将训练图像调整为最大尺寸 250px。

定性的例子生成的随机图像样本 如图1,6所示 ， 还有更多的例子在SM中. 对于每个 例子，我们展示大量随机抽样与原有的长宽的图像，还有减少和扩大每个轴尺寸的图像。由此可见，在所有的例子中,那些生成样本描述新的物体的现实结构和配置，同时保留训练图像的视觉内容。我们的模型成功地保存了对象的全局结构，例如山脉(图 1)、气球或金字塔(图 6)，以及精细的纹理信息。因为网络有一个有限的接受域(小于整个图像),它可以生成新的子块中不存在的组合训练图像此外,我们注意到,在许多情况下反射和阴影是实际合成,可以看到在图 6 和图 1(图 8)的第一个例子。请注意，SinGANs 架构是与分辨率无关的，因此可以用于高分辨率图像，如图 7 所示(见 SM 中的 4M pix 结果)。在这里，所有尺度的结构都很好地生成了，从天空、云和山脉的整体布局，到雪的精细纹理。

在测试时间尺度的影响

我们的多尺度体系结构允许通过选择在测试时开始生成的尺度来控制样本之间的变化量。从第 n 个刻度开始，我们把噪波映射调整到这个刻度 $\{z_N^{rec}, \cdots, z_{n+1}^{rec}\}$，并且仅对 $\{z_n, \cdots, z_0\}$ 使用随机抽签。其效果如图 8 所示。可以看出，在最大的尺度上开始生成(n = N)，会导致全球结构的大变异性。在某些情况下，一个大的突出的物体，如斑马图像，这可能导致不切实际的样本。

然而，从更小的尺度开始生成，可以保持全局结构不变，而只改变更细的图像特征(例如斑马的 s 条纹)。参见 SM 获得更多的例子。

训练图像 **从单个图像中的随机采样**

图 6:**随机图像样本**。在单一图像上训练后，我们的模型可以生成真实的随机图像样本，描述新的结构和对象配置，同时保持训练图像的子块分布。由于我们的模型是完全卷积的，因此生成的图像可能具有任意大小和纵横比。请注意，我们的目标不是图像重定向，我们的图像样本是随机的和优化的，以维护子块统计，而不是保留突出的对象。更多结果和图像重定向方法的定性比较请参见 SM。

图 7:**高分辨率图像生成**。我们的模型生成的随机样本，训练在 243×1024 的图像上(右上角);新的全局结构以及精细的细节被真实地生成。参见 SM 中的 4Mpix 示例。

训练中量表的作用

图 9 显示了用更少尺度的作用，在最粗糙的水平上有效的接受域更小，只允许捕获精细的纹理。随着尺度数量的增加，更大支撑的结构出现，而且全局对象的安排保存得更好。

1、 定量评价

为了量化生成图像的真实性以及它们如何捕获训练图像的内部统计数据，我们使用了两个指标:(i) Amazon Mechanical Turk (AMT)“真实/虚假”用户研究，(ii) Frechet Inception 距离[23]的新单图像版本。

AMT 感知研究

我们遵循了[26,58]的协议，并在两种情况下进行感知实验。(i)配对(真与假):研究人员向参与者展示了 50 个实验序列，每个实验中，一张假图像(由 SinGAN 生成)与它的真实训练图像进行 1 秒钟的对比。工作人员被要求挑选出假照片。(ii)非配对(真或假):工作人员看到一张图片 1 秒钟，然后被问及这是否是假的。总共有 50 张真实的图像和 50 张不相关的假图像被随机分配给每个工作单位。

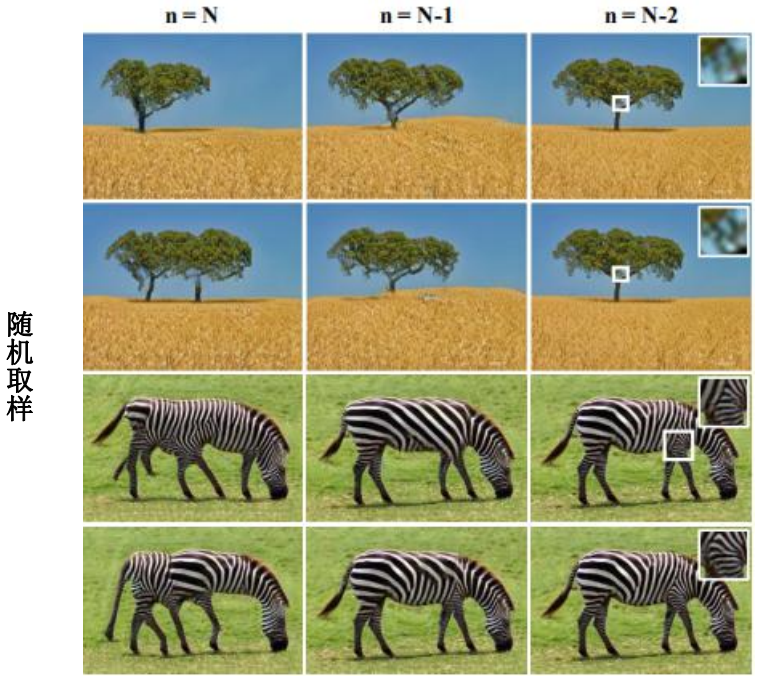


图 8：**来自不同尺度的生成(在推断时)**. 我们展示了从给定级别 n 开始分层生成的效果。对于我们的全生成方案(n = n)，最粗级的输入是随机噪声。为了从更小的比例 n 生成，我们插入向下采样的原始图像 xn 作为该比例的输入。这使得我们可以控制生成结构的规模，例如，我们可以保持斑马的形状和姿势，只改变它的条纹纹理，从 $n = N - 1$ 开始生成。

我们对两种类型的生成过程重复了这两个协议:从最大的(Nth)尺度开始生成，从 $N-1$ 尺度开始生成(如图 8 所示)。为了量化生成图像的多样性，对于每个训练示例，我们计算每个像素超过 100 个生成图像的强度值的标准差(std)，在所有像素上取平均值，然后根据训练图像的强度值的标准差进行标准化。

真实的图片是从数据库[59]中随机选取的，来自山脉、丘陵、沙漠和天空的子类别。在这

四个测试中，我们有 50 个不同的参与者。在所有测试中，前 10 个测试都是包含反馈的教程。结果见表 1。

正如预期的那样，在不配对的情况下，混淆率始终较大，在这种情况下没有可供比较的参考。此外，很明显，混淆率随着生成图像的多样性而降低。然而，即使改变了大型结构，我们生成的图像也很难与真实图像区分开来(50%的分数意味着完全混淆了真实图像和虚假图像)。完整的测试图像包含在 SM 中。

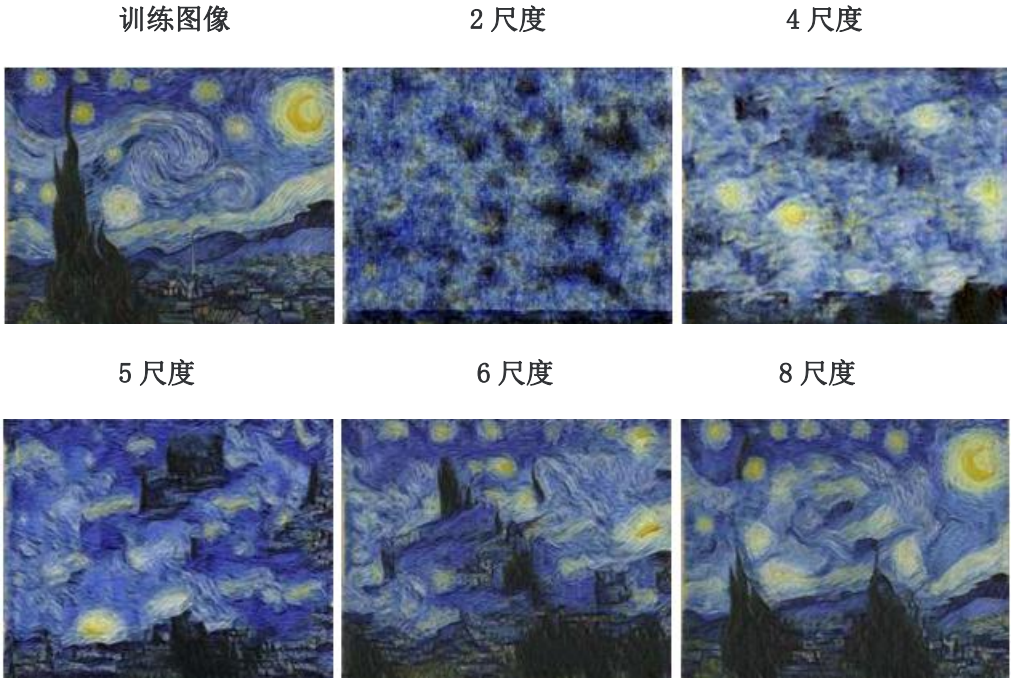


图 9:**使用不同数量的量表进行训练的效果**。SinGAN 架构中的尺度数量对结果有很大的影响。只有少量比例的模型只能捕获纹理。随着尺度数量的增加，SinGAN 成功地捕捉到了更大的结构以及场景中物体的整体布局。

1st Scale	Diversity	Survey	Confusion
N	0.5	paired	$21.45\% \pm 1.5\%$
		unpaired	$42.9\% \pm 0.9\%$
$N-1$	0.35	paired	$30.45\% \pm 1.5\%$
		unpaired	$47.04\% \pm 0.8\%$

表 1:**“真/假”AMT 检验**。我们报告了两个生成过程的混淆率:从最大尺度 N 开始(生成具有较大多样性的样本)，从第二个最大尺度 N-1 开始(保留原始图像的全局结构)。在每种情况下，我们都进行了配对研究(显示真-vs-假图像对)，和一个未配对的(无论是假或真图像显示)。方差由辅助程序[14]估计。

单幅图像 Frechet 先启距离

接下来，我们量化了 SinGAN 对 x 的内部统计数据的捕获程度。GAN 评估的一个常用度量是 Frechet 先启距离(FID)[23]，它度量生成图像的深度特征分布与真实图像的分布之间的偏

差。然而，在我们的设置中，我们只有一个真实的图像，并且对它的内部补丁统计非常感兴趣。因此，我们提出了单图像 FID (SIFID)度量。在初始化网络[49]中，我们不是使用最后一个池化层之后的激活向量(每个图像一个向量)，而是在第二个池化层之前使用卷积层输出的深层特征的内部分布(图中每个位置一个向量)。我们的 SIFID 是真实图像和生成的样本中这些特征的统计数据之间的 FID。

图 10:**超分辨率**。当 SinGAN 被训练在一个低分辨率的图像上时，我们能够超级分辨。这是通过迭代地对图像进行采样并将其输入到 SinGAN 的最精细的比例生成器来实现的。可见，SinGAN 的视觉质量优于 SOTA 内部方法 ZSSR[46]和 DIP[51]。它也比 EDSR[32]好，可以与 SRGAN[30]相比，后者是在大型集合上训练的外部方法。括号中显示了相应的 PSNR 和 NIQE[40]。

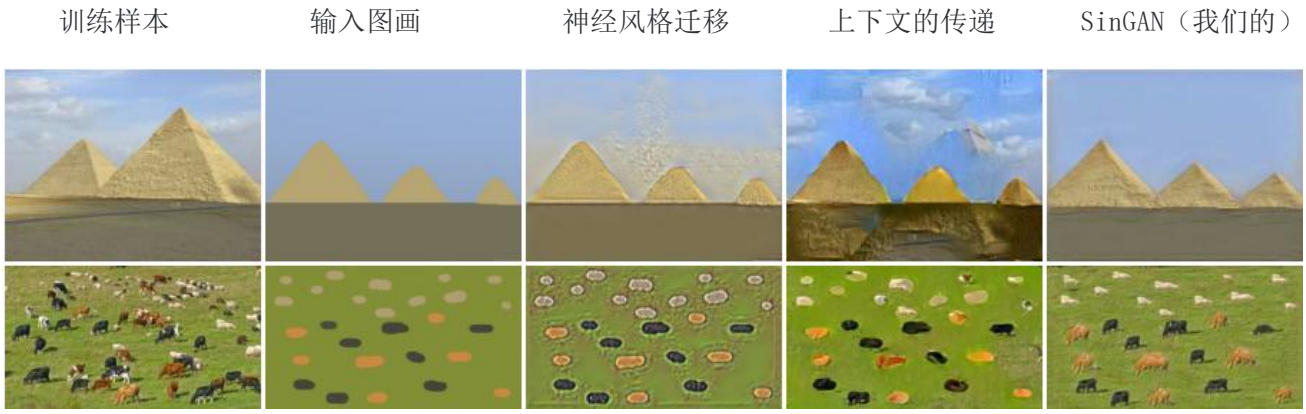


图 11:**图画到图像**。我们在目标图像上训练 SinGAN，并在测试时将一个向下采样的图画注入到一个粗糙的水平。我们生成的图像保留了剪贴画的布局和一般结构，同时生成与训练图像匹配的真实纹理和精细细节。著名的样式转移方法[17,38]在此任务中失败。

1st Scale	SIFID	Survey	SIFID/AMT Correlation
N	0.09	paired	−0.55
		unpaired	−0.22
$N-1$	0.05	paired	−0.56
		unpaired	−0.34

表 2:**单个图像 FID (SIFID)**。我们将 FID 指标应用于单个图像，并报告 50 幅图像的平均分，对于完整的生成(第一行)，以及从第二个最大尺度(第二行)开始。与 AMT 结果的相关性表明，SIFID 与人类的排名高度一

致。

从表 2 可以看出,从规模 N-1 中生成的 SIFID 平均值低于从规模 N 中生成的 SIFID 平均值,这与用户研究结果一致。我们还报告了 SIFID 分数和假图像的混淆率之间的相关性。请注意,这两者之间存在显著的(反)相关性,这意味着一个小的 SIFID 通常可以很好地指示出较大的混淆率。成对测试的相关性更强,因为 SIFID 是成对的措施(设有一对 $x_n, \tilde{x}_n$)。

四、应用

我们将探讨 SinGAN 在许多图像处理任务中的应用。为此，我们在培训后使用我们的模型，不进行架构更改或进一步调优，并对所有应用程序采用相同的方法。该思想是利用这样一个事实，即在推理时，SinGAN 只能生成与训练图像具有相同子块分布的图像。因此，可以通过在 $n < N$ 的某个尺度将图像(可能是向下采样的版本)注入到生成金字塔中，并通过生成器将其前馈，从而使其子块分布与训练图像的子块分布匹配，从而进行操作。不同的注射规模导致不同的效果。我们会考虑之后的应用。(更多结果和注射规模效应见 SM)。

超分辨率

将输入图像的分辨率提高一个因子 s。我们在低分辨率上训练我们的模型(LR)图像,重建减肥 $\alpha = 100$ 和金字塔的比例因子 $r = \sqrt[k]{s}$ 对于 $k \in N$。由于小型结构往往在自然场景[18]的尺度上反复出现，因此在测试时，我们通过一个 r 因子对 LR 图像进行上采样，并将其(连同噪声)注入最后一个生成器 G_0。我们重复 k 次以获得最终的高分辨率输出。示例结果如图 10 所示。可以看出，我们重建的视觉质量超过了目前最先进的内部方法[51,46]，也超过了以最大信噪比[32]为目标的外部方法。有趣的是，它可以与外部训练的 SRGAN 方法[30]相媲美，尽管它只暴露在一张图像中。在[4]之后，我们在 BSD100 数据集[35]上比较表 3 中 5 种方法的失真程度(RMSE)和感知质量(NIQE[40])是两个根本冲突的需求[5]。可以看出,SinGAN 擅长感知;其 NIQE 分数仅略低于 SRGAN，其 RMSE 稍好一些。

	External methods		Internal methods		
	SRGAN	EDSR	DIP	ZSSR	SinGAN
RMSE	16.34	12.29	13.82	13.08	16.22
NIQE	3.41	6.50	6.35	7.13	3.71

表 3:**超分辨率评价**。在[5]之后，我们报告了失真(RMSE)和感知质量(NIQE)[40]，越低越好)在 BSD100[35]上。由此可以看出，SinGAN 的性能与 SRGAN[30]类似。

图画到图像

将剪贴画转换成逼真的图像。这是通过向下采样的剪贴画图像并将其输入一个粗尺度(例如N-1 或 N-2)来实现的。从图 2 和图 11 可以看出，我们保留了画面的整体结构，真实地生成了与原图匹配的纹理和高频信息。我们的方法在视觉质量上优于风格迁移方法[38,17](图 11)。

调和

现实地混合粘贴对象与背景图像。我们在背景图像上训练 SinGAN，并在测试时注入原始粘贴的复合材料的下采样版本。在这里，我们将生成的图像与原始背景相结合。从图 2 和图 13 可以看出，我们的模型对粘贴对象的纹理进行了裁剪以匹配背景，并且经常比[34]更好地保留了对象的结构。缩放 2、3、4 通常会在保持对象的结构和转移背景纹理之间取得良好的平衡。

编辑

生成一个无缝的合成，其中图像区域已复制和粘贴到其他位置。这里，我们再次将复合材料的下采样版本注入到粗尺度之一。然后，我们将编辑区域的 SinGANs 输出与原始图像结合起来。如图 2 和图 12 所示，SinGAN 重新生成了精细的纹理，并无缝地缝合了粘贴的部分，产生了比 Photoshop 的内容感知效果更好的效果。

单一图像的动画

创建一个简短的视频剪辑与现实的物体运动，从一个单一的输入图像。自然图像往往包含重复，这显示不同的同一动态对象[55]的“快照”(例如，一群鸟的图像显示了一只鸟的所有翅膀姿势)。使用 SinGAN，我们可以沿着图像中物体的所有表象的流形前进，从而从一个单一的图像合成运动。我们发现，对于许多类型的图像，一个现实的效果是通过 z 空间中的随机漫步实现的，从 z^{rec} 开始的第一帧在所有的生成尺度(见 SM 视频)。

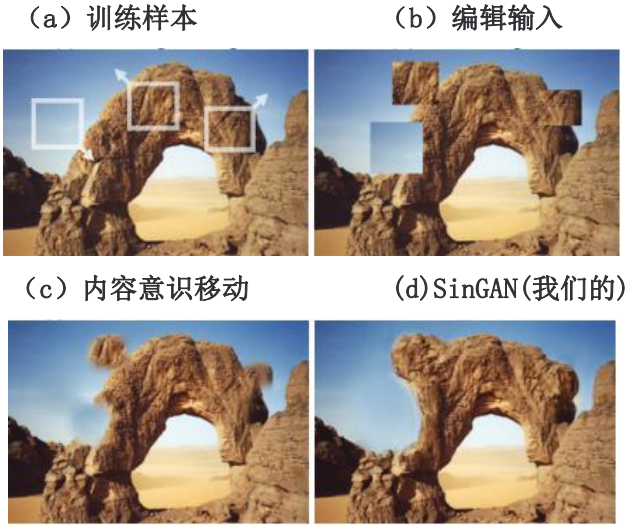


图 12:**编辑**。我们从原始图像(a)中复制并粘贴一些补丁，然后将编辑后的图像(b)的下采样版本输入到我们

的模型的中级水平(在(a)上进行预训练)。在生成的图像(d)中，这些局部编辑被转换成连贯的、逼真的结构。(c) 比较 Photoshop 的内容意识移动。

图 13:**协调**。我们的模型能够保持粘贴对象的结构，同时调整其外观和纹理。专用的协调方法[34]过度混合的对象与背景。

五、结论

我们介绍了一个新的无条件生成方案 SinGAN，它是从一个单一的自然图像中学习来的。我们展示了它超越纹理的能力，并为自然复杂的图像生成多样的真实样本。与外部训练的生成方法相比，内部学习在语义多样性方面存在固有的局限性。例如，如果训练图像包含一条狗，我们的模型将不会生成不同犬种的样本。然而，我们的实验证明，SinGAN 可以提供一个非常强大的工具，为广泛的图像处理任务。

致谢 Idan Kligvasser 的宝贵见解。这项研究得到了以色列科学基金会(grant 852/17)和奥伦多夫基金会的支持。

A Fine-Grained Facial Expression Database for End-to-End Multi-Pose Facial Expression Recognition

Wenxuan Wang
Fudan University
wxwang17@fudan.edu.cn

Qiang Sun
Fudan University
18110860051@fudan.edu.cn

Tao Chen
Fudan University
eetchen@fudan.edu.cn

Chenjie Cao
Ping An OneConnect
caochenjie948@pingan.com

Ziqi Zheng
Ping An OneConnect
zhengziqi356@pingan.com.cn

Guoqiang Xu
Ping An OneConnect
xuguoqiang371@pingan.com.cn

Han Qiu
Ping An OneConnect
hannaqiu@pingan.com.cn

Yanwei Fu *
Fudan University
Corresponding Author
yanweifu@fudan.edu.cn *

Abstract

The recent research of facial expression recognition has made a lot of progress due to the development of deep learning technologies, but some typical challenging problems such as the variety of rich facial expressions and poses are still not resolved. To solve these problems, we develop a new Facial Expression Recognition (FER) framework by involving the facial poses into our image synthesizing and classification process. There are two major novelties in this work. First, we create a new facial expression dataset of more than 200k images with 119 persons, 4 poses and 54 expressions. To our knowledge this is the first dataset to label faces with subtle emotion changes for expression recognition purpose. It is also the first dataset that is large enough to validate the FER task on unbalanced poses, expressions, and zero-shot subject IDs. Second, we propose a facial pose generative adversarial network (FaPE-GAN) to synthesize new facial expression images to augment the data set for training purpose, and then learn a LightCNN based Fa-Net model for expression classification. Finally, we advocate four novel learning tasks on this dataset. The experimental results well validate the effectiveness of the proposed approach.

1. Introduction

Facial expression [5], as the most important facial attribute, reflects the emotion status of a person, and contains meaningful communication information. Facial expression recognition (FER) is widely used in multiple applications such as psychology, medicine, security and education [5]. In psychology, it can be used for depression recognition for analyzing psychological distress. On the other hand, detecting a student's concentration or frustration is also helpful in improving the educational approach.

Facial expression recognition mainly contains four steps: face detection, face alignment, feature extraction and facial expression classification. (1) In the first step, the face is detected from the image with each labelled by a bounding box. (2) In the second step, the face landmarks are generated to align the face. (3) In the third step, the features that contain facial related information are extracted in either handcrafted way, *e.g.*, SIFT, [4] Gabor wavelets [3, 22] and LBP [29] or learned way by a neural network. (4) In the fourth step, various classifiers such as SVM, KNN and MLP can be adopted for facial expression classification.

The recent renaissance of deep neural networks delivers the human level performance towards several vision tasks, such as object classification, detection and segmentation [19, 18, 27]. Inspired by this, some deep network methods [15, 23, 35] have been proposed to address the facial expression recognition. In FER task, facial expression is usually assumed to contain six discrete primary emotions: anger, disgust, fear, happy, sad and surprise according to Ekman's theory. With an additional neutral emotion, the seven emotions compose the main part of most common emotion datasets, including CK+ [20, 14], JAFEE [22], FER2013 [26] and FERG [2].

However, one most challenging problem of FER in fact is lacking of a large-scale dataset of high quality images, that can be employed to train the deep networks and in-

vestigate the impacting factors for the FER task. Another disadvantage of these datasets, *e.g.*, JAFFE and FER2013 dataset, is the little diversity of expression emotions, which cannot really express the versatile facial expression emotions in the real world life.

To this end, we create a new dataset F^2ED (Fine-grained Facial Expression Database) with 54 emotion types, which include larger number of emotions with subtle changes, such as calm, embarrassed, pride, tension and so on. Further, we also consider the influence of face pose changes on the expression recognition, and introduce the pose as another attribute for each expression. Four orientations (poses) including front, half left, half right and bird view are labelled, and each has a balanced number of examples to avoid training bias.

On this dataset, we can further investigate how the poses, expressions, and subject IDs affect the FER performance. Critically, we propose four novel learning tasks over this dataset as shown in Fig. 1(c). They are expression recognition with the standard balanced setting (ER-SS), unbalanced expression (ER-UE), unbalanced poses (ER-UP), and zero-shot ID (ER-ZID). Similar to the typical zero-shot learning setting [16], the zero-shot ID setting means that the testing faces of persons have not appeared in the training set. To tackle these four learning tasks, we further design a novel framework that can augment training data, and then train the classification network. Extensive experiments on our dataset, as well as JAFEE [22], FER2013 [26] show that (1) our dataset is large enough to be used to pre-train a deep network as the backbone network; (2) the unbalanced poses, expressions and zero-shot IDs indeed negatively affect the FER task; (3) the data augmentation strategy is helpful to learn a more powerful model yielding better performance. These three points are also the main contributions of this paper.

2. Related Work

2.1. Facial expression recognition

Extensive FER works based on neural networks have been proposed [15, 31, 36]. Khorrami *et al.* [15] trains a CNN for FER task, visualizes the learned features and finds that these features strongly correspond to the FAUs proposed in [6]. Attentional CNN [23] on FER is proposed to focus on the most salient parts of faces by adding a spatial transformer.

Generative Adversarial Net (GAN) [9] based models have also been investigated in solving the FER task. Particularly, GAN is usually composed of a generator and a discriminator. In order to weaken the influence of pose and occlusion, the pose-invariant model [35] is proposed by generating different pose and expression faces based on GAN. Qian *et al.* [28] propose a generative adversarial network (GAN) designed specifically for pose normalization in re-id. Yan *et al.* [34] propose a de-expression model to generate neutral expression images from source images by Conditional cGAN [24], and use the residual information in the intermediate layer in GAN to classify the expression.

2.2. Previous Datasets

CK+. The extended Cohn-Kanade (CK+) database [20]is an updated version of CK database [14]. In CK+ database, there are 593 video sequences from 123 subjects. Of the 593 video sequences, 327 are selected according to the FACS coded emotion labels. The last frame of the selected video is labeled as one of the eight emotions: angry, contempt, disgust, fear, happy, sad, surprise and neutral.

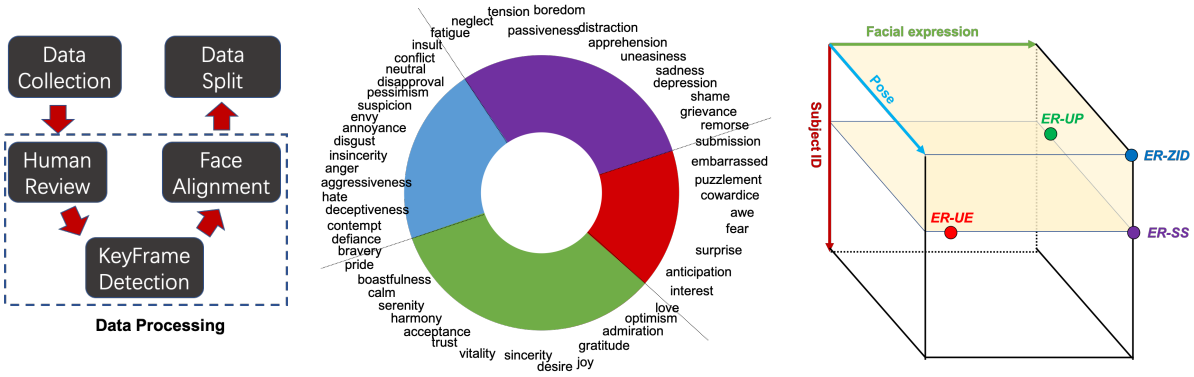
JAFFE. The Japanese Female Facial Expression (JAFFE) database [22] contains 213 images of 256×256 pixels resolution. The images are taken from 10 Japanese female models in a controlled environment. Each image is rated with one of the following 6 emotion adjectives: angry, disgust, fear, happy, sad and surprise.

FER2013. The Facial Expression Recognition 2013 database [26] contains 35887 images of 48×48 resolution. These images are taken in the wild setting which means more challenging conditions such as occlusion and pose variations are included. They are labelled as one of the seven emotions as described above. The dataset is split into 28709 training images, 3589 validation images and 3589 test images.

KDEF. The dataset of Karolinska Directed Emotional Faces [21] contains 4900 images of 562×762 pixels resolution. The images are taken from 140 persons (70 male, 70 female) from 5 angles with 7 emotions. The angles contain full left profile, half left profile, front, full right profile and half right profile. The emotion set contains 7 expressions: afraid, angry, disgusted, happy, sad, surprised and neutral.

2.3. Learning paradigms

Zero-shot learning recognize the new visual categories that have not been seen in the labelled training examples [16]. The problem is usually solved by transferring learning from source domain to the target domain. Semantic attributes that describe a new object can be utilized in zero-shot learning. Xu *et al.*[33] propose a zero-shot video emotion recognition. In this paper, we propose a novel FER task on the persons that are not in the training set. On the other hand, class imbalance is a common problem, especially in deep learning [12, 8]. For the first time, we propose a dataset that is large enough to help to evaluate the influence of unbalanced poses, expressions, and person IDs over the FER task. To alleviate this issue, we investigate synthesizing more data by GAN-based data augmentation inspired by recent works on Person Re-ID[28] and Facial expression recognition [35].



(a) Data processing (b) Expression classes (c) Problem Context

Figure 1. (a) We show the flow of data processing of F^{2}ED dataset. (b) F^{2}ED has 54 different facial expression classes, and we organize them into four large classes. (c) F^{2}ED dataset can be applied to various problem contexts. ER-SS: Expression recognition in the standard setting, ER-UE: Expression recognition with unbalanced expression, ER-UP: Expression recognition with unbalanced poses, ER-ZID: Expression recognition with zero-shot ID.

3. Fine-Grained Facial Expression Database

To the best of our knowledge, we contribute the largest fine-grained facial expression dataset to the community. Specifically, our F^{2}ED dataset has the largest number of images (totally 219719 images) with 119 identities and 54 kinds of fine-grained facial emotions. Each person is captured from four different views of cameras as shown in Fig. 3. Furthermore, in Tab. 1, our dataset is compared against the existing dataset – CK+, JAFFE, FER2013, KDEF. We show that our F^{2}ED is orders of magnitude larger than these existing datasets in terms of expression classes and number of total images.

3.1. The collection of F^{2}ED

We create the F^{2}ED dataset in 3 steps as in Fig. 1(a).

Data Collection. It takes us totally six months to collect video data. We invite more than 200 different candidates who are unfamiliar with our research topics. Each candidate is captured by four cameras placed at four different orientations to collect videos for persons as shown in Fig. 3 (a). The four orientations are front, half left, half right and bird view. The half left and half right cameras have a horizontal angle of 45 degrees with the front of the person, respectively. The bird view camera has a vertical angle of 30 degrees with the front of the person. Each camera takes 25 frames per second. The whole video capturing process is designed as a normal conversation between the candidate and two psychological experts. Totally, we aim at capturing 54 different types of expressions [17], *e.g.*, acceptance, angry, bravery, calm, disgust, envy, fear, neutral and so on. The conversation will follow some scripts which are calibrated by psychologists, and thus can induce/inspire one particular type of expression successfully conveyed by the candidates. For each candidate, we only save 5 minutes' video segment for each type of emotion.

Data Processing. With gathered expression videos, we further generate the final image dataset by human review, key images generation and face alignment. Specifically, the human review step is very important to guarantee the general quality of recorded expressions. Three psychologists are invited to help us review the captured emotion videos. Particularly, each captured video will be labeled by these psychologists. We only save the videos that have consistent labels by the psychologists. Thus totally about 119 identities' videos are preserved finally. Then key frames are extracted from each resulting video and face detection and alignment are conducted by the toolboxes of Dlib and MTCNN [36] over each frame. Critically, the face bounding boxes are cropped from the original images and resized to a resolution of 256×256 pixels. Finally we get the dataset F^{2}ED of totally 219719 images.

3.2. Statistics and Meta-information of F^{2}ED

Data Information. There are 4 types of face information in our dataset, including person identity, facial expression, pose and landmarks.

Person Identity. Totally we have 119 persons, including 37 male and 82 female aging from 18 to 24. Most of them are university students. Each person expresses his/her emotions under guidance and the video is taken when the person's emotion is observed.

Facial expression. Our dataset is composed of 54 types of emotions, based on the theory of Lee [17]. In this work, it expands the emotion set of Plutchik by including more complex mental states based on seven eye features. The seven features include temporal wrinkles, wrinkles below eyes, nasal wrinkles, brow slope, brow curve, brow distance and eye apertures. The 54 emotions can be clustered

dataset	#expression	#subject	#pose	#image	#sequence	Resolution	Pose list	Condition
CK+	8	123	1	327	593	490×640	F	Controlled
JAFFE	7	10	1	213	-	256×256	F	Controlled
FER2013	7	-	-	35887	-	48×48	-	In-the-wild
KDEF	7	140	5	4900	-	562×762	FL,HL,F,FR,HR	Controlled
F^2ED	54	119	4	219719	5418	256×256	HL,F,HR,BV	Controlled

Table 1. Comparison F^2ED with existing facial expression database. In the pose list, F : front, FL : full left, HL: half left, FR: full right, HR: half right, BV: bird view

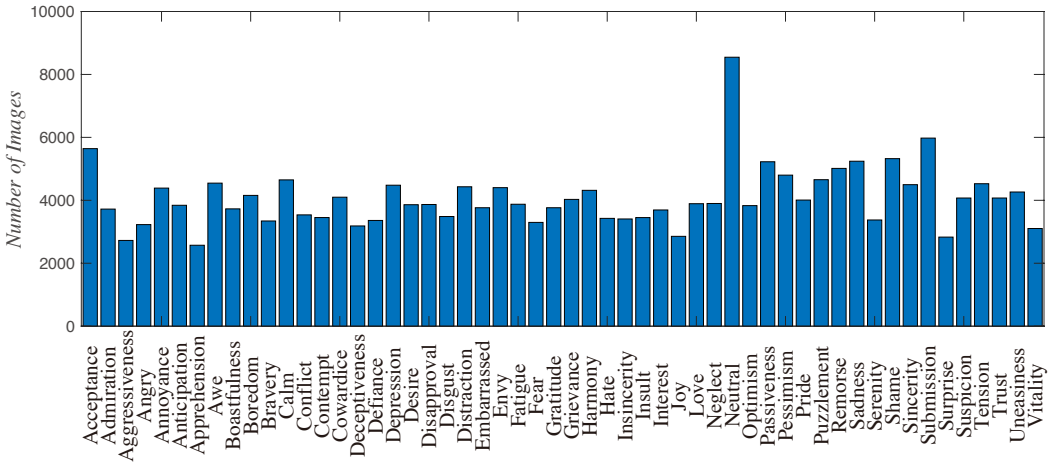
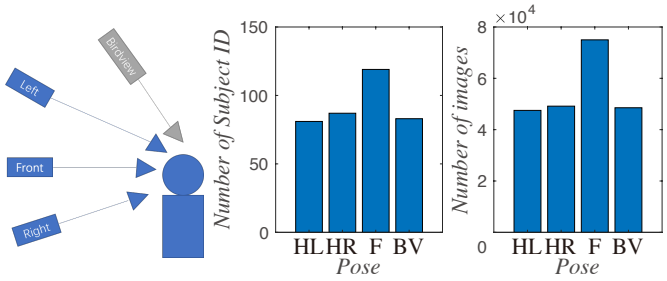


Figure 2. Image distribution of different expressions.



(a) (b)

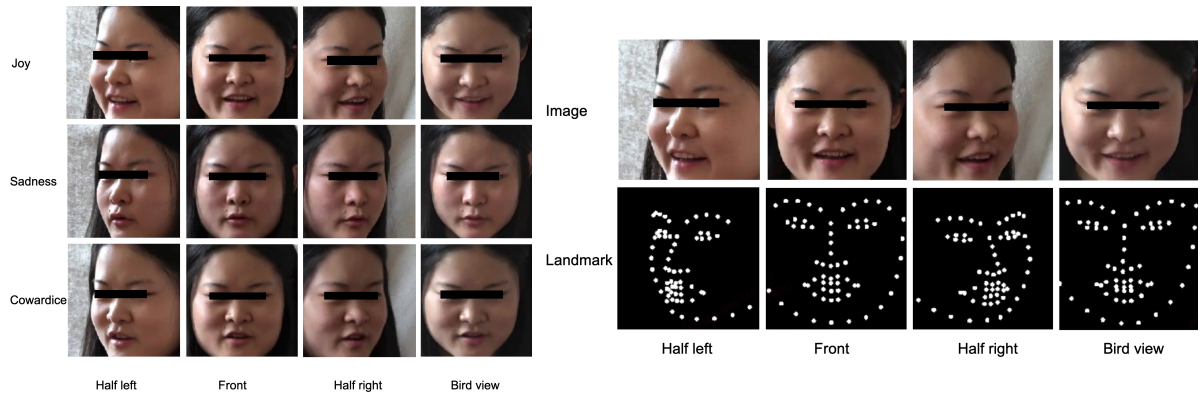
Figure 3. (a) Cameras used to collect facial expressions. (b) Distributions of subject ID and images over poses.

into 4 groups by k-means clustering algorithm as shown in Fig. 1(b). We also compute data distribution in Fig. 2.

Pose. As an important type of meta-information, poses often cause facial appearance changes. In real world applications, facial pose variations are mainly introduced by the relative position and orientation changes of the cameras to persons. In F^2ED, we collect videos from 4 orientations: half left, front, half right and bird view. Fig. 4(a) gives some examples of the F^2ED of different poses. In F^2ED we have 47053 half left, 49152 half right, 74985 front and 48529 bird view images. The distributions of subject ID and image number over poses are compared in Fig. 3 (b).

Facial Landmarks. Facial landmarks define the contour of facial components, including eye, nose, mouth and cheek. First we extract the facial landmarks with 68 points into position annotation text files by the Dlib. Then we convert the landmark position text file into images in a mask style. The example landmark images are shown in Fig. 4(b).

Tab. 1 shows the comparison between our F^2ED with existing facial expression database. As shown in the table, our dataset contains 54 subtle expression types, while other datasets only contain 7 or 8 expression types. For the person number, CK+, KDEF and F^2ED are nearly the same. The current public facial expression datasets are usually collected in two ways: in the wild or in the controlled environment. The FER2013 is collected in the wild, so the number of pose can not be determined. The rest datasets are collected in a controlled environment, where the number of pose for CK+ and JAFFE is 1, KDEF is 5 and F^2ED is 4. Our F^2ED is the only one that contains the bird view pose images which is very useful in real world scenario. For image number, F^2ED contains 219719 images, which is 6 times larger than the second largest dataset. All datasets have a similar resolution except FER2013 which has only a 48×48 resolution. CK+ and F^2ED are generated from 593 video sequences and 5418 video sequences.



(a) Face examples (b) Facial landmark examples

Figure 4. (a) There are some facial examples of F^{2}ED with different poses and emotions. (b) We give the facial landmark examples as the meta-information of F^{2}ED.

4. Learning on F^{2}ED

4.1. Learning tasks

In the F^{2}ED, we consider the expression learning over different types of variants as shown in Fig. 1(c); and further study the influence of different poses and subjects over the FER. To the best of our knowledge, this is the first exploration on this type of tasks. Particularly, we are interested in the following tasks for this dataset.

Expression recognition in the standard setting (ER-SS). The first and most important task is to directly learn the supervised classifiers on F^{2}ED. Particularly, as shown in Fig. 3(b) and Fig. 2, our dataset has balanced number of pose and emotion classes. We thus randomly shuffle our dataset and split it into 175000, 19719 and 25000 images for the train, validation and test set, respectively. The classifiers should be trained and validated on the train and validation sets, and predicted over the test set.

Expression recognition with unbalanced expression distribution (ER-UE). We further compare the results of learning classifiers with unbalanced facial expressions. In real word scenario, some facial expressions are rare, *e.g.*, cowardice. Thus it is imperative to investigate the FER in such an unbalanced expression setting. Specifically, we take 20% of total facial expressions as the rare classes. Among these rare classes, 90% of the images are kept as the testing instances, the rest 10% are used as the train set. The other 80% classes are treated as the normal emotion classes, and all of them are used for training. Thus, totally we have 178989 and 140730 images for the train and test set, respectively. For expression type analysis, there are 54 expression types in train set and 11 expression types in test set. On average, the occurrence frequency of testing expression class is only $1/10$ of that of training classes. In our setting, we assume that the model works with the prior knowledge that there are 54 rather than 11 expression classes in testing, which makes the chance of ER-UE task keep $1/54$.

Expression recognition with unbalanced poses (ER-UP). The learning task is further conducted with unbalanced poses. In this setting, we assume that the half left pose is rare in the train set. Thus the 10% of the half left pose images are used as the train set, and the rest 90% are used as test set. The other three types of poses – the half right, front, bird view pose images are used as the train set. Thus we get 177372 training images and 42347 testing images. For pose type analysis, there are 4 poses in train set and 1 pose in test set. This task aims to predict the expressions with rare poses in training set.

Expression recognition with zero-shot ID (ER-ZID). We aim at recognizing the expression types of the persons that have not been seen before. Particularly, we randomly pick the images from 21 and 98 persons as train and test set, respectively. This results in 189306 training and 30413 testing images. The task is to recognize expressions with zero-shot ID, referring to the disjoint subject ID in train and test sets. This enables us to verify whether the model can learn the person invariant feature for emotion classification.

4.2. Learning methods

We propose an end-to-end framework to address the four learning tasks in Fig. 5. Particularly, to tackle the issues of learning unbalanced number of images, our key idea is to employ the GAN based models for data augmentation to produce balanced training set. Our framework has the components of Facial Pose GAN (FaPE-GAN), and Face classification Networks (Fa-Net). The former one is an image synthesis network, and the latter is a classification network.

FaPE-GAN. It is trained by a combination of the training images and synthesized face images of new poses. The facial poses are normally represented by a landmark set. As shown in Fig. 5, this network firstly takes the face image $\mathbf{I}_i$ and the pose image $\mathbf{I}_{\mathcal{P}}$ as input, then the generator pro-

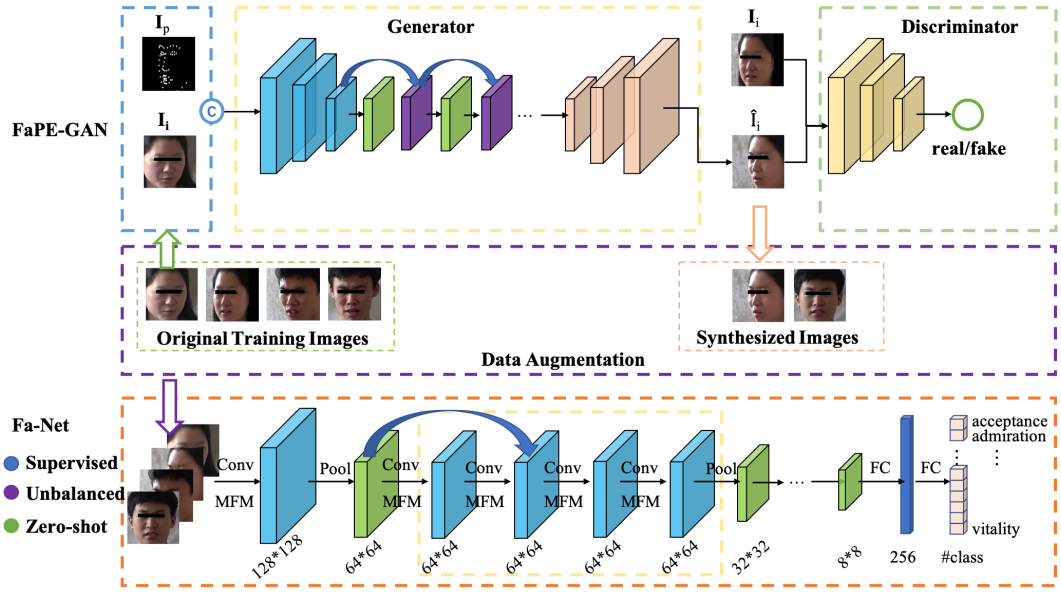


Figure 5. Overview of our framework. It includes the FaPE-GAN and Fa-Net component. FaPE-GAN can synthesize face images with input image and target pose. The Fa-Net is the classification network which is trained by the augmented face images and original face images. The Fa-Net can be applied in supervised, unbalanced and zero-shot learning.

duces the fake image $\hat{\mathbf{I}}_i$ of the same person with the pose of $\mathbf{I}_\mathcal{P}$, *i.e.*, $\hat{\mathbf{I}}_i = G_{FaPE}(\mathbf{I}_i, \mathbf{I}_\mathcal{P})$, and the discriminator tries to differentiate the fake target image $\hat{\mathbf{I}}_i$ from the real input image $\mathbf{I}_i$. Despite the pose may be changed in $\hat{\mathbf{I}}_i$, our FaPE-GAN still aims to keep the face identity of $\mathbf{I}_i$. Critically, we introduce the adversarial loss as,

$$\min_G \max_D \mathcal{L}_{GAN} = \mathbb{E}_{\mathbf{I_i}\sim p_d(\mathbf{I}_i)}\left[\log D\left(\mathbf{I_i}\right)\right] \quad (1)$$

$$+\left[\log\left(1 - D\left(G_{FaPE}\left(\mathbf{I_i}, \mathbf{I}_\mathcal{P}\right)\right)\right)\right] \quad (2)$$

where $p_d(\mathbf{I}_i)$ are the distributions of real images $\mathbf{I}_i$. The training process iteratively updates the parameters of generator G_{FaPE} and discriminator D. The generator loss can be formulated as,

$$\mathcal{L}_{G_{FaPE}} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{L_1} \quad (3)$$

where we have $\mathcal{L}_{L_1} = \mathbb{E}_{\mathbf{I}_t\sim p_d(\mathbf{I}_t)}\left[\left|\mathbf{I}_t - \hat{\mathbf{I}}_i\right|\right]$, $\mathbf{I}_t$ is the real target image and $\hat{\mathbf{I}}_i = G_{Dec}\left(G_{Enc}\left(\mathbf{I}_i, \mathbf{I}_\mathcal{P}\right)\right)$ is the reconstructed image, with the input image $\mathbf{I}_i$ and facial pose $\mathbf{I}_\mathcal{P}$. [24]. The hyperparameter λ is used to balance the two terms. The discriminator loss is formulated as, $\mathcal{L}_D = -\mathcal{L}_{GAN}$. The training process iteratively optimizes the loss functions of $\mathcal{L}_{G_{FaPE}}$ and $\mathcal{L}_D$. Fig. 6 shows two examples generated by FaPE-GAN .

Fa-Net. The same classification network is utilized to address all the four learning tasks in Sec. 4.1. Particularly, the backbone network is LightCNN [32]. The G_{FaPE} can synthesize plenty of additional face images in alleviating the issues of unbalanced training images. The augmented faces and original input faces are thus used to train our classification network.

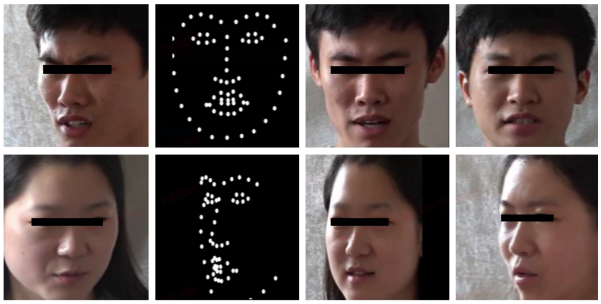


Figure 6. GAN output examples

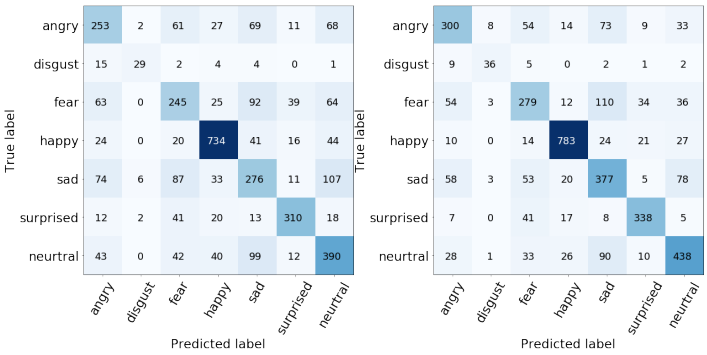
5. Experiments

Extensive experiments are conducted on F^2ED to evaluate the learning tasks defined in Sec. 4.1. Furthermore, the tasks of facial emotion recognition are also evaluated on FER2013 and JAFFE dataset.

Implementation details. The λ is set to 10, and the Adam optimizer is used in learning the FaPE-GAN with the learning rate of $2e-4$. The β_1 and β_2 are set as 0.5 and 0.999 respectively. The training epoch number is set to 100. For the facial expression classification network, We use the SGD optimizer with a momentum of 0.9 and decrease the learning rate by 0.457 every 10 steps. The max epoch number is

Model	Acc.
Bag of Words [13]	67.4%
VGG+SVM [7]	66.3%
GoogleNet [8]	65.2%
Mollahosseini *et al* [25]	66.4%
DNNRL [10]	70.6%
Attention CNN [23]	70.0%
Fa-Net	71.1%

Table 2. Accuracy on FER2013 test set in supervised learning setting



(a) (b)

Figure 7. (a) The confusion matrix on FER 2013 for Fa-Net without pre-training. (b) The confusion matrix on FER2013 for Fa-Net pre-trained on F^2ED

set to 100. The learning rate and batch size varies depending on the dataset size. To train the classification model, we set the learning rate/batch size as 0.01/128, $2e-3$/64 and $5e-4$/32, on F^2ED, FER2013 and JAFFE, respectively.

5.1. Results on FER2013 dataset

Settings. Following the setting of ER-SS, we conduct the experiments on FER2013 by using the entire 28709 training images and 3589 validation images to train/validate our model, which is further tested on the rest 3589 test images. The FER classification accuracy is reported as the evaluation metric to compare different competitors.

Competitors. Our model is compared against several competitors, including Bag of Words [13], VGG+SVM [7], GoogleNet [8], Mollahosseini *et al* [25], DNNRL [10] and Attention CNN [23]. Classifiers based on hand-crafted features, or specially designed architectures for FER, are investigated here. These methods can achieve the state-of-the-art results on this dataset.

Results on FER2013. To show the efficacy of our dataset, our classification network – Fa-Net is pre-trained on our F^2ED, and then fine-tuned on the training set of FER2013 dataset. The results show that our model can achieve the accuracy of 71.1%, which is superior to other state-of-the-art methods, as compared in Tab. 2. Tab. 4 shows that the Fa-Net pre-trained on F^2ED can improve the expression recognition performance by 8.8% comparing to the one without pre-training. The confusion matrix in Fig. 7 shows that pre-training increases the scores on all expression types. It demonstrates that the F^2ED dataset with large expression variations from more persons can pre-train a deep network with good initialization parameters. Note that our Fa-Net is not specially designed for FER task, since our Fa-Net is built upon the backbone – LightCNN, one typical face recognition architecture.

5.2. Results on JAFFE dataset

Settings. For the setting of ER-SS, we follow the split setting of the deep-emotion paper[23] to use 120 images for training, 23 images for validation, and keep totally 70 images for test (7 emotions per face ID).

Model	Acc.
Fisherface[1]	89.2%
Salient Facial Patch[11]	92.6%
CNN+SVM[30]	95.3%
Attention CNN[23]	92.8%
Fa-Net	95.7%

Table 3. Accuracy on JAFFE test set in supervised learning setting.

Competitors. Our model is compared against several competitors, including Fisherface[1], Salient Facial Patch [11], CNN+SVM[30] and Attention CNN [23]. These methods are tailored for the tasks of FER.

As listed in Tab. 3, our model achieved the accuracy of 95.7%, outperforming all the other competitors. Remarkably, our model surpasses the Attention CNN by 2.9% in the same data split setting. The accuracy of CNN+SVM is slightly lower than our model by 0.4%, even though their model is trained and tested on the entire dataset. This shows the efficacy of our dataset in pre-training the network. Tab. 4 further shows that Fa-Net pre-trained on the F^2ED has clearly improved the performance by 12.8%. The confusion matrix in Fig. 8 shows that the pre-trained Fa-Net only makes 3 wrong predictions and surpasses the one without pre-training on all expression types.

5.3. Results on F^2ED

Results on our dataset. We conduct the four different learning tasks on our dataset, namely, supervised (ER-SS), unbalanced expression (ER-UE), unbalanced pose (ER-UP) and zero-shot ID (ER-ZID) by the data split setting described in Sec. 4.1. Note that, since our Fa-Net is built upon the general face recognition backbone – LightCNN, it can thus be served as the main network in our experiments.

ER-SS task. Our model has achieved the accuracy of 73.6% as shown in Tab. 5. It shows that our F^2ED is well annotated so it can be used for classification task. Consider-

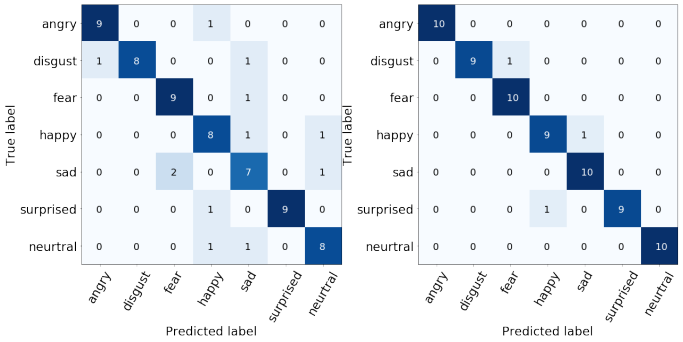


Figure 8. (a) The confusion matrix on FER 2013 for Fa-Net without pre-training. (b) The confusion matrix on FER2013 for Fa-Net pre-trained on F^{2}ED

Dataset	Pre-trained	Acc.
FER2013		62.3%
	✓	71.1%
JAFFE		82.9%
	✓	95.7%

Table 4. Results of the Fa-Net model with and without pre-trained on our F^{2}ED.

ing the large scale of the facial expression dataset, this performance is already very good and difficult to obtain which demonstrates that LightCNN is a good backbone for facial recognition. By using FaPE-GAN for data augmentation, the performance of our model is further improved by 0.9% comparing to the Fa-Net without GAN, which means that GAN is useful to generate more diversified examples for training.

ER-UE task. The accuracy of direct classification is 30.8% as shown in Tab. 5. This shows that the propose ER-UE task is very difficult, as the FER task greatly suffers from the unbalanced emotion data. Particularly, in our setting, only 10% examples from the 11 facial expression types appear in the training set, and the classifiers are thus confused by the other 43 emotion classes in the training stage. Furthermore, we also show that the data augmentation strategy endowed by our FaPE-GAN can indeed help to improve the performance of FER: the performance is improved by 3.5% which is larger than the 0.9% improvement in supervised learning setting. This indicates that the data augmentation is more effective in the data sparse condition such as unbalanced learning.

ER-UP task. Towards this task, our Fa-Net can hit the accuracy of 39.9% as shown in Tab. 5. Again, we argue that the proposed ER-UP is a very hard task, since this accuracy is only slightly better than the performance of ER-UE. This shows that the unbalanced pose data may also negatively affect the performance of FER task. Essentially, there are 54 types of expressions which are more diversified than the pose. Our data augmentation can still work in such a setting, and the synthesized data can help to train the Fa-Net, and alleviate the problem of unbalanced poses. As a result, it improves the performance of Fa-Net by 3.6%.

model/acc	ER-SS	ER-UE	ER-UP	ER-ZID
Fa-Net	72.7	27.3	36.3	6.7
FaPE-GAN+Fa-Net	73.6	30.8	39.9	7.1

Table 5. Accuracy on F^{2}ED for Fa-Net with and without data augmentation in supervised(ER-SS), unbalanced expression(ER-UE), unbalanced pose(ER-UP) and zero-shot ID(ER-ZID) setting

ER-ZID task. Surprisingly, the learning task proposed in this setting is the most challenging one compared with the other learning tasks. As shown in Tab. 5, we notice that our model only achieves an accuracy of 7.1% while the chance in fact is 1.9% ($\frac{1}{54}$ as described before), since the zero-shot task is much more difficult than the unbalanced task. This indicates that the generalization ability of FER is subject to other persons that the model has never seen before. Actually, this is the most desirable property of the FER model, since one can not assume the faces of test persons always appear in the training set. In our ER-ZID task, only 21 persons in the test set are never seen in the training set. Interestingly, our FaPE-GAN based data augmentation still contributes a 0.4% performance improvement over the baseline. This suggests the data augmentation may be still a potential useful strategy to facilitate the training of classification network.

Overall, our classification model with FaPE-GAN based data augmentation has clearly surpasses the one without FaPE-GAN on all 4 task types.

6. Conclusion

In this work, we introduce F^{2}ED, a new facial expression database containing 54 different emotion types and more than 200k examples. Furthermore we propose an end-to-end deep neural network based facial expression recognition framework, which uses a facial pose generative adversarial network to augment the data set. We perform supervised, zero-shot and unbalanced learning tasks on our F^{2}ED dataset, and the results show that our model has achieved the state-of-the-art. Subsequently, we fine-tune our model pre-trained on F^{2}ED on the existing FER2013 and JAFFE database, and the results demonstrate the efficacy of our F^{2}ED dataset.

References

[1] Z. Abidin and A. Harjoko. A neural network based facial expression recognition using fisherface. *International Journal of Computer Applications*, 59(3), 2012. 5.2

[2] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. Modeling stylized character expressions via deep learning.

In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016. 1

[3] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573. IEEE, 2005. 1

[4] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo. 3d facial expression recognition using sift descriptors of automatically detected keypoints. *The Visual Computer*, 27(11):1021, 2011. 1

[5] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016. 1

[6] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 2.1

[7] M.-I. Georgescu, R. T. Ionescu, and M. Popescu. Local learning with deep and handcrafted features for facial expression recognition. *arXiv preprint arXiv:1804.10892*, 2018. 5.1

[8] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis. Deep learning approaches for facial emotion recognition: A case study on fer-2013. In *Advances in Hybridization of Intelligent Methods*, pages 1–16. Springer, 2018. 2.3, 5.1

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2.1

[10] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao. Deep neural networks with relativity learning for facial expression recognition. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016. 5.1

[11] S. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2015. 5.2

[12] C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 2.3

[13] R. T. Ionescu, M. Popescu, and C. Grozea. Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on challenges in representation learning, ICML*, 2013. 5.1

[14] T. Kanade, Y. Tian, and J. F. Cohn. Comprehensive database for facial expression analysis. In *fg*, page 46. IEEE, 2000. 1, 2.2

[15] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015. 1, 2.1

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 1, 2.3

[17] D. H. Lee and A. K. Anderson. Reading what the mind thinks from how the eye sees. *Psychological Science*, 28(4):494, 2017. 3.1, 3.2

[18] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *ICCV*, 2017. 1

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 1

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010. 1, 2.2

[21] D. Lundqvist, A. Flykt, and A. Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630, 1998. 2.2

[22] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998. 1, 2.2

[23] S. Minaee and A. Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv preprint arXiv:1902.01019*, 2019. 1, 2.1, 5.1, 5.2

[24] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv: Learning*, 2014. 2.1, 4.2

[25] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. 5.1

[26] C. Pierre-Luc and C. Aaron. Challenges in representation learning: Facial expression recognition challenge, 2013. 1, 2.2

[27] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5399–5408, 2017. 1

[28] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–667, 2018. 2.1, 2.3

[29] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009. 1

[30] Y. Shima and Y. Omori. Image augmentation for classifying facial expression images by using deep neural network pre-trained with object image database. In *Proceedings of the 3rd International Conference on Robotics, Control and Automation*, pages 140–146. ACM, 2018. 5.2

[31] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 365–374. ACM, 2017. 2.1

[32] W. Xiang, H. Ran, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics Security*, PP(99):1–1, 2015. 4.2

[33] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, 9(2):255–270, 2018. 2.3

[34] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018. 2.1

[35] F. Zhang, T. Zhang, Q. Mao, and C. Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018. 1, 2.1, 2.3

[36] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2.1, 3.1

班级 1602051

学号 16020510038

本科毕业设计（论文）

外文资料翻译

毕业设计题目	基于单幅图像的面部表情生成算法研究
外文资料题目	Facial Expression Database
学院	人工智能学院
专业	智能科学与技术
学生姓名	邢博伟
指导教师姓名	毛莎莎

一个用于端到端的多姿态面部表情数据库表情识别

摘要

近年来，随着深度学习技术的发展，面部表情识别的研究取得了很大的进展，但仍存在一些典型的挑战性问题，如丰富的面部表情和姿态的多样性等。为了解决这些问题，我们开发了一种新的面部表情识别(FER)框架，将面部姿态包含到我们的图像合成和分类过程中。这项工作有两个主要的新奇之处。首先，我们创建了一个新的面部表情数据集，包含超过 20 万的图像，119 个人，4 个姿势和 54 个表情。据我们所知，这是第一个用于表情识别的面部表情识别数据集。它也是第一个足够大的数据集，可以在不平衡的姿态、表达式和零镜头主题 ID 上验证 FER 任务。其次，我们提出了一个面部姿态生成对抗网络(FaPE-GAN)来合成新的面部表情图像，以扩充训练数据集，然后学习一个基于 LightCNN 的表情分类的 Fa-Net 模型。最后，我们提出了四种新的学习任务。实验结果验证了该方法的有效性。

一、 简介

面部表情[5]作为最重要的面部属性，反映了一个人的情绪状态，包含着有意义的交流信息。面部表情识别在心理学、医学、安全、教育等领域有着广泛的应用。在心理学中，它可以用于分析心理痛苦的抑郁识别。另一方面，检测学生的注意力或沮丧感也有助于改进教育方法。

面部表情识别主要包括四个步骤:人脸检测、人脸对齐、特征提取和面部表情分类。(1) 在第一步中，从图像中检测人脸，每个人脸都由一个边框标记。(2) 在第二步中，生成脸部地标来对齐脸部。(3) 第三步是手工提取包含面部相关信息的特征，如 SIFT、[4]Gabor 小波[3,22]、LBP [29]等或通过神经网络学习。（4）第四步，利用 SVM、KNN、MLP 等多种分类器进行面部表情分类。

最近，深度神经网络的复兴为一些视觉任务提供了人类水平的性能，如对象分类、检测和分割[19,18,27]。受此启发，一些针对面部表情识别的深度网络方法[15,23,35]被提出。在 FER 任务中，根据 Ekman 的理论，面部表情通常被认为

包含六种离散的基本情绪:愤怒、厌恶、恐惧、高兴、悲伤和惊讶。加上一个额外的中性情绪，这七种情绪构成了最常见的情绪数据集的主要部分，包括CK+[20,14]、JAFEE[22]、FER2013[26]和FERG[2]。

然而，事实上，一个最具挑战性的问题是缺乏一个大数据集的高质量图像，可以用来训练深层网络和研究影响因素的FER任务。这些数据集的另一个缺点，如JAFFE和FER2013数据集，是表达情绪的多样性太少，不能真正表达现实生活中多变的面部表情情绪。

为此，我们创建了一个包含54种情绪类型的新数据集f2ed(细粒度面部表情数据库)，其中包括大量的情绪，这些情绪都有细微的变化，如平静、尴尬、骄傲、紧张等。此外，我们还考虑了人脸姿态变化对表情识别的影响，并将姿态作为每个表情的另一个属性引入。四个方向(姿势)包括前面，一半左，一半右和鸟瞰图被标记，每个都有一个平衡的例子，以避免训练偏差。

在这个数据集上，我们可以进一步研究姿态、表达式和主题id如何影响FER性能。关键的是，我们在这个数据集上提出了四个新的学习任务，如图1(c)所示。它们与标准平衡设置表达式识别 (ER-SS)，表达式 (ER-UE)，不平衡 不平衡造成 (ER-UP)，零样本学习ID (ER-ZID).与典型的零样本学习设置[16]相似，零样本ID设置意味着训练集中没有出现人的测试面。为了解决这四个学习任务，我们进一步设计了一个可以扩充训练数据的新框架，然后训练分类网络。在我们的数据集上进行了大量的实验，JAFEE [22]， FER2013[26]表明(1)我们的数据集足够大，可以将深度网络作为骨干网络进行预训练;(2)姿态不平衡、表情不平衡、零拍id确实对FER任务有负向影响;(3)数据扩充策略有助于学习更强大、性能更好的模型。这三点也是本文的主要贡献。

二、 相关工作

1、 人脸表情识别

基于神经网络的广泛的FER工作已经被提出[15,31,36]。Khorrami等人通过[15]训练CNN的FER任务，将学习到的特征可视化，发现这些特征与[6]中提出的FAUs有很强的对应关系。在FER上的Attention CNN[23]被提议通过增加一个空间变压器来聚焦于人脸最突出的部分。

在求解FER任务时，还研究了基于生成对抗网(GAN)[9]的模型。特别是，GAN

通常由发生器和鉴别器组成。为了减弱位姿和遮挡的影响，提出了位姿不变模型[35]。根据 GAN 生成不同的姿态和表情脸。Qian 等人提出了一种生成式对抗网络(GAN)，专门针对 reid 中的位姿归一化设计。Yan 等人提出了一种去表达模型，利用条件 cGAN[24]从源图像中生成中性表达图像，利用 GAN 中间层的残差信息对表达进行分类。

2、 已有数据集

CK+.

扩展的 Cohn-Kanade (CK+)数据库[20]是 CK 数据库[14]的更新版本。在 CK+数据库中，有来自 123 名受试者的 593 个视频序列。在 593 个视频序列中，有 327 个是根据 FACS 编码的情绪标签选择的。所选视频的最后一帧被标记为八种情绪之一:愤怒、轻蔑、厌恶、恐惧、快乐、悲伤、惊讶和中性。

JAFFE.

日本女性面部表情数据库[22]包含 213 张 256×256 像素分辨率的图像。这些照片是在一个受控的环境下从 10 名日本女模特身上拍摄的。每张图片都用以下 6 个情绪形容词中的一个来打分:愤怒、厌恶、恐惧、快乐、悲伤和惊讶。

FER2013.

2013 面部表情识别数据库[26]包含了 35887 张 48 分辨率的图像。这些图像是在野外拍摄的，这意味着更有挑战性的条件，如遮挡和姿势变化都包括在内。它们被列为上述七种情绪之一。数据集分为 28709 张训练图像、3589 张验证图像和 3589 张测试图像。

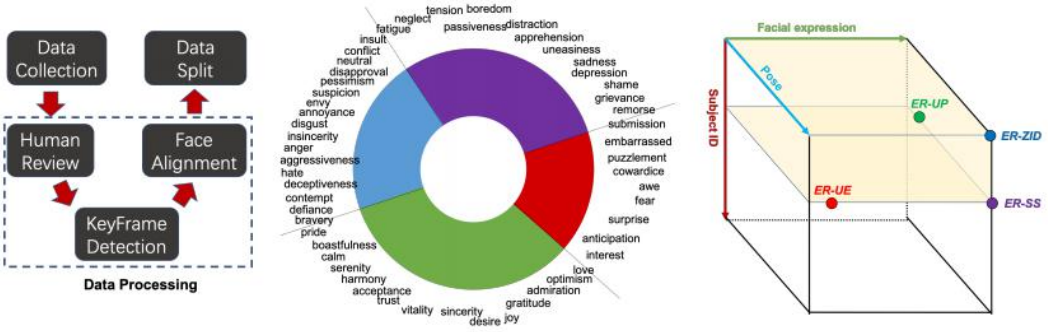
KDEF.

Karolinska 的情绪面孔[21]数据集包含 562×762 像素分辨率的 4900 张图像。图片来自 140 人(70 男，70 女)，5 个角度，7 种情绪。这些角度包括完全左侧面，半左侧面，前面，完全右侧面和半右侧面。情绪集包含 7 种表情:害怕、生气、厌恶、高兴、悲伤、惊讶和中性。

3、 学习模式

零镜头学习识别新的视觉类别，没有看到的标记训练例子[16]。该问题通常

通过将学习从源域转移到目标域来解决。描述一个新对象的语义属性可以用于零样本学习。徐等人提出了一种零镜头视频情感识别。在这篇文章中，我们提出了一个新颖的任务，针对的是不在训练集中的人。另一方面，阶级不平衡是一个常见的问题，尤其是在深度学习中[12,8]。我们第一次提出了一个足够大的数据集来帮助评估不平衡的姿势、表情和人员 ID 对 FER 任务的影响。为了缓解这一问题，我们研究了综合更多的数据，基于 GAN 的数据增强的灵感来自于最近的人的 ID[28]和面部表情识别[35]的工作。



(a) 数据预处理　　(b) 表情类别　　(c) 问题内容

图 1 所示。(a)展示了 f2ed 数据集的数据处理流程。(b) f2ed 有 54 个不同的面部表情类，我们把它们分成四个大类。(c) f2ed 数据集可应用于各种问题环境。ER-SS:标准设置下的表情识别，ER-UE:表情不平衡识别，ER-UP:表情姿势不平衡识别，ER-ZID:表情零角度识别。

三、 精细的面部表情数据库

我们所知,我们贡献最大的细粒度的面部表情数据集社区。具体地说,我们的数据集 F 2 ed 最多的图像(完全 219719 图像)与 119 年身份和 54 种细粒度的面部情绪。捕捉每个人从四种不同的相机视图,如图 3 所示。此外,在选项卡。1,我们的数据集是对现有数据集 CK +相比,贾菲,FER2013 KDEF。我们表明,F 2 ed 数量级比这些现有的数据集的表达类和数量的总图像。

1、F^2ED 集合

我们按照图 1(a)的步骤创建 F^2ED 数据集。

数据收集

我们一共花了 6 个月的时间来收集视频数据。我们邀请了 200 多个不熟悉我

们研究主题的候选人。如图 3 (a)所示，每个候选人被放置在四个不同方向的四个摄像头捕捉，以为人们收集视频。四个方向分别是前、半左、半右和鸟瞰图。半左半右的摄像头与人的前面分别有 45 度的水平角度。鸟瞰图相机与人的前面有一个 30 度的垂直角度。每台相机每秒拍摄 25 帧。整个视频的拍摄过程被设计成候选人和两位心理专家之间的正常对话。总的来说，我们的目标是捕捉 54 种不同类型的表情，如接受、生气、勇敢、冷静、厌恶、嫉妒、恐惧、中立等等。对话将遵循一些由心理学家校准的脚本，因此可以诱导/激发候选人成功传达的一种特定类型的表达。对于每个候选人，每种情绪我们只保存 5 分钟的视频片段。

数据处理

通过采集表情视频，通过人工点评、关键图像生成和人脸对齐，进一步生成最终的图像数据集。具体来说，人工检查步骤对于保证记录表达式的总体质量非常重要。三位心理学家被邀请来帮助我们回顾被捕捉到的情感视频。特别是，每一个捕捉到的视频都会被这些心理学家贴上标签。我们只保存心理学家标记一致的视频。最终总共保留了 119 个身份认证视频。然后从每个得到的视频中提取关键帧，通过每帧上的 Dlib 和 MTCNN[36]的工具箱进行人脸检测和对齐。关键的是，人脸包围框是从原始图像中裁剪出来的，并将其调整到 256×256 像素的分辨率。最后得到了 219719 幅图像的 F^2ED 数据集。

2、F^2ED 的统计数据和元信息

数据信息。在我们的数据集中有 4 种类型的面部信息，包括人的身份、面部表情、姿势和地标。

用户的身份。公司现有员工 119 人，其中男 37 人，女 82 人，年龄 18 ~ 24 岁。他们大多数是大学生。每个人都在指导下表达自己的情绪，当观察到这个人的情绪时，就会拍下视频。

面部表情。根据 Lee[17]的理论，我们的数据集由 54 种情绪组成。在这部作品中，它扩展了情绪分类的情感集，包括了基于七个眼睛特征的更复杂的精神状态。这七个特征包括颞部的皱纹、眼部以下的皱纹、鼻部的皱纹、眉毛的倾斜度、眉毛的曲线、眉毛的距离和眼孔。通过 k-means 聚类算法将 54 种情绪聚类为 4 组，如图所示图 1 (b)。我们还计算了图 2 中的数据分布。

dataset	#expression	#subject	#pose	#image	#sequence	Resolution	Pose list	Condition
CK+	8	123	1	327	593	490 × 640	F	Controlled
JAFFE	7	10	1	213	-	256 × 256	F	Controlled
FER2013	7	-	-	35887	-	48 × 48	-	In-the-wild
KDEF	7	140	5	4900	-	562 × 762	FL,HL,F,FR,HR	Controlled
F^2ED	54	119	4	219719	5418	256 × 256	HL,F,HR,BV	Controlled

表 1。与现有的面部表情数据库进行比较。在姿态列表中，F: 前面，FL: 全左，HL: 半左，FR: 全右，HR: 半右，BV:鸟瞰图

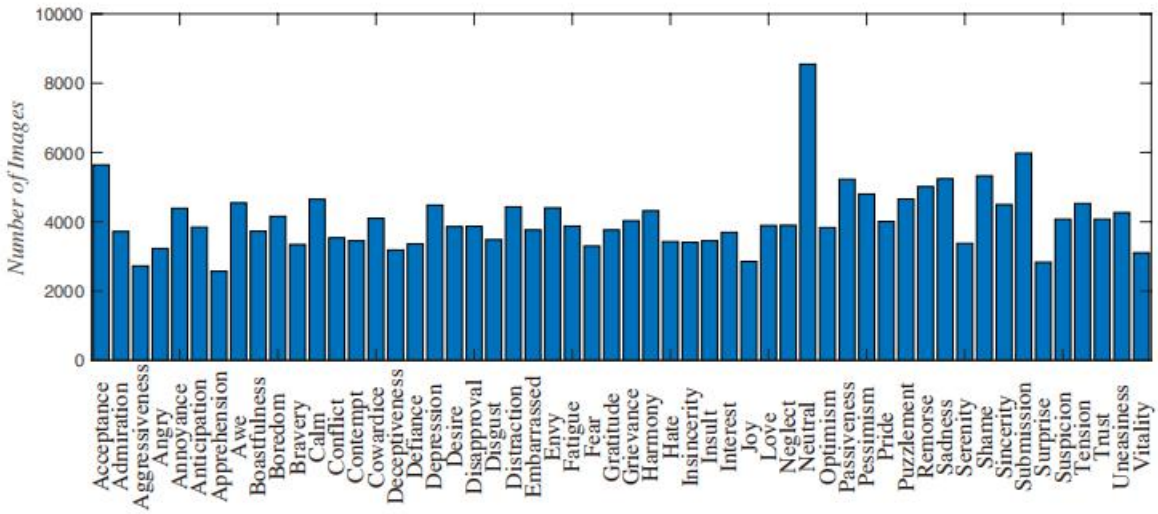


图 2 不同表情的图像分布

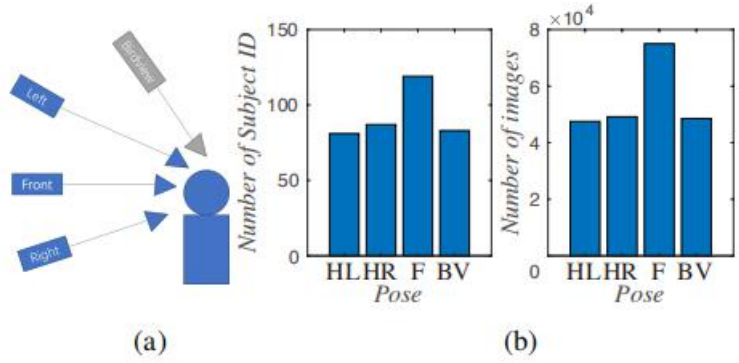


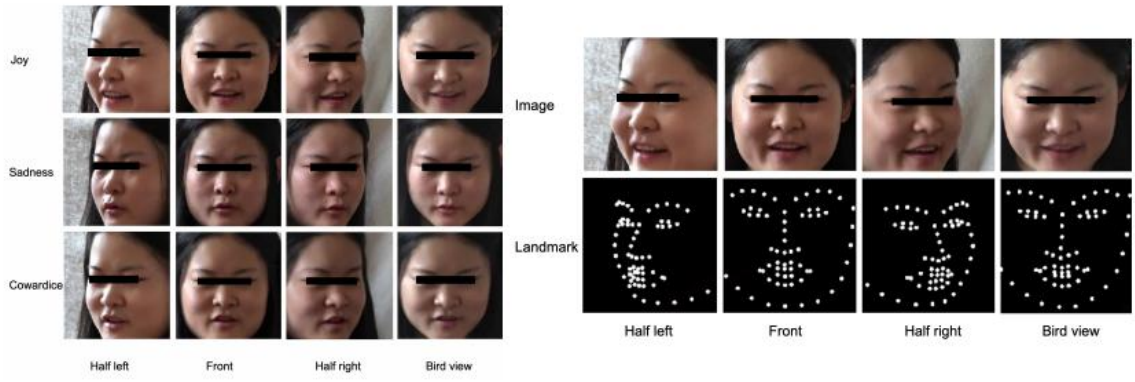
图 3 (a)用来收集面部表情的照相机 (b)拍摄对象 ID 和照片的分布

姿势。姿势作为一种重要的元信息类型，常引起面部表情的变化。在实际应用中，人脸姿态的变化主要是由相机相对位置和方向的变化引起的。在 F^2ED 中，我们收集了 4 个方向的视频:半左、前、半右和鸟瞰图。图 4(a)给出了不同姿态下的 F^2ED 的例子。在 F^2ED 我们有 47053 一半左，49152 一半右，74985 前面和 48529 鸟瞰图。图 3 (b)对比了受试者 ID 和图像编号在位姿上的分布。

面部地标。面部地标定义了面部成分的轮廓，包括眼睛、鼻子、嘴巴和脸颊。首先利用 Dlib 将 68 个点的人脸标志提取到位置标注文本文件中。然后将地标位置文本文件转换为掩模样式的图像。示例地标图像如图 4(b)所示。

表 1 显示了我们的F^2ED与现有面部表情数据库的比较。如表所示，我们的数据集包含 54 种微妙的表达式类型，而其他数据集只包含 7 或 8 种表达式类型。对于人员编号，CK+、KDEF 和F^2ED几乎相同。当前公众面部表情数据集的采集通常有两种方式:野外采集和受控环境采集。FER2013 是在野外采集的，所以姿态的数量无法确定。其余数据集在受控环境中收集，其中 CK+和 JAFFE 的 pose 数为 1,KDEF 为 5,f2ed 为 4。我们的F^2ED是唯一一个包含鸟瞰姿态图像，这是非常有用的现实世界场景。在图像数量上，F^2ED包含 219719 张图像，是第二大数据集的 6 倍。所有的数据集都有相似的分辨率，除了 FER2013 只有 48 - 48 的分辨率。CK+和F^2ED分别来自 593 个视频序列和 5418 个视频序列。

四、 学习F^2ED



（a）人脸样本　　（b）人脸标志样本

图 4 (a) F^2ED面部的有不同的姿势和情绪。(b)我们给出了面部地标的例子作为F^2ED的元信息。

1、 学习任务

在F^2ED中，我们考虑了图 1(c)所示的不同类型变异的表达学习;并进一步研究不同姿势、不同受试者对 FER 的影响。据我们所知，这是对这类任务的首次探索。特别是，我们对这个数据集的以下任务感兴趣。

标准设置中的表达式识别(ER-SS)。第一个也是最重要的任务是直接学习F^2ED上的监督分类器。特别是，如图 3(b)和图 2 所示，我们的数据集平衡了姿态和情绪类的数量。因此，我们随机洗牌我们的数据集，并把它分为 175000,19719 和 25000 图像的火车，验证和测试集，分别。分类器应在训练集和验证集上进行训练和验证，并在测试集上进行预测。

表达式分布不平衡的表达式识别(ER-UE)。我们进一步比较了学习不平衡面部表情分类器的结果。在真实的场景中，一些面部表情是很少见的，例如:懦弱。因此，在这样一个不平衡的表达式设置中，研究 FER 是很有必要的。具体来说，我们把 20%的面部表情作为稀有类。在这些罕见的类中，90%的图像作为测试实例，其余 10%作为训练集。其余 80%的类作为正常的情感类，全部用于训练。因此，我们总共有 178989 和 140730 的图像分别为火车和测试集。对于表达式类型分析，在训练集中有 54 种表达式类型，在测试集中有 11 种表达式类型。平均而言，测试表达式类的出现频率只有训练班的 1/10。在我们的设置中，我们假设模型在测试中有 54 个而不是 11 个表达式类的先验知识下工作，这使得 ER-UE 任务的机会保持 1/54。

具有不平衡姿势的表情识别(ER-UP)。学习任务进一步以不平衡姿势进行。在这种情况下，我们假设左半摆的姿势在火车上是罕见的。因此,10%的一半左图像作为训练集,作为测试集,其余 90%。其他三种类型的姿势对了一半,前面,鸟视图构成图像作为训练集。因此我们得到 177372 42347 训练图像和测试图像。在位姿类型分析中，训练集中有 4 个位姿，测试集中有 1 个位姿。

零样本学习 ID (ER-ZID)的表达式识别。我们的目标是识别以前没有见过的人的表情类型。特别地，我们随机选取了 21 人和 98 人的图像分别作为训练集和测试集。结果在 189306 训练和 30413 测试图像。任务是识别零枪击 ID 的表达式，指的是训练集和测试集中不相交的主题 ID。这使得我们可以验证模型是否可以学习人的不变特征进行情绪分类。

2、 学习方法

我们提出了一个端到端框架来解决图 5 中的四个学习任务。特别地，为了解决学习图像数量不平衡的问题，我们的关键思想是使用基于 GAN 的模型进行数据扩充，生成平衡的训练集。我们的框架包括面部姿态 GAN (FaPE-GAN)和面部分

类网络(Fa-Net)。前者是一个图像合成网络，后者是一个分类网络。FaPE-GAN。将训练图像与合成的新姿态人脸图像相结合进行训练。

面部姿态通常由地标集合表示，如图 5 所示，这个网络首先将人脸图像 I_i 和姿势图像 I_p 作为输出，然后生成器产生同一个人的假图像 $\hat{I}_i$ 的姿势 I_p，例如，$\hat{I}_i = G_{F_aPE}(\mathrm{I}_i, \mathrm{I}_p)$，鉴别器试图区分假目标图像 I_i 和真实输入图像 $\hat{I}_i$。尽管姿势可能会在 I_i 中改变，我们的 FaPEGAN 仍然致力于保持 I_i 的面部特征。在此基础上，我们引入了对抗性损失。

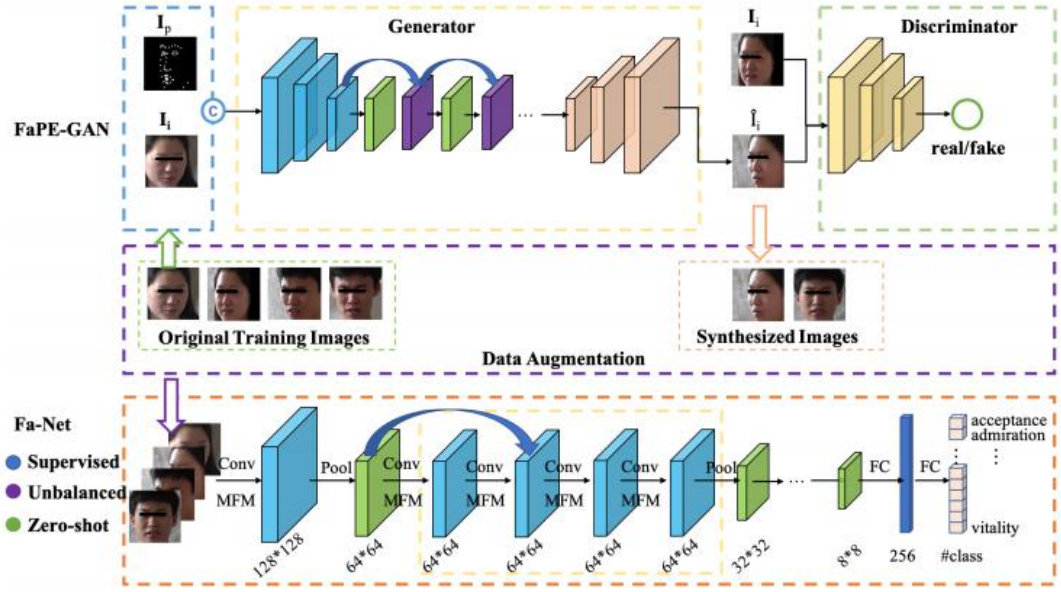


图 5。我们的框架概述。它包括 FaPE-GAN 和 Fa-Net 组件。通过输入图像和目标姿态，FaPE-GAN 可以合成人脸图像。Fa-Net 是由增强人脸图像和原始人脸图像训练而成的分类网络。该算法可用于监督学习、非平衡学习和零学习。

$$\min_G \max_D \mathcal{L}_{GAN} = \mathbb{E}_{\mathbf{I_i} \sim p_d(\mathbf{I_i})} [\log D(\mathbf{I_i})] \quad (1)$$

$$+ [\log (1 - D(G_{FaPE}(\mathbf{I_i}, \mathbf{I_P})))] \quad (2)$$

其中 $p_d(\mathrm{I}_i)$ 为真实图像 I_i 的分布。训练过程迭代更新发生器 G_{FaPE} 和鉴别器 D 的参数，发生器损耗可表示为

$$\mathcal{L}_{G_{FaPE}} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{L_1} \quad (3)$$

其中，我们有 $\zeta_{L_1} = E_{I_{t\sim P_d(I_t)}}\left[\left|I_t - \hat{I}_i\right|\right]$，$I_t$ 是那个真实的目标图像，而且 $\hat{I}_i = G_{D_{ec}}(\mathrm{G}_{E_{nc}}(\mathrm{I}_i, \mathrm{I}_p))$ 是重构图像，伴随着 1 图像 I_i 和面部姿势 I_p [24]。超参数 λ 被用作平衡两项。鉴别器损耗公式为: $\zeta_D = -\zeta_{GAN}$。训练过程迭代优化 ζ_{GF_aPE} 和 ζ_D 的损耗函数，如图 6 所示为 FaPE-GAN 生成的两个例子。

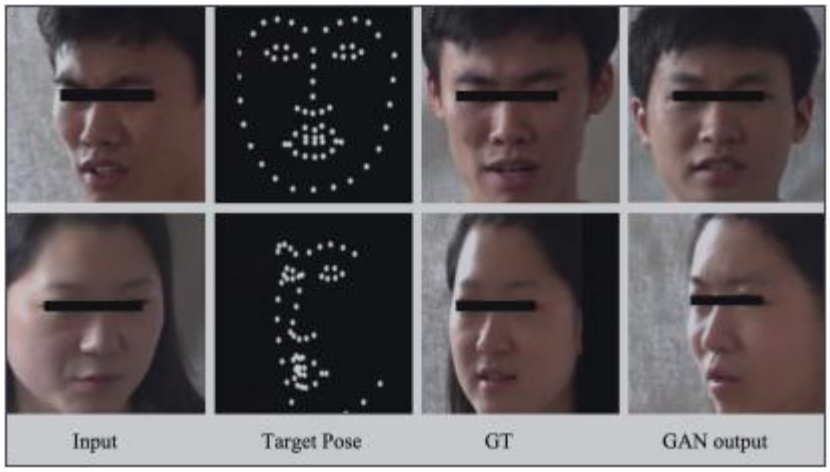


图 6 GAN 的输出样例

Fa-Net。使用相同的分类网络来处理第 4.1 节中的所有四个学习任务。特别是骨干网络是 LightCNN[32]。G_{FaPE} 可以合成大量附加的人脸图像，缓解了训练图像不平衡的问题。利用增广面和原始输入面对分类网络进行训练。

五、 实验

在 F^2ED 上进行了大量的实验来评估 4.1 节中定义的学习任务。此外，还利用 FER2013 和 JAFFE 数据集对人脸情感识别任务进行了评价。

Model	Acc.
Bag of Words [13]	67.4%
VGG+SVM [7]	66.3%
GoogleNet [8]	65.2%
Mollahosseini *et al* [25]	66.4%
DNNRL [10]	70.6%
Attention CNN [23]	70.0%
Fa-Net	71.1%

表 2 在监督学习环境下 FER2013 测试集的准确性

实现细节。λ是设置为 10,亚当优化器是用于学习的学习速率 FaPE-GAN 2 e-4。β1 和β2 分别设置为 0.5 和 0.999。训练历数设置为 100。对于面部表情分类网络，我们使用了动量为 0.9 的 SGD 优化器，每 10 步将学习率降低 0.457。最大历元数设置为 100。学习速率和批大小取决于数据集的大小。为了训练分类模型，我们将学习率/批量大小设置为 0.01/128,2e e 3/64 和 5e4 /32，分别在 F2ED, FER2013 和 JAFFE。

1、 关于 FER2013 数据集的结果

设置。以 ER-SS 为背景，我们在 FER2013 上进行实验，使用全部 28709 张训练图像和 3589 张验证图像对我们的模型进行训练/验证，并在剩下的 3589 张测试图像上进行测试。将 FER 分类精度作为比较不同竞争对手的评价指标。

竞争对手。我们的模型对比了几个竞争对手，包括 Bag of Words [13], VGG+SVM [7], GoogleNet [8], Mollahosseini 等[25],DNNRL[10]和 Attention CNN[23]。基于手工特征或特别为 FER 设计的架构的分类器在这里被研究。这些方法可以在这个数据集上获得最新的结果。

结果 FER2013。为了证明我们的数据集的有效性，我们的分类网络 Fa-Net 在我们的 f2ed 上进行了预训练,然后在 FER2013 数据集的训练集上进行了微调。结果表明，与表 2 相比，我们的模型可以达到 71.1%的准确率，优于其他最先进的方法。由表 4 可知，在 F^2ED 上进行预处理的 FaNet 与未进行预处理的 FaNet 相比，可以提高 8.8%的表达识别性能。图 7 中的混淆矩阵显示，训练前提高了所有表达类型的得分。结果表明，具有较大表达量的 F^2ED 数据集可以对具有良好初始化参数的深度网络进行预训练。请注意，我们的 Fa-Net 并不是专门为 FER 任务而设计的,因为我们的 Fa-Net 是建立在典型的人脸识别架构 LightCNN 的基础上的。

2、 JAFFE 数据集的结果

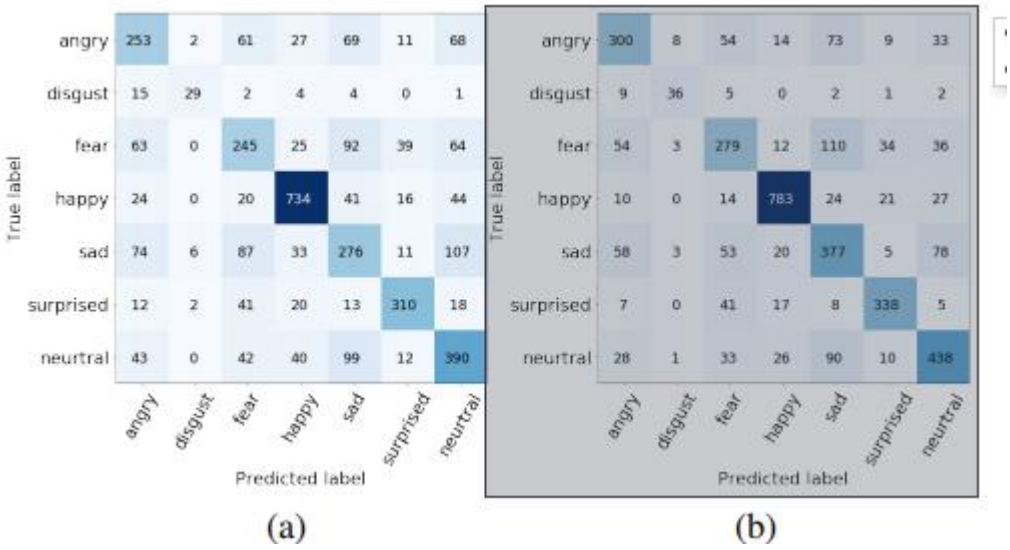


图 7　(a)未经培训的 Fa- Net 的 2013 年 FER 混淆矩阵。(b) Fa- Net 上的混淆矩阵在 F^2ED 上进行了预训练

Model	Acc.
Fisherface[1]	89.2%
Salient Facial Patch[11]	92.6%
CNN+SVM[30]	95.3%
Attention CNN[23]	92.8%
Fa-Net	95.7%

表 3 在监督学习环境下，JAFFE 测试集的准确性。

设置。对于 ER-SS 的设置，我们按照 deep-emotion paper[23]的 split 设置，使用 120 张图像进行训练，23 张图像进行验证，共保留 70 张图像进行测试(每张人脸 ID 7 张情绪)。

竞争对手。我们的模型与多个竞争对手进行了对比，包括菲舍尔面部[1]、显著面部补丁[11]、CNN+SVM[30]，注意 CNN[23]。这些方法是为 FER 的任务量身定制的。

如表 3 所示，我们的模型准确率达到了 95.7%，超过了其他所有竞争对手。值得注意的是，在相同的数据分割设置下，我们的模型比 CNN 的注意力高 2.9%。虽然 CNN+SVM 的模型经过了整个数据集的训练和测试，但其准确率比我们的模型略低 0.4%。这显示了我们的数据集在网络预训练方面的有效性。表 4 进一步表明，在 f2ed 上预先训练的 Fa-Net 性能明显提高了 12.8%。从图 8 中的混淆矩阵

可以看出，经过预处理的 Fa-Net 只做出了 3 个错误的预测，并且在所有的表达类型上都超过了未经预处理的 Fa-Net。

3、 F^2ED 的结果

数据集上的结果。我们在数据集上进行了四种不同的学习任务，即通过 4.1 节中描述的数据分割设置进行监督(ER-SS)、不平衡表达(ER-UE)、不平衡位姿(ER-UP)和零样本学习 ID (ER-ZID)。。请注意，由于我们的 Fa-Net 是建立在一般人脸识别中枢 LightCNN 的基础上，因此在我们的实验中，它可以作为主要的网络。

ER-SS 任务。我们的模型达到了 73.6%的准确率，如表 5 所示。这表明我们的 F2ED 有很好的注释，可以用于分类任务。考虑到人脸表情数据集的规模较大，这种性能已经很好，并且很难获得，这说明 LightCNN 是人脸识别的一个很好的主干。通过使用 FaPE-GAN 进行数据扩充，我们的模型的性能比不使用 GAN 的 Fa-Net 进一步提高了 0.9%，这意味着 GAN 可以产生更加多样化的训练样本。

ER-UE 任务。如表 5 所示，直接分类的准确率为 30.8%。这说明提出的 ER-UE 任务是非常困难的，因为 FER 任务在很大程度上受到情绪数据不平衡的影响。特别是在我们的设置中，11 种面部表情类型中只有 10%的例子出现在训练集中，因此分类器被训练阶段的另外 43 个情感类混淆了。此外，我们还证明了我们的 FaPE-GAN 所赋予的数据扩充策略确实有助于提高 FER 的性能:性能提高了 3.5%，大于监督学习设置下的 0.9%。这表明，在非平衡学习等数据稀疏条件下，数据扩充更有效。

ER-UP 任务。对于这个任务，我们的 Fa-Net 可以达到 39.9%的准确率，如表 5 所示。再次，我们认为提议的 ER-UP 是一项非常困难的任务，因为这种准确性仅略优于 ER-UE 的性能。这说明不平衡位姿数据也会对 FER 任务的性能产生负面影响。从本质上说，有 54 种表情比姿势更多样化。我们的数据扩充仍然可以在这样的设置下工作，并且合成的数据可以帮助训练 Fa-Net，并缓解不平衡位姿的问题。因此，它将 Fa-Net 的性能提高了 3.6%。

model/acc	ER-SS	ER-UE	ER-UP	ER-ZID
Fa-Net	72.7	27.3	36.3	6.7
FaPE-GAN+Fa-Net	73.6	30.8	39.9	7.1

表 5 具有和不具有监督(ER-SS)、不平衡表达式(ER-UE)、不平衡姿态(ER-UP)和零样本学习 ID(ER-ZID)设置的数据扩充功能的 Fa-Net 的 F^2ED 准确性

ER-ZID 任务。令人惊讶的是，在这种情况下提出的学习任务与其他学习任务相比是最具挑战性的。如表 5 所示，我们注意到我们的模型只有 7.1%的准确率，而实际的概率是 1.9%(如前所述为 154)，因为零扫描比不平衡任务要困难得多。这说明 FER 的泛化能力受制于模型中从未见过的其他人。实际上,拿来的这是最可取的属性模型,从一个不能假设测试人的面孔总是出现在训练集。在我们 ER-ZID 任务,只有 21 人在测试设置中从未见过培训 set.有趣的是，我们 FaPE-GAN 基础数据增加基线仍贡献 0.4%的性能提升。这表明，数据扩充可能仍然是一种潜在的有用的促进分类网络训练的策略。

总的来说，我们的基于 FaPE-GAN 数据扩充的分类模型明显优于没有数据扩充的分类模型所有 4 种任务类型的 FaPE-GAN。

六、 结论

在这项工作中，我们引入了 F^2ED 一个新的面部表情数据库，包含 54 种不同的情感类型和超过 200k 个例子。在此基础上，我们提出了一种基于端到端深度神经网络的人脸表情识别框架，该框架利用人脸姿态生成对抗网络来扩充数据集。结果表明，我们的模型达到了最先进的水平。随后，我们在现有的 FER2013 和 JAFFE 数据库上对 F^2ED 预训练模型进行了微调，结果证明了 F^2ED 数据集的有效性。

SinGAN: Learning a Generative Model from a Single Natural Image

Tamar Rott Shaham
Technion

Tali Dekel
Google Research

Tomer Michaeli
Technion

Figure 1: **Image generation learned from a single training image.** We propose *SinGAN*–a new unconditional generative model trained on a *single natural image*. Our model learns the image's patch statistics across multiple scales, using a dedicated multi-scale adversarial training scheme; it can then be used to generate new realistic image samples that preserve the original patch distribution while creating new object configurations and structures.

Abstract

We introduce SinGAN*, an unconditional generative model that can be learned from a single natural image. Our model is trained to capture the internal distribution of patches within the image, and is then able to generate high quality, diverse samples that carry the same visual content as the image. SinGAN contains a pyramid of fully convolutional GANs, each responsible for learning the patch distribution at a different scale of the image. This allows generating new samples of arbitrary size and aspect ratio, that have significant variability, yet maintain both the global structure and the fine textures of the training image. In contrast to previous single image GAN schemes, our approach is not limited to texture images, and is not conditional (*i.e. *it generates samples from noise). User studies confirm that the generated samples are commonly confused to be real images. We illustrate the utility of SinGAN in a wide range of image manipulation tasks.*

1. Introduction

Generative Adversarial Nets (GANs) [19] have made a dramatic leap in modeling high dimensional distributions of visual data. In particular, unconditional GANs have shown remarkable success in generating realistic, high quality samples when trained on class specific datasets (*e.g.*, faces [33], bedrooms[47]). However, capturing the distribution of highly diverse datasets with multiple object classes (*e.g.* ImageNet [12]), is still considered a major challenge and often requires conditioning the generation on another input signal [6] or training the model for a specific task (*e.g.* super-resolution [30], inpainting [41], retargeting [45]).

Here, we take the use of GANs into a new realm – *unconditional* generation learned from a *single natural image*. Specifically, we show that the internal statistics of patches within a single natural image typically carry enough information for learning a powerful generative model. *SinGAN*, our new single image generative model, allows us to deal with general natural images that contain complex structures and textures, without the need to rely on the existence of a database of images from the same class. This is achieved by a pyramid of fully convolutional light-weight GANs, each is responsible for capturing the distribution of patches at a different scale. Once trained, *SinGAN* can produce diverse high quality image samples (of arbitrary dimensions), which semantically resemble the training image, yet contain new object configurations and structures (Fig. 1).

Modeling the internal distribution of patches within a single natural image has been long recognized as a powerful prior in many computer vision tasks [64]. Classical examples include denoising [65], deblurring [39], super resolution [18], dehazing [2, 15], and image editing [37, 21, 9, 11, 50]. The most closley related work in this context is [48], where a bidirectional patch similarity measure is defined and optimized to guarantee that the patches of an image after manipulation are the same as the original ones. Motivated by these works, here we show

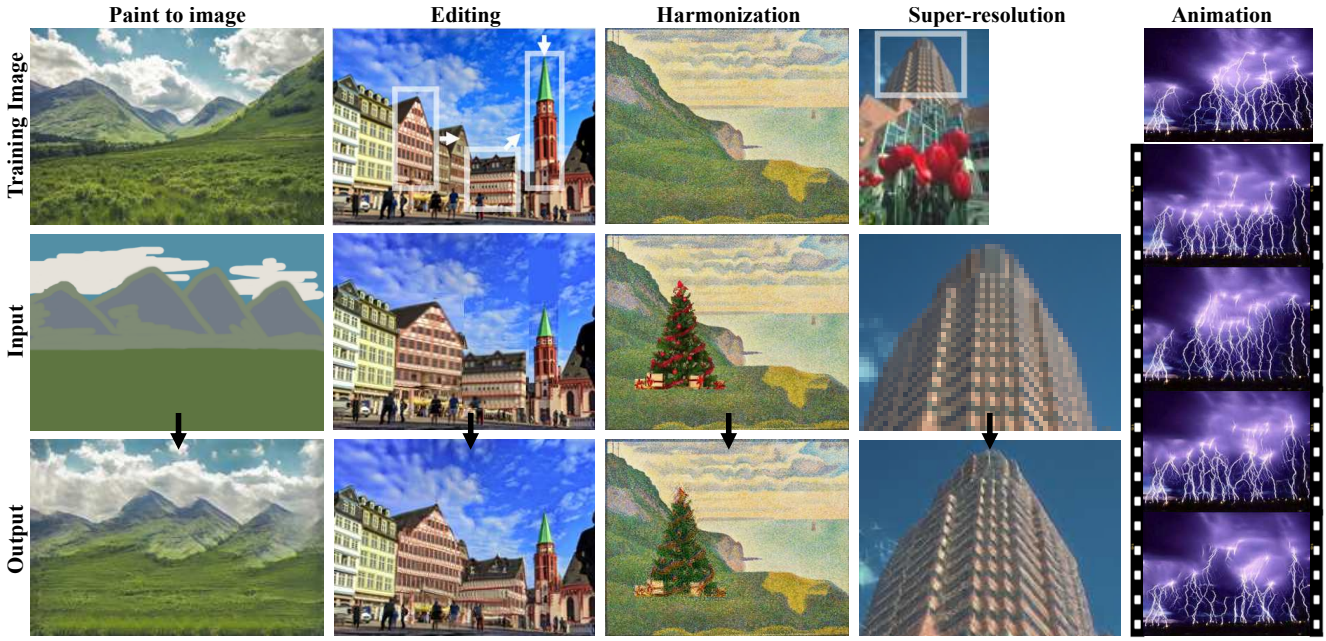


Figure 2: **Image manipulation.** SinGAN can be used in various image manipulation tasks, including: transforming a paint (clipart) into a realistic photo, rearranging and editing objects in the image, harmonizing a new object into an image, image super-resolution and creating an animation from a single input. In all these cases, our model observes only the training image (first row) and is trained in the same manner for all applications, with no architectural changes or further tuning (see Sec. 4).

how SinGAN can be used within a simple unified learning framework to solve a variety of image manipulation tasks, including paint-to-image, editing, harmonization, super-resolution, and animation from a single image. In all these cases, our model produces high quality results that preserve the internal patch statistics of the training image (see Fig. 2 and our project webpage). All tasks are achieved with *the same* generative network, without any additional information or further training beyond the original training image.

1.1. Related Work

Single image deep models Several recent works proposed to "overfit" a deep model to a single training example [51, 60, 46, 7, 1]. However, these methods are designed for specific tasks (*e.g.*, super resolution [46], texture expansion [60]). Shocher *et al.* [44, 45] were the first to introduce an internal GAN based model for a single natural image, and illustrated it in the context of retargeting. However, their generation is conditioned on an input image (*i.e.*, mapping images to images) and is not used to draw random samples. In contrast, our framework is purely generative (*i.e.* maps noise to image samples), and thus suits many different image manipulation tasks. *Unconditional* single image GANs have been explored only in the context of texture generation [3, 27, 31]. These models do not generate meaningful samples when trained on non-texture images (Fig. 3). Our method, on the other hand, is not restricted to texture and can handle general natural images (*e.g.*, Fig. 1).

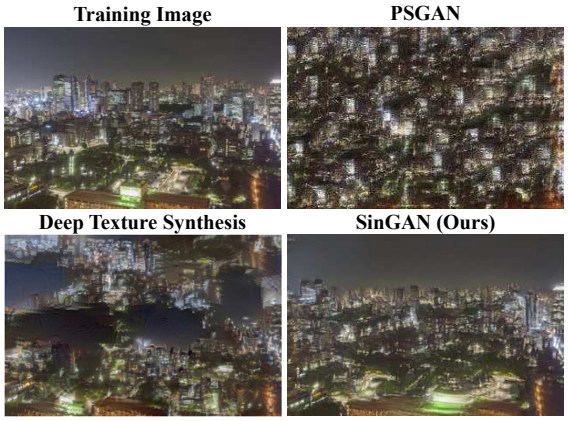


Figure 3: **SinGAN vs. Single Image Texture Generation.** Single image models for texture generation [3, 16] are not designed to deal with natural images. Our model can produce realistic image samples that consist of complex textures and non-reptititve global structures.

Generative models for image manipulation The power of adversarial learning has been demonstrated by recent GAN-based methods, in many different image manipulation tasks [61, 10, 62, 8, 53, 56, 42, 53]. Examples include interactive image editing [61, 10], sketch2image [8, 43], and other image-to-image translation tasks [62, 52, 54]. However, all these methods are trained on class specific datasets, and here too, often condition the generation on another input signal. We are not interested in capturing common features among images of the same class, but rather con-

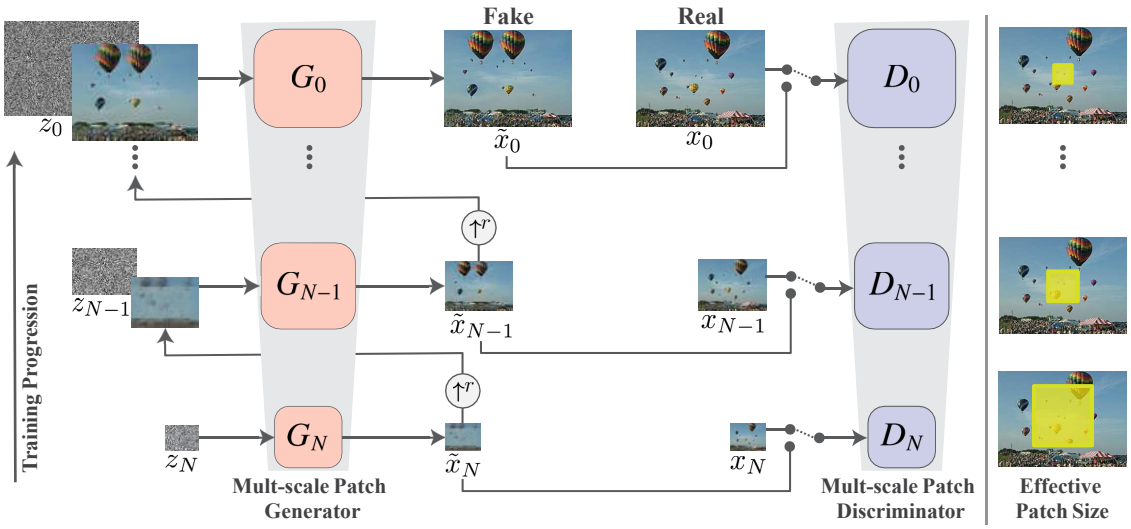


Figure 4: **SinGAN's multi-scale pipeline.** Our model consists of a pyramid of GANs, where both training and inference are done in a coarse-to-fine fashion. At each scale, G_n learns to generate image samples in which all the overlapping patches cannot be distinguished from the patches in the down-sampled training image, x_n, by the discriminator D_n; the effective patch size decreases as we go up the pyramid (marked in yellow on the original image for illustration). The input to G_n is a random noise image z_n, and the generated image from the previous scale $\tilde{x}_n$, upsampled to the current resolution (except for the coarsest level which is purely generative). The generation process at level n involves all generators $\{G_N \dots G_n\}$ and all noise maps $\{z_N, \dots, z_n\}$ up to this level. See more details at Sec. 2.

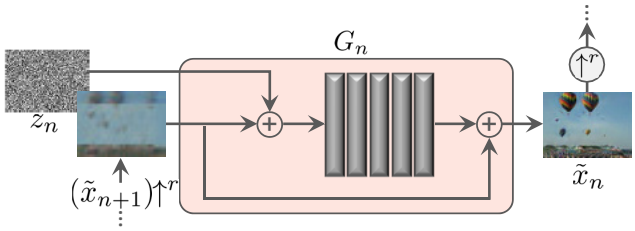


Figure 5: **Single scale generation.** At each scale n, the image from the previous scale, $\tilde{x}_{n+1}$, is upsampled and added to the input noise map, z_n. The result is fed into 5 conv layers, whose output is a residual image that is added back to $(\tilde{x}_{n+1}) \uparrow^r$. This is the output $\tilde{x}_n$ of G_n.

sider a different source of training data – all the overlapping patches at multiple scales of a single natural image. We show that a powerful generative model can be learned from this data, and can be used in a number of image manipulation tasks.

2. Method

Our goal is to learn an *unconditional* generative model that captures the internal statistics of a *single* training image x. This task is conceptually similar to the conventional GAN setting, except that here the training samples are patches of a single image, rather than whole image samples from a database.

We opt to go beyond texture generation, and to deal with more general natural images. This requires capturing the statistics of complex image structures at many different scales. For example, we want to capture global properties such as the arrangement and shape of large objects in the image (*e.g.* sky at the top, ground at the bottom), as well as fine details and texture information. To achieve that, our generative framework, illustrated in Fig. 4, consists of a hierarchy of patch-GANs (Markovian discriminator) [31, 26], where each is responsible for capturing the patch distribution at a different scale of x. The GANs have small receptive fields and limited capacity, preventing them from memorizing the single image. While similar multi-scale architectures have been explored in conventional GAN settings (*e.g.* [28, 52, 29, 52, 13, 24]), we are the first explore it for internal learning from a single image.

2.1. Multi-scale architecture

Our model consists of a pyramid of generators, $\{G_0, \dots, G_N\}$, trained against an image pyramid of x: $\{x_0, \dots, x_N\}$, where x_n is a downsampled version of x by a factor r^n, for some $r > 1$. Each generator G_n is responsible of producing realistic image samples w.r.t. the patch distribution in the corresponding image x_n. This is achieved through adversarial training, where G_n learns to fool an associated discriminator D_n, which attempts to distinguish patches in the generated samples from patches in x_n.

The generation of an image sample starts at the coarsest scale and sequentially passes through all generators up to the finest scale, with noise injected at every scale. All the generators and discriminators have the same receptive field and thus capture structures of decreasing size as we go up the generation process. At the coarsest scale, the generation is purely generative, *i.e.* G_N maps spatial white Gaussian noise z_N to an image sample $\tilde{x}_N$,

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

The effective receptive field at this level is typically $\sim 1/2$ of the image's height, hence G_N generates the general layout of the image and the objects' global structure. Each of the generators G_n at finer scales ($n < N$) adds details that were not generated by the previous scales. Thus, in addition to spatial noise z_n, each generator G_n accepts an upsampled version of the image from the coarser scale, *i.e.*,

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

All the generators have a similar architecture, as depicted in Fig. 5. Specifically, the noise z_n is added to the image $(\tilde{x}_{n+1})\uparrow^r$, prior to being fed into a sequence of convolutional layers. This ensures that the GAN does not disregard the noise, as often happens in conditional schemes involving randomness [62, 36, 63]. The role of the convonlutional layers is to generate the missing details in $(\tilde{x}_{n+1})\uparrow^r$ (residual learning [22, 57]). Namely, G_n performs the operation

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

where ψ_n is a fully convolutional net with 5 conv-blocks of the form Conv(3×3)-BatchNorm-LeakyReLU [25]. We start with 32 kernels per block at the coarsest scale and increase this number by a factor of 2 every 4 scales. Because the generators are fully convolutional, we can generate images of arbitrary size and aspect ratio at test time (by changing the dimensions of the noise maps).

2.2. Training

We train our multi-scale architecture sequentially, from the coarsest scale to the finest one. Once each GAN is trained, it is kept fixed. Our training loss for the nth GAN is comprised of an adversarial term and a reconstruction term,

$$\min_{G_n}\max_{D_n} \mathcal{L}_{\text{adv}}(G_n, D_n) + \alpha\mathcal{L}_{\text{rec}}(G_n). \tag{4}$$

The adversarial loss $\mathcal{L}_{\text{adv}}$ penalizes for the distance between the distribution of patches in x_n and the distribution of patches in generated samples $\tilde{x}_n$. The reconstruction loss $\mathcal{L}_{\text{rec}}$ insures the existence of a specific set of noise maps that can produce x_n, an important feature for image manipulation (Sec. 4). We next describe $\mathcal{L}_{\text{adv}}, \mathcal{L}_{\text{rec}}$ in detail. See Supplementary Materials (SM) for optimization details.

Adversarial loss Each of the generators G_n is coupled with a Markovian discriminator D_n that classifies each of the overlapping patches of its input as real or fake [31, 26]. We use the WGAN-GP loss [20], which we found to increase training stability, where the final discrimination score is the average over the patch discrimination map. As opposed to single-image GANs for textures (*e.g.*, [31, 27, 3]), here we define the loss over the whole image rather than over random crops (a batch of size 1). This allows the net to learn boundary conditions (see SM), which is an important feature in our setting. The architecture of D_n is the same as the net ψ_n within G_n, so that its patch size (the net's receptive field) is 11×11.

Reconstruction loss We want to ensure that there exists a specific set of input noise maps, which generates the original image x. We specifically choose $\{z_N^{\text{rec}}, z_{N-1}^{\text{rec}} \ldots, z_0^{\text{rec}}\} = \{z^*, 0, \ldots, 0\}$, where z^* is some fixed noise map (drawn once and kept fixed during training). Denote by $\tilde{x}_n^{\text{rec}}$ the generated image at the nth scale when using these noise maps. Then for $n < N$,

$$\mathcal{L}_{\text{rec}} = \|G_n(0, (\tilde{x}_{n+1}^{\text{rec}})\uparrow^r) - x_n\|^2, \tag{5}$$

and for $n = N$, we use $\mathcal{L}_{\text{rec}} = \|G_N(z^*) - x_N\|^2$.

The reconstructed image $\tilde{x}_n^{\text{rec}}$ has another role during training, which is to determine the standard deviation σ_n of the noise z_n in each scale. Specifically, we take σ_n to be proportional to the root mean squared error (RMSE) between $(\tilde{x}_{n+1}^{\text{rec}})\uparrow^r$ and x_n, which gives an indication of the amount of details that need to be added at that scale.

3. Results

We tested our method both qualitatively and quantitatively on a variety of images spanning a large range of scenes including urban and nature scenery as well as artistic and texture images. The images that we used are taken from the Berkeley Segmentation Database (BSD) [35], Places [59] and the Web. We always set the minimal dimension at the coarsest scale to 25px, and choose the number of scales N s.t. the scaling factor r is as close as possible to $4/3$. For all the results, (unless mentioned otherwise), we resized the training image to maximal dimension 250px.

Qualitative examples of our generated random image samples are shown in Fig. 1, Fig. 6, and many more examples are included in the SM. For each example, we show a number of random samples with the same aspect ratio as the original image, and with decreased and expanded dimensions in each axis. As can be seen, in all these cases, the generated samples depict new realistic structures and configuration of objects, while preserving the visual content of the training image. Our model successfully preservers global structure of objects, *e.g.* mountains (Fig. 1), air balloons or pyramids (Fig. 6), as well as fine texture information. Because the network has a limited receptive field (smaller than the entire image), it can generate new combinations of patches that do not exist in the training image Furthermore, we observe that in many cases reflections and shadows are realistically synthesized, as can be seen in Fig. 6 and Fig. 1 (and the first example of Fig. 8). Note that SinGAN's architecture is resolution agnostic and can thus be used on high resolution images, as illustrated in Fig. 7 (see 4Mpix results in the SM). Here as well, structures at all scales are nicely generated, from the global arrangement of sky, clouds and mountains, to the fine textures of the snow.

Effect of scales at test time Our multi-scale architecture allows control over the amount of variability between samples, by choosing the scale from which to start the generation at test time. To start at scale n, we fix the noise maps up

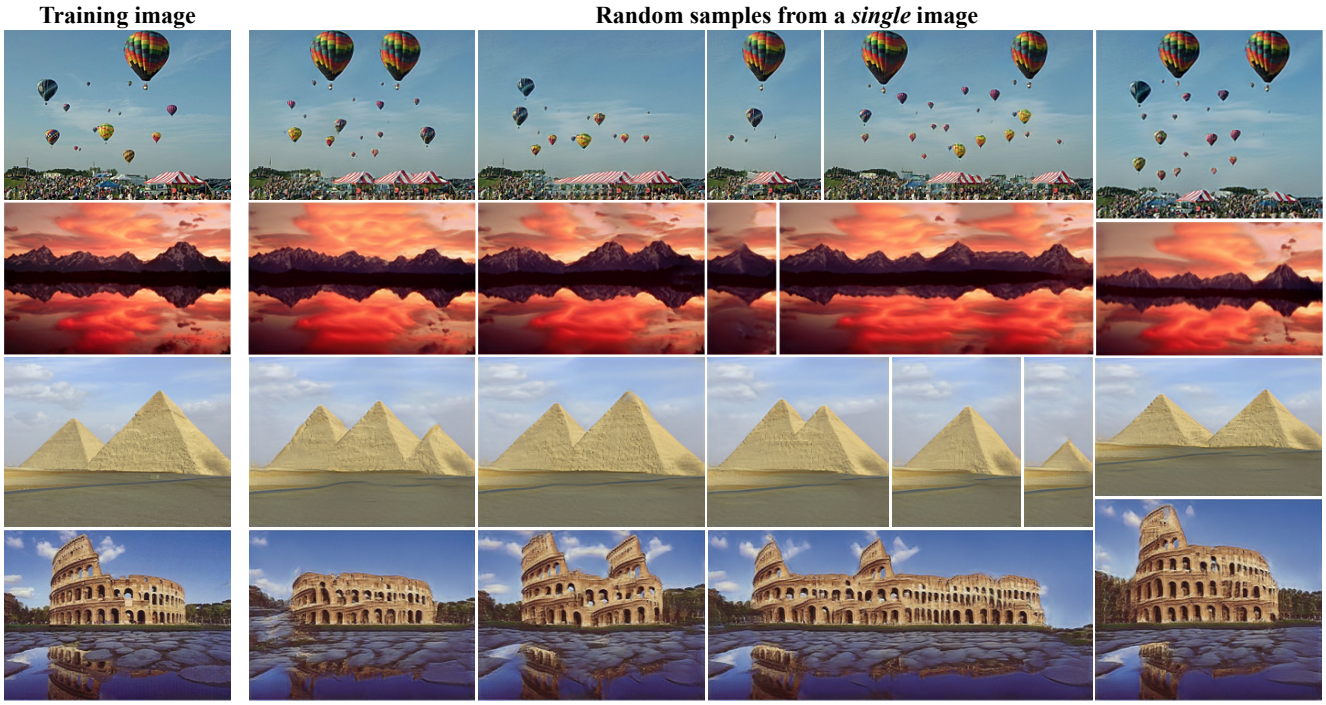


Figure 6: **Random image samples.** After training SinGAN on a single image, our model can generate realistic random image samples that depict new structures and object configurations, yet preserve the patch distribution of the training image. Because our model is fully convolutional, the generated images may have arbitrary sizes and aspect ratios. Note that our goal is not image retargeting – our image samples are random and optimized to maintain the patch statistics, rather than preserving salient objects. See SM for more results and qualitative comparison to image retargeting methods.

Figure 7: **High resolution image generation.** A random sample produced by our model, trained on the 243×1024 image (upper right corner); new global structures as well as fine details are realistically generated. See 4Mpix examples in SM.

to this scale to be $\{z_N^{\text{rec}}, \ldots, z_{n+1}^{\text{rec}}\}$, and use random draws only for $\{z_n, \ldots, z_0\}$. The effect is illustrated in Fig. 8. As can be seen, starting the generation at the coarsest scale ($n = N$), results in large variability in the global structure. In certain cases with a large salient object, like the Zebra image, this may lead to unrealistic samples. However, starting the generation from finer scales, enables to keep the global structure intact, while altering only finer image features (*e.g.* the Zebra's stripes). See SM for more examples.

Effect of scales during training Figure 9 shows the effect of training with fewer scales. With a small number of scales, the effective receptive field at the coarsest level is smaller, allowing to capture only fine textures. As the number of scales increases, structures of larger support emerge, and the global object arrangement is better preserved.

3.1. Quantitative Evaluation

To quantify the realism of our generated images and how well they capture the internal statistics of the training image, we use two metrics: (i) Amazon Mechanical Turk (AMT) "Real/Fake" user study, and (ii) a new single-image version of the Fréchet Inception Distance [23].

AMT perceptual study We followed the protocol of [26, 58] and performed perceptual experiments in 2 settings.

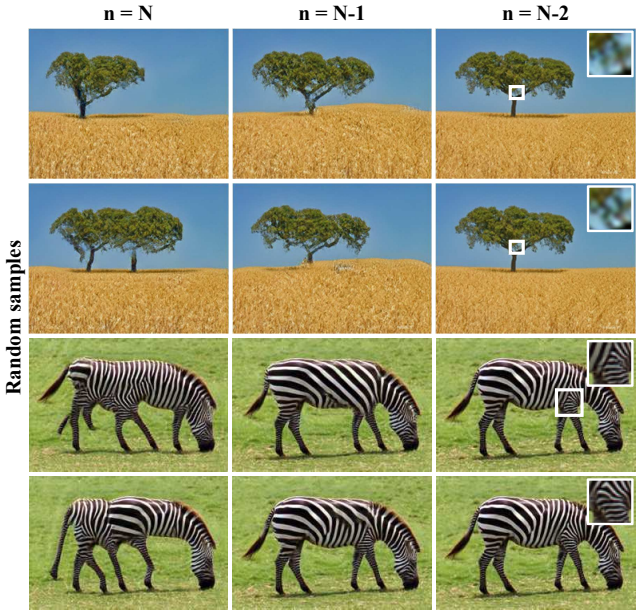


Figure 8: **Generation from different scales (at inference).** We show the effect of starting our hierarchical generation from a given level n. For our full generation scheme ($n = N$), the input at the coarsest level is random noise. For generation from a finer scale n, we plug in the downsampled original image, x_n, as input to that scale. This allows us to control the scale of the generated structures, *e.g.*, we can preserve the shape and pose of the Zebra and only change its stripe texture by starting the generation from $n = N-1$.

(i) Paired (real vs. fake): Workers were presented with a sequence of 50 trials, in each of which a fake image (generated by SinGAN) was presented against its real training image for 1 second. Workers were asked to pick the fake image. (ii) Unpaired (either real or fake): Workers were presented with a *single* image for 1 second, and were asked if it was fake. In total, 50 real images and a disjoint set of 50 fake images were presented in random order to each worker.

We repeated these two protocols for two types of generation processes: Starting the generation from the coarsest (Nth) scale, and from scale $N-1$ (as in Fig. 8). This way, we assess the realism of our results in two different variability levels. To quantify the diversity of the generated images, for each training example we calculated the standard deviation (std) of the intensity values of each pixel over 100 generated images, averaged it over all pixels, and normalized by the std of the intensity values of the training image.

The real images were randomly picked from the "places" database [59] from the subcategories Mountains, Hills, Desert, Sky. In each of the 4 tests, we had 50 different participants. In all tests, the first 10 trials were a tutorial including a feedback. The results are reported in Table 1.

As expected, the confusion rates are consistently larger in the unpaired case, where there is no reference for comparison. In addition, it is clear that the confusion rate decreases with the diversity of the generated images. However, even when large structures are changed, our generated images were hard to distinguish from the real images (a score of 50% would mean perfect confusion between real and fake). The full set of test images are included in the SM.

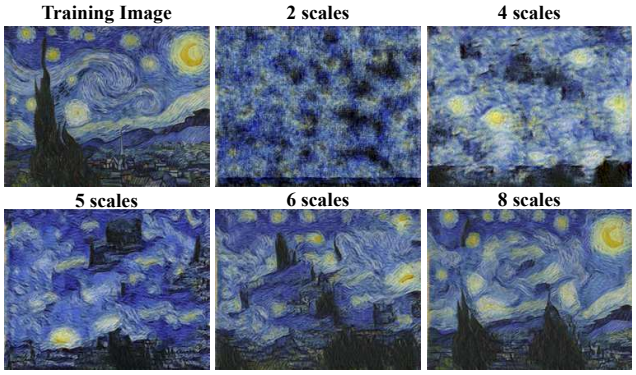


Figure 9: **The effect of training with a different number of scales.** The number of scales in SinGAN's architecture strongly influences the results. A model with a small number of scales only captures textures. As the number of scales increases, SinGAN manages to capture larger structures as well as the global arrangement of objects in the scene.

1st Scale	Diversity	Survey	Confusion
N	0.5	paired	$21.45\% \pm 1.5\%$
		unpaired	$42.9\% \pm 0.9\%$
$N-1$	0.35	paired	$30.45\% \pm 1.5\%$
		unpaired	$47.04\% \pm 0.8\%$

Table 1: **"Real/Fake" AMT test.** We report confusion rates for two generation processes: Starting from the coarsest scale N (producing samples with large diversity), and starting from the second coarsest scale $N-1$ (preserving the global structure of the original image). In each case, we performed both a paired study (real-vs.-fake image pairs are shown), and an unpaired one (either fake or real image is shown). The variance was estimated by bootstrap [14].

Single Image Fréchet Inception Distance We next quantify how well SinGAN captures the internal statistics of x. A common metric for GAN evaluation is the Fréchet Inception Distance (FID) [23], which measures the deviation between the distribution of deep features of generated images and that of real images. In our setting, however, we only have a single real image, and are rather interested in its *internal* patch statistics. We thus propose the Single Image FID (SIFID) metric. Instead of using the activation vector after the last pooling layer in the Inception Network [49] (a single vector per image), we use the internal distribution of deep features at the output of the convolutional layer just before the second pooling layer (one vector per location in the map). Our SIFID is the FID between the statistics of those features in the real image and in the generated sample.

As can be seen in Table 2, the average SIFID is lower for

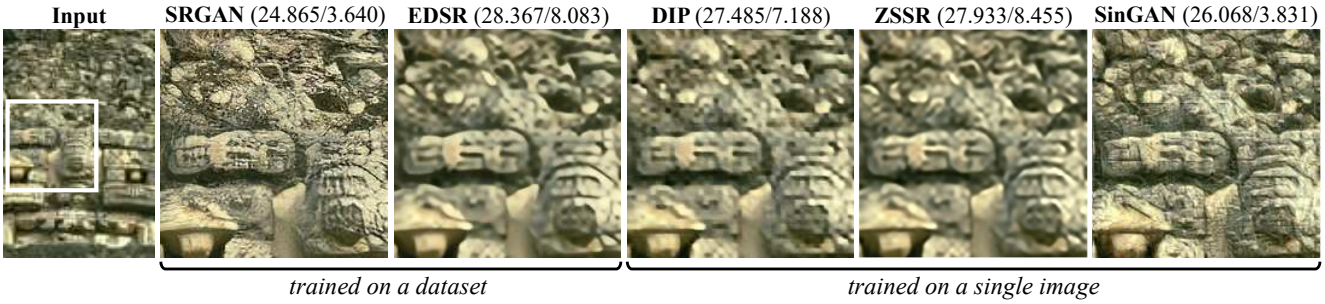


Figure 10: **Super-Resolution.** When SinGAN is trained on a low resolution image, we are able to super resolve. This is done by iteratively upsampling the image and feeding it to SinGAN's finest scale generator. As can be seen, SinGAN's visual quality is better than the SOTA internal methods ZSSR [46] and DIP [51]. It is also better than EDSR [32] and comparable to SRGAN [30], external methods trained on large collections. Corresponding PSNR and NIQE [40] are shown in parentheses.

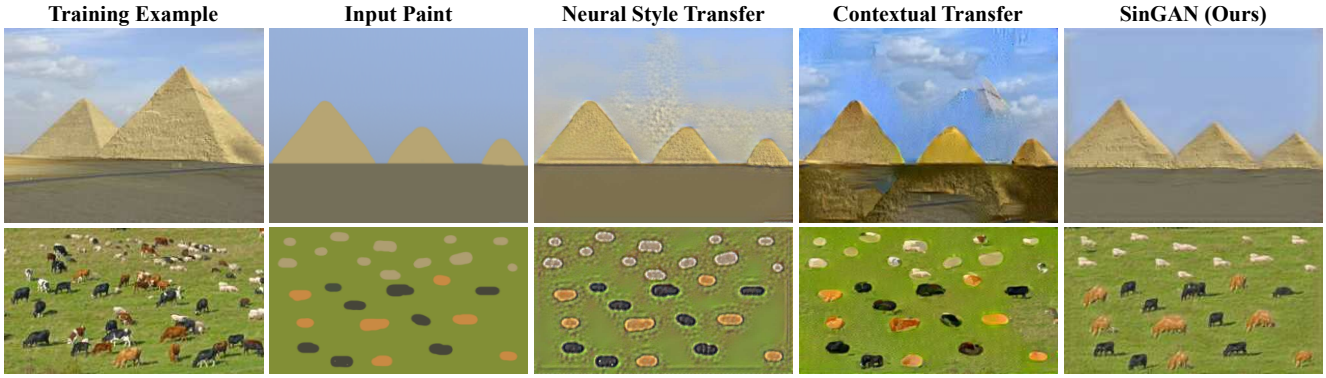


Figure 11: **Paint-to-Image.** We train SinGAN on a target image and inject a downsampled version of the paint into one of the coarse levels at test time. Our generated images preserve the layout and general structure of the clipart while generating realistic texture and fine details that match the training image. Well-known style transfer methods [17, 38] fail in this task.

1st Scale	SIFID	Survey	SIFID/AMT Correlation
N	0.09	paired	-0.55
		unpaired	-0.22
$N-1$	0.05	paired	-0.56
		unpaired	-0.34

Table 2: **Single Image FID (SIFID).** We adapt the FID metric to a single image and report the average score for 50 images, for full generation (first row), and starting from the second coarsest scale (second row). Correlation with AMT results shows SIFID highly agrees with human ranking.

generation from scale $N-1$ than for generation from scale N, which aligns with the user study results. We also report the correlation between the SIFID scores and the confusion rates for the fake images. Note that there is a significant (anti) correlation between the two, implying that a small SIFID is typically a good indicator for a large confusion rate. The correlation is stronger for the paired tests, since SIFID is a paired measure (it operates on the pair $x_n, \tilde{x}_n$).

4. Applications

We explore the use of SinGAN for a number of image manipulation tasks. To do so, we use our model *after training*, with no architectural changes or further tuning and follow the same approach for all applications. The idea is to utilize the fact that at inference, SinGAN can only produce images with the same patch distribution as the training image. Thus, manipulation can be done by injecting (a possibly downsampled version of) an image into the generation pyramid at some scale $n < N$, and feed forwarding it through the generators so as to match its patch distribution to that of the training image. Different injection scales lead to different effects. We consider the following applications (see SM for more results and the injection scale effect).

Super-Resolution *Increase the resolution of an input image by a factor s.* We train our model on the low-resolution (LR) image, with a reconstruction loss weight of $\alpha = 100$ and a pyramid scale factor of $r = \sqrt[k]{s}$ for some $k \in \mathbb{N}$. Since small structures tend to recur across scales of natural scenes [18], at test time we upsample the LR image by a factor of r and inject it (together with noise) to the last generator, G_0. We repeat this k times to obtain the final high-res output. An example result is shown in Fig. 10. As can be seen, the visual quality of our reconstruction exceeds that of state-of-the-art *internal* methods [51, 46] as well as of *external* methods that aim for PSNR maximization [32].

	External methods		Internal methods		
	SRGAN	EDSR	DIP	ZSSR	SinGAN
RMSE	16.34	12.29	13.82	13.08	16.22
NIQE	3.41	6.50	6.35	7.13	3.71

Table 3: **Super-Resolution evaluation**. Following [5], we report distortion (RMSE) and perceptual quality (NIQE [40], lower is better) on BSD100 [35]. As can be seen, SinGAN's performance is similar to that of SRGAN [30].

Interestingly, it is comparable to the externally trained SRGAN method [30], despite having been exposed to only a single image. Following [4], we compare these 5 methods in Table 3 on the BSD100 dataset [35] in terms of distortion (RMSE) and perceptual quality (NIQE [40]), which are two fundamentally conflicting requirements [5]. As can be seen, SinGAN excels in perceptual quality; its NIQE score is only slightly inferior to SRGAN, and its RMSE is slightly better.

Paint-to-Image *Transfer a clipart into a photo-realistic image.* This is done by downsampling the clipart image and feeding it into one of the coarse scales (*e.g.* $N-1$ or $N-2$). As can be seen in Figs. 2 and 11, the global structure of the painting is preserved, while texture and high frequency information matching the original image are realistically generated. Our method outperforms style transfer methods [38, 17] in terms of visual quality (Fig. 11).

Harmonization *Realistically blend a pasted object with a background image.* We train SinGAN on the background image, and inject a downsampled version of the naively pasted composite at test time. Here we combine the generated image with the original background. As can be seen in Fig. 2 and Fig. 13, our model tailors the pasted object's texture to match the background, and often preserves its structure better than [34]. Scales 2,3,4 typically lead to good balance between preserving the object's structure and transferring the background's texture.

Editing *Produce a seamless composite in which image regions have been copied and pasted in other locations.* Here, again, we inject a downsampled version of the composite into one of the coarse scales. We then combine SinGAN's output at the edited regions, with the original image. As shown in Fig. 2 and Fig. 12, SinGAN re-generates fine textures and seamlessly stitches the pasted parts, producing nicer results than Photoshop's Content-Aware-Move.

Single Image Animation *Create a short video clip with realistic object motion, from a single input image.* Natural images often contain repetitions, which reveal different "snapshots" in time of the same dynamic object [55] (*e.g.* an image of a flock of birds reveals all wing postures of a single bird). Using SinGAN, we can travel along the manifold of all appearances of the object in the image, thus synthesizing motion from a single image. We found that for many types of images, a realistic effect is achieved by a random walk in z-space, starting with z^{rec} for the first frame at all generation scales (see SM video).

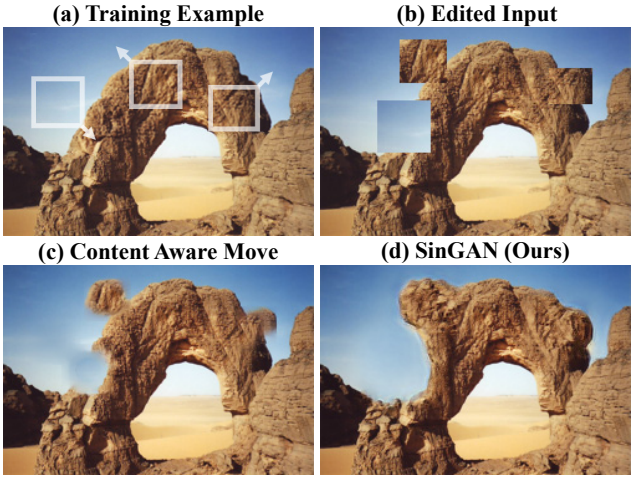


Figure 12: **Editing.** We copy and paste a few patches from the original image (a), and input a downsampled version of the edited image (b) to an intermediate level of our model (pretrained on (a)). In the generated image (d), these local edits are translated into coherent and photo-realistic structures. (c) comparison to Photoshop content aware move.

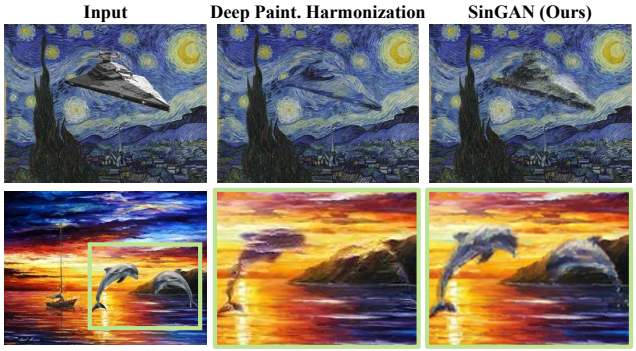


Figure 13: **Harmonization.** Our model is able to preserve the structure of the pasted object, while adjusting its appearance and texture. The dedicated harmonization method [34] overly blends the object with the background.

5. Conclusion

We introduced SinGAN, a new unconditional generative scheme that is learned from a single natural image. We demonstrated its ability to go beyond textures and to generate diverse realistic samples for natural complex images. Internal learning is inherently limited in terms of *semantic* diversity compared to externally trained generation methods. For example, if the training image contains a single dog, our model will not generate samples of different dog breeds. Nevertheless, as demonstrated by our experiments, SinGAN can provide a very powerful tool for a wide range of image manipulation tasks.

Acknowledgements Thanks to Idan Kligvasser for valuable insights. This research was supported by the Israel Science Foundation (grant 852/17) and the Ollendorff foundation.

References

[1] Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi. Surprising effectiveness of few-image unsupervised feature learning. *arXiv preprint arXiv:1904.13132*, 2019. 2

[2] Yuval Bahat and Michal Irani. Blind dehazing using internal patch recurrence. In *2016 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2016. 1

[3] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial GAN. *arXiv preprint arXiv:1705.06566*, 2017. 2, 4

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *European Conference on Computer Vision Workshops*, pages 334–355. Springer, 2018. 8

[5] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 8

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 2

[8] Wengling Chen and James Hays. Sketchygan: towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018. 2

[9] Taeg Sang Cho, Moshe Butman, Shai Avidan, and William T Freeman. The patch transform and its applications to image editing. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[10] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3511–3520, 2018. 2

[11] Tali Dekel, Tomer Michaeli, Michal Irani, and William T Freeman. Revealing and modifying non-local variations in a single image. *ACM Transactions on Graphics (TOG)*, 34(6):227, 2015. 1

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[13] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 3

[14] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992. 6

[15] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):12, 2011. 1

[16] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015. 2

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 7, 8

[18] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 349–356. IEEE, 2009. 1, 7

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 4

[21] Kaiming He and Jian Sun. Statistics of patch offsets for image completion. In *European Conference on Computer Vision*, pages 16–29. Springer, 2012. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5, 6

[24] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5077–5086, 2017. 3

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 3, 4, 5

[27] Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. Texture synthesis with spatial generative adversarial networks. *Workshop on Adversarial Training, NIPS*, 2016. 2, 4

[28] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 3

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 7, 8

[31] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 2, 3, 4

[32] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 7

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 1

[34] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep painterly harmonization. *arXiv preprint arXiv:1804.03189*, 2018. 8

[35] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *null*, page 416. IEEE, 2001. 4, 8

[36] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 4

[37] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1368–1376. IEEE, 2018. 1

[38] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 7, 8

[39] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *European Conference on Computer Vision*, pages 783–798. Springer, 2014. 1

[40] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 7, 8

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1

[42] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional GANs for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2

[43] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017. 2

[44] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and remapping the "DNA" of a natural image. *arXiv preprint arXiv: arXiv:1812.00231*, 2018. 2

[45] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. InGAN: Capturing and Remapping the "DNA" of a Natural Image. *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[46] Assaf Shocher, Nadav Cohen, and Michal Irani. Zero-Shot Super-Resolution using Deep Internal Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 2, 7

[47] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 1

[48] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 1

[49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 6

[50] Tal Tlusty, Tomer Michaeli, Tali Dekel, and Lihi Zelnik-Manor. Modifying non-local variations across multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6276–6285, 2018. 1

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. *arXiv preprint arXiv:1711.11585*, 2017. 2, 3

[53] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. 2016. 2

[54] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[55] Xuemiao Xu, Liang Wan, Xiaopei Liu, Tien-Tsin Wong, Liansheng Wang, and Chi-Sing Leung. Animating animal motion from still. *ACM Transactions on Graphics (TOG)*, 27(5):117, 2008. 8

[56] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 2

[57] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 4

[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 5

[59] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 4, 6

[60] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. *arXiv preprint arXiv:1805.04487*, 2018. 2

[61] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision (ECCV)*, pages 597–613. Springer, 2016. 2

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 2, 4

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017. 4

[64] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984. IEEE, 2011. 1

[65] Maria Zontak, Inbar Mosseri, and Michal Irani. Separating signal from noise using patch recurrence across scales. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1195–1202, 2013. 1

西安电子科技大学
毕业设计（论文）指导情况登记表

专业	智能科学与专业	题目	基于单幅图像的面部表情生成算法研究		
学生姓名	邢博伟	学号	16020510038	教师姓名	毛莎莎
指导次数	指 导 内 容	指 导 效 果		指导时间	指导教师签名
1	初步了解毕业设计课题	讲解非常详细		2019.12.6	毛莎莎
2	指导毕业设计课题研究目的和研究意义	探讨很深入		2019.12.13	毛莎莎
3	研究毕设课题的国内外研究现状和发展趋势	讲解非常耐心		2019.12.20	毛莎莎
4	了解 SinGAN，毛老师为我讲解相关文献	解释通俗易懂		2019.12.27	毛莎莎
5	帮助解决复现 SinGAN 代码过程出现的问题	非常和蔼可亲		2020.3.6	毛莎莎
6	指导搭建深度学习环境	指导特别细心		2020.3.13	毛莎莎
7	讨论解决 SinGAN 对人脸失真的问题	帮助很大		2020.3.20	毛莎莎

指 导 次 数	指 导 内 容	指 导 效 果	指导 时间	指导教师 签 名
8	继续讨论针对 SinGAN 对人脸图像失真的解决办法	得出不错的效果	2020.3.27	毛莎莎
9	讨论选择哪个网络模型与 SinGAN 相结合	得出结果选择 StarGAN	2020.4.3	毛莎莎
10	帮助解决远程服务器的问题	老师超级耐心，人非常好	2020.4.10	毛莎莎
11	创新搭建合成模型 SinGANimation	老师的指导有很大帮助	2020.4.17	毛莎莎
12	与导师讨论论文结构框架	老师介绍很详细	2020.4.24	毛莎莎
13	与导师讨论对算法进一步优化改进	老师的指导非常有用	2020.5.3	毛莎莎
14	毛老师为我修改论文	老师提出了多处修改意见，非常细心	2020.5.10	毛莎莎
15	根据盲审意见，与导师讨论再次修改论文	老师很耐心指出论文的不足	2020.5.17	毛莎莎
16	与导师讨论毕业答辩事项	老师提出很多答辩建议	2020.5.27	毛莎莎

说明：1.本表由学生填写；

2.学生提出的问题或导师指导的问题均填入指导内容栏内，经教师指导将解决结果记在效果栏内，并请指导教师签名作为记载。