

硕 士 研 究 生 读 书 报 告



题目 基于生成模型的实时交互式图像编辑

作者姓名 张 涛

作者学号 22351266

指导教师 李启雷

学科专业 电子信息

所在学院 软件学院

提交日期 二○二四年一月

# 基于生成模型的实时交互式图像编辑

## 引言

近些年来，深度学习相关的人工智能技术得到了飞速的发展，基于判别式深度学习模型的相关工作和工程层出不穷，目标检测，语言分割等相关应用也为人们的生活带来了极大的方便。这时，人们开始思考，既然深度学习能够在判别式任务中获得如此大的成果，那么能不能让深度学习模型做一些生成式的任务，辅助人们完成一些有创意性的工作呢？生成模型就应运而生了。

生成模型的能力正如其名，它能够借助深度学习模型强大的挖掘数据信息的能力，学习到数据的隐式分布，并且能够凭借学习到的隐式分布从而生成出与输入数据分布类似的输出，从而达到以假乱真的生成效果。生成模型近些年也得到了飞速的发展，带动了一场AIGC的浪潮，在计算机视觉领域，生成模型被广泛应用于各类图像生成，视频生成，甚至3D物体生成等；在自然语言处理领域，生成模型也有被用于目前正火热的GPT模型中，可以说在各个领域中总能看到生成模型的影子。

由于生成模型具备优秀的以假乱真的能力，生成模型也常被用于图像编辑任务中。图像编辑在日常的生活，以及设计类的工作中是非常常见的任务。因此，生成模型在图像编辑中的发展具有重要的实际意义。在SIGGRAH2023会议上的一篇论文DragGAN[1]引发了业内和业外众多人员的关注，它可以基于生成模型实现实时交互式的精准图像编辑，将生成模型完成图像编辑任务的水准又推向一个新的高度。本次报告主要讲解DragGAN论文的工作，同时也回顾一些以往基于生成模型实现图像编辑优秀的相关工作，保持前沿性的同时，也不失全面性。

## 相关工作

在DragGAN被提出以前，已经有相当多关于基于生成模型做图像编辑任务的工作发表，以下是一些典型的例子。

### 基于监督学习的方法

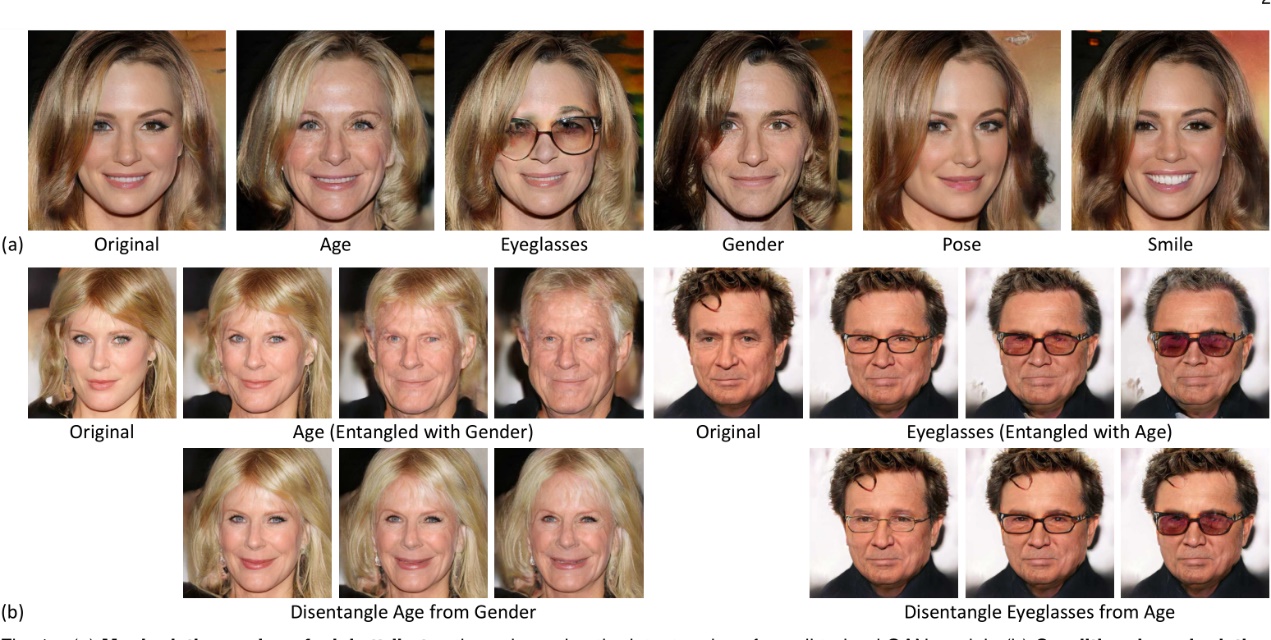
基于监督学习完成图像编辑的方法，需要用户给数据划分一系列标签，通过训练让生成模型学习到这些标签的属性，随后给训练好的模型输入图片，并给定想要获取的标签，就能够使得图像向指定方向的属性进行编辑。一个比较典型的例子就是InterfaceGAN[2]，这个模型能够实现输入一张人像，并给定一个标签，从而实现使得原始的人像获得该标签的属性。

图 2.1 InterfaceGAN 示例图

### 基于语义图的方法

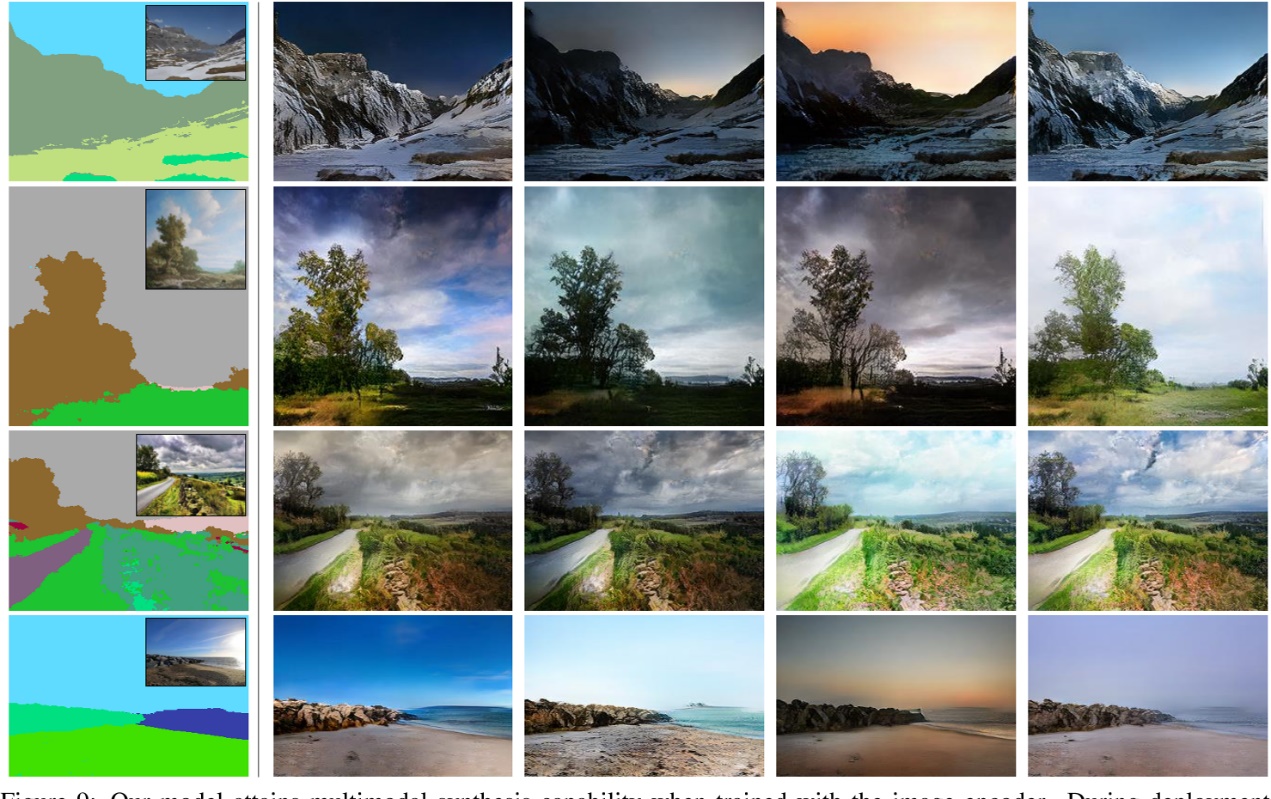
基于语义图的方法，从视觉上比较形象。用户可以通过输入一张语义图的方式，拟定图片的一个大体框架，随后由生成模型来完成细节部分的工作。一个典型的例子就是SPADE[3]，用户可以通过自由绘制语义图，模型根据用户给定的语义图扩展成精美的风景图，非常直观地完成图片的编辑。如下图所示，图中最左侧就是输入给模型的语义图，右侧则是模型生成的图片。

图 2.2 SPADE示例图

### 基于文本驱动的方法

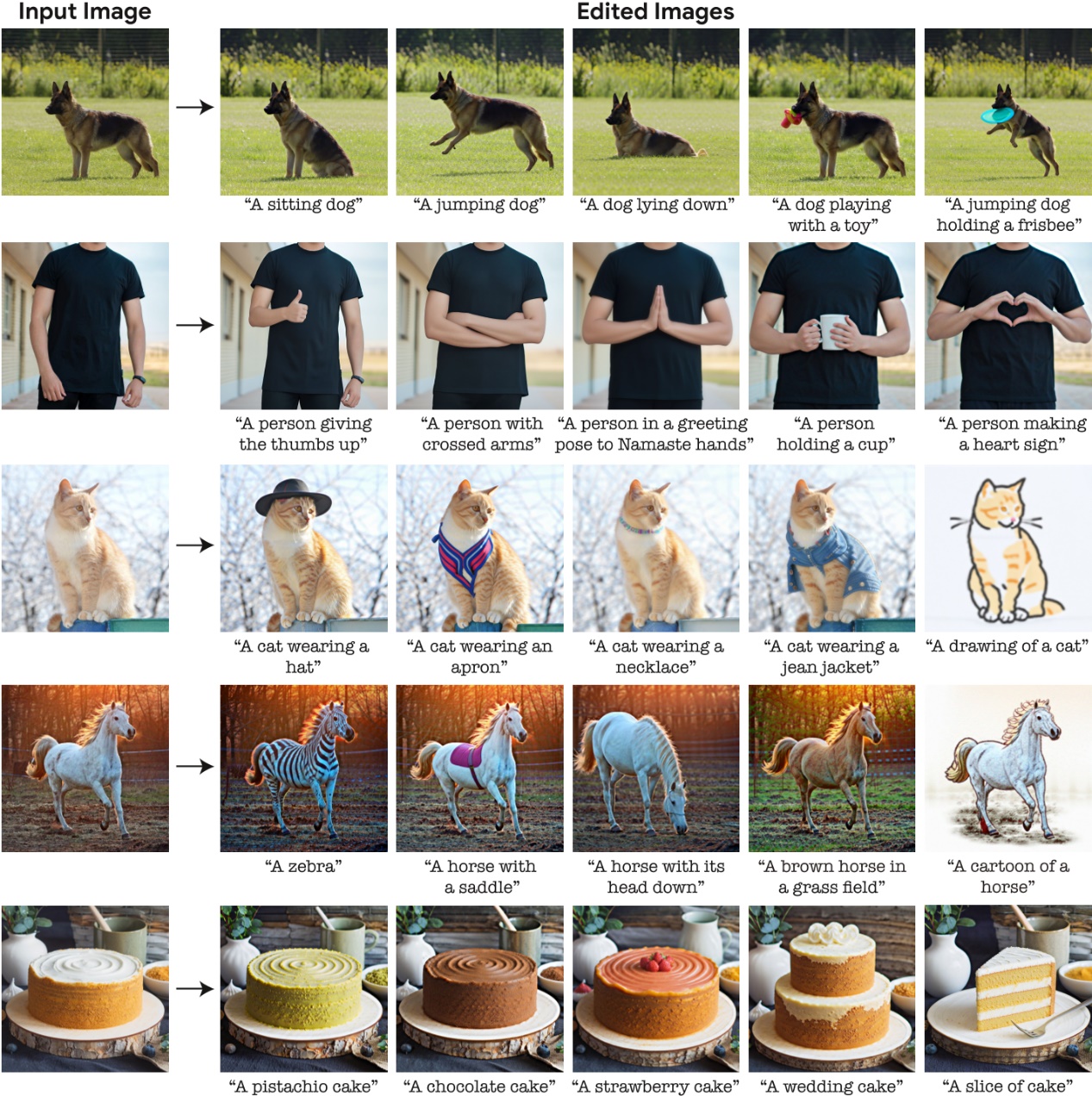
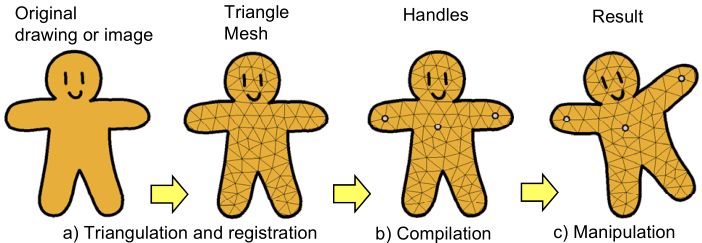
近些年，得益于大模型和多模态方法的飞速发展，以及扩散模型的提出，基于文本驱动的方法成为了一种非常主流的图像编辑方法。在这种方法中，用户可以通过输入文本要求这样极具语义的方法，就能够轻松地实现图像的编辑。目前非常火爆的Midjourney，StableDiffusion，DALL等都是基于这种方式。下图是近期的一篇基于文本驱动的方式完成图像编辑工作Imagic[4]的示例图，可以看到这种方式可以简单地输入一句话就能够实现图像的编辑。

图 2.3 Imagic示例图

### 基于拖拽式的方法

基于拖拽式的方法，其实早在2005年一篇图形学相关的论文的论文As-rigid-as-possible[5]中就已经提出来了。在这篇论文中，采用的方法是将物体网格化并且假设物体具有一致的刚度，从而来实现形变，下图是其采取的方法的示意图。可以看到，这种方法由于假设了物体具有一致的刚度，导致了其物体不符合本身的结构，因为很多时候物体是具有骨架的。同时，它也不能生成一些新的内容，即原本被遮罩住的部分，无法通过这样的拉拽来清除掉遮罩。相对而言，新提出的基于拖拽式的DragGAN解决了这些问题，DragGAN的拖拽是更加符合语义，更加贴近显示的拖拽式的图片编辑，它的图像编辑能够符合物体的结构，并且能够在拖拽时想象出被遮挡住的内容。

图 2.4 As-rigid-as-possible方法示意图

## DragGAN

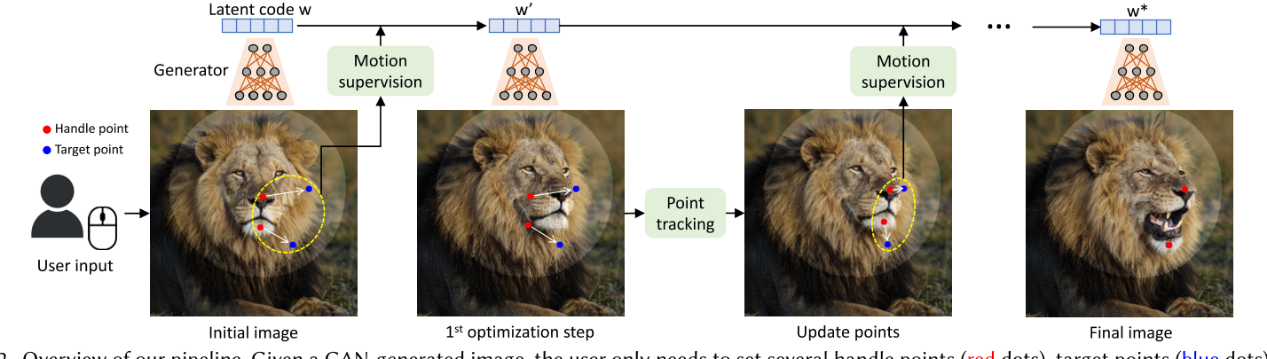
尽管以往的工作都能比较出色地完成图像编辑的任务，然而它们均存在一些缺点。例如基于监督学习的方法需要大量的标签属性才能生成比较多样的图片；基于语义图的方法当用户需要较大范围地改变图中的物体时需要重新绘制语义图；基于文本驱动的方法不能做到比较精确地图像编辑。因此，DragGAN基于以上方法的缺点，结合以往拖拽式的方法，首次提出了基于生成模型的拖拽式的实时交互式的图像编辑方法，兼顾实时便捷的同时，还能做到非常精确的图像编辑。下面将详细介绍DragGAN的图像编辑方法。

### 总体流程

DragGAN模型方面采用了StyleGAN[6]模型，StyleGAN模型是一个基于样式的GAN模型。我们知道，GAN模型能够通过模型中的生成器实现从一个隐向量w到一张图片的完整转换，因此我们能够通过对隐向量w做一些扰动或者变换就能够实现对GAN模型生成出的模型进行编辑。

为了使得DragGAN模型能够实现拖拽式的图像编辑，作者为DragGAN模型设计两个用于拖拽的点。如下图所示，红色的点代表要被拖拽的点，蓝色的点代表拖拽的目标位置。因此，在DragGAN进行图像编辑的过程中，我们所需要做的就是当红色的点被拖拽向蓝色点运动时，不断地更新隐向量w来更新图片达到图片的编辑效果。

具体而言，DragGAN将图片更新的过程分解为多步，一点一点地让红色点靠近蓝色点，从而实现图像的编辑。在每一步中又可以拆解为两大过程，分别是运动监督和点跟踪，下面将详细介绍这两个过程。

图 3.1 DragGAN的总体流程

### 运动监督

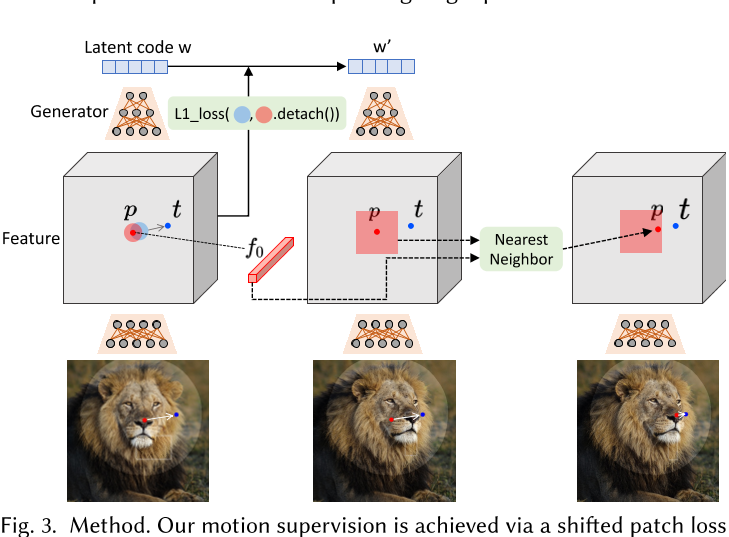
DragGAN模型在运动监督时并没有采用非常传统的光流模型来实现，因为如果采用光流模型来实现运动监督的话并不能达到实时的效果。对于运动监督，DragGAN采用了一个更为简明的方法，就是利用DragGAN模型中隐藏层的feature map来实现。以往的工作已经指出了GAN模型中隐藏层的feature map对生成图片的位置信息是敏感的，因此在红色点向蓝色点的每一小步的移动过程中，DragGAN首先会确定一个小的范围，随后分别在红色点及其一个小的邻域的点确定两个小范围的区域，为了实现拖拽的效果，DragGAN用红色点的小范围区域作为蓝色点区域的ground truth进行loss计算，反向回传使得隐向量w更新。此时用更新的隐向量w生成出的图片就是这一小步拖拽所得到的图片了。运动监督的图示如下图红框中所示。

图 3.2 DragGAN运动监督图示

### 点跟踪

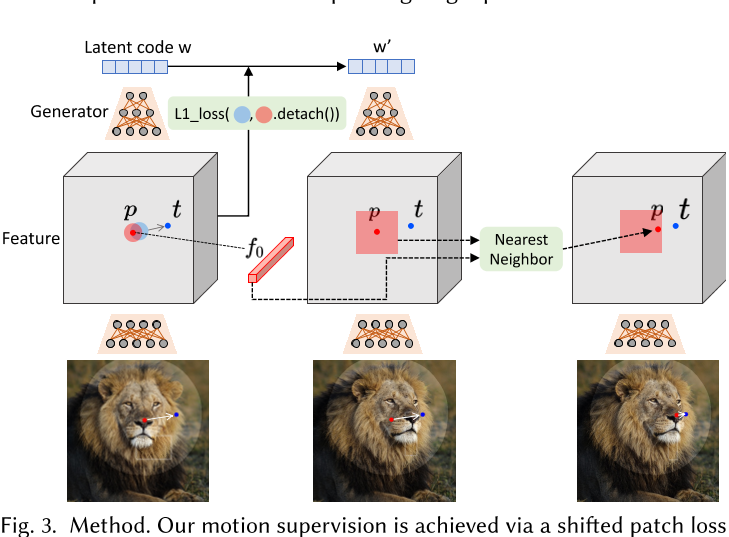
DragGAN点跟踪的实现也非常地简明，对于一个运动监督后生成的图片，我们需要将红色的锚点重新定位，新的锚点应该与原来图片中的锚点具有相同的语义。因此，我们只需要在feature map中红色点的一个小领域中去寻找一个新的点，这个点的feature map值应当是这个领域中与原始图片红色点的feature map值最接近，找到这个点后将其红色锚点重新定位到该点就完成了该小步的点跟踪的过程。点跟踪的图示如下图红框所示。

图 3.3 DragGAN点跟踪图示

### StyleGAN

DragGAN模型中运用到的模型是预训练好的StyleGAN，并没有对模型进行较大的改动，这里简要介绍一下StyleGAN。StyleGAN是基于样式的生成器，它摒弃了传统GAN的输入方式，将原来输入到GAN的传统的隐向量w改为通过全连接层嵌入的隐向量空间，模型的输入变为固定的，而嵌入的隐向量空间w作为模型的风格影响，输入到模型的AdaIN[7]模块中实现风格的控制。这种基于样式的GAN模型，能够自动学习高层次属性和生成图像的随机差异之间的无监督分离，并且使得图像合成具有直观的、针对不同尺度的控制。StyleGAN模型的结构图如下图所示。

图 3.4 StyleGAN的结构图

### 效果评估

为了体现DragGAN模型的强大功能，DragGAN的作者在论文中也做了大量其他的消融实验。

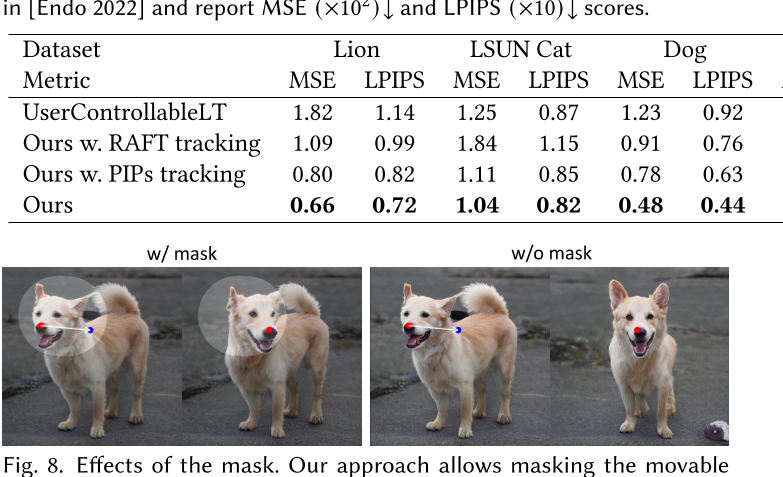
例如，通过遮罩实现拖拽只影响图中的一部分，如下图所示，左图是启用了遮罩的，右图是不启用遮罩的。可以看到当使用了遮罩时，左图仅会移动小狗的头部，而不使用遮罩时则是使得小狗整体转向。

图 3.5 DragGAN遮罩消融实验

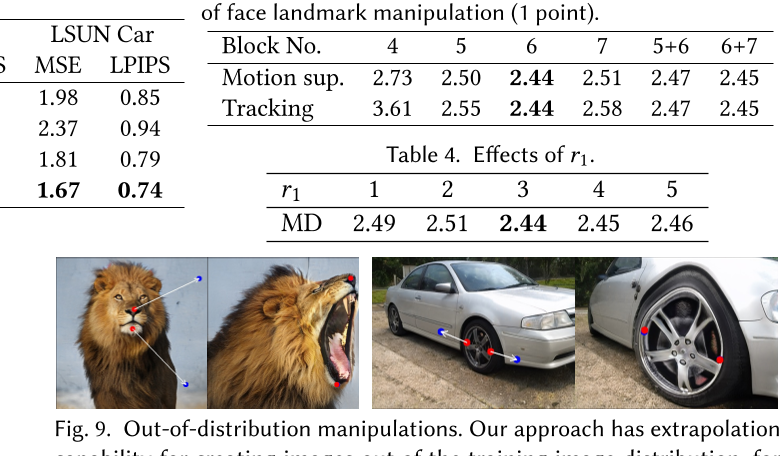
对于又遮罩的图片，例如左图想让狮子张开血盆大口，DragGAN也能很好地实现，并且能够精细地生成狮子口中的牙齿。

图 3.6 DragGAN细节生成消融实验

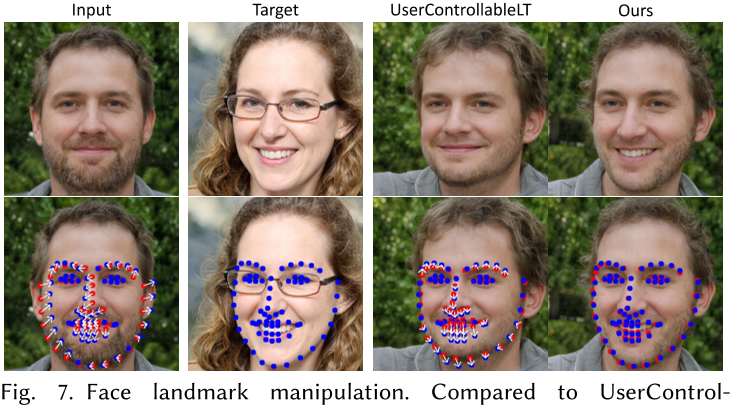
DragGAN还指出，该方法能够很好地用于基于关键点检测的图像编辑，例如下图中，通过关键点检测后得到人脸的关键点，此时利用DragGAN就能非常方便地将一个人像的姿势迁移到另一个人像中。

图 3.7 DragGAN关键点消融实验

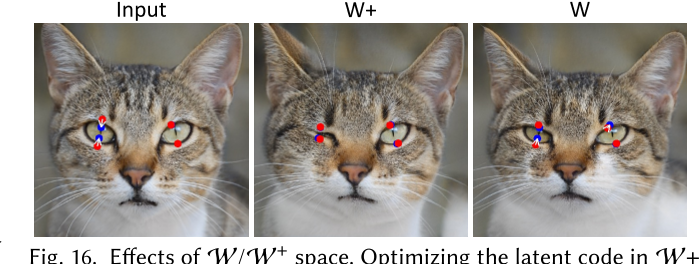
另外，DragGAN还指出若采用的StyleGAN采用不同隐向量空间W+/W会得到不同的效果。W+和W隐向量的不同之处在隐向量W输入到风格AdaIN中经过的仿射变换是否相同。论文指出W+隐向量生成的图像编辑所形成的效果会比W的大一些。

图 3.8 DragGAN W+和W隐向量空间不同生成效果

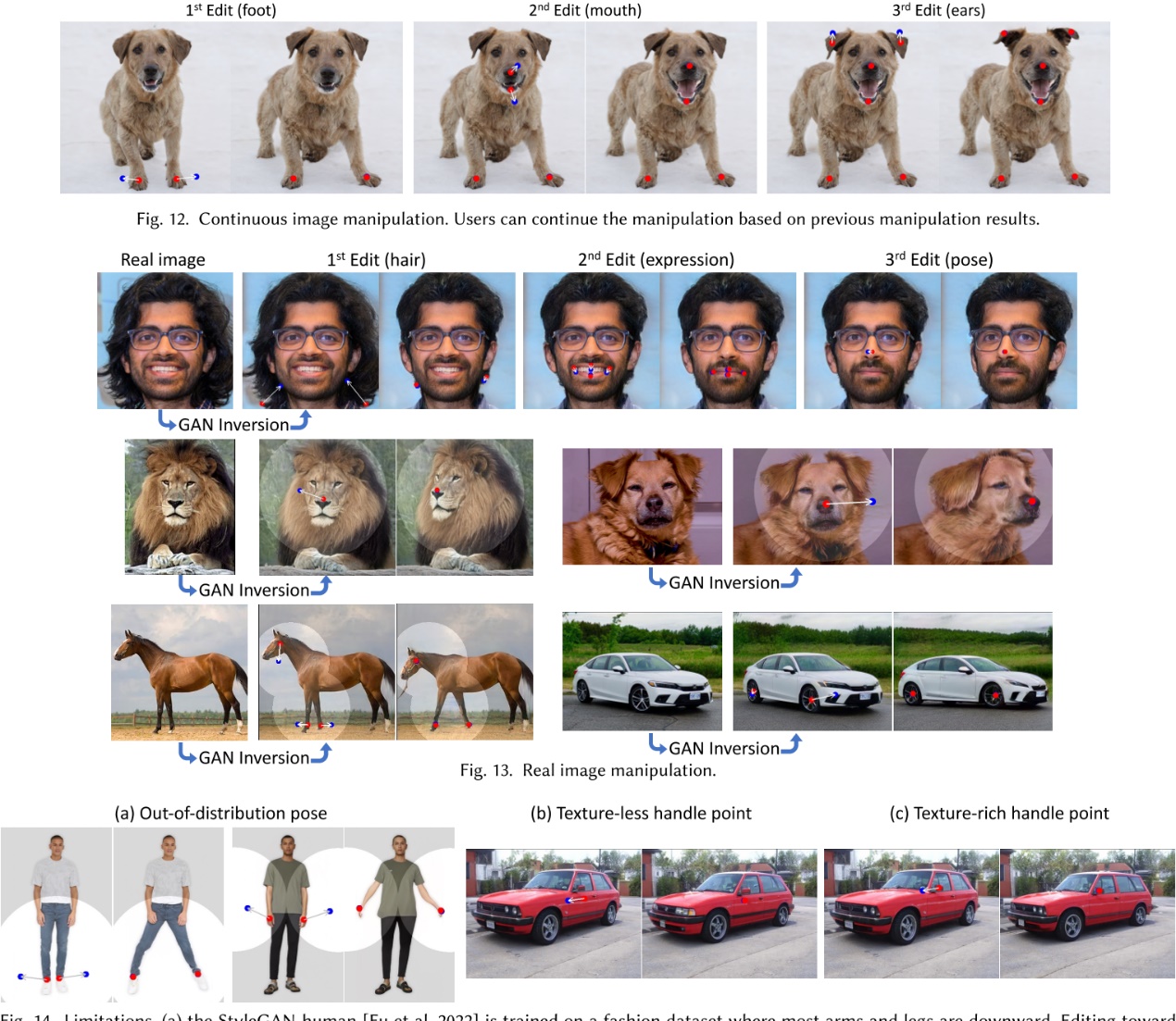
最后，论文也指出了一些不足之处，由于DragGAN采用的是预训练好的StyleGAN模型，对于真实图片的输入，由于首先要经过一个Inversion过程将真实图片转为隐向量，如果转换得到的隐向量超过了预训练好的StyleGAN模型的隐向量空间，那么可能会得到不好的编辑效果。

图 3.9 DragGAN对真实图片的编辑

## 总结

本次报告较为详细介绍了DragGAN的工作原理及其效果，同时也回顾了以往一些优秀的图像编辑算法。图像编辑作为生活中非常常见的需求，同时也是设计方面的常见任务，其重要性是不言而喻的，通过生成模型能够方便人们快速地完成图像的编辑，具有非常高的实际意义，相信在未来生成模型在图像编辑方面能够继续稳步发展，让人工智能技术不断造福人类。

## 参考文献

1. Pan X, Tewari A, Leimkühler T, et al. Drag your gan: Interactive point-based manipulation on the generative image manifold[C]//ACM SIGGRAPH 2023 Conference Proceedings. 2023: 1-11.
2. SHEN Y, YANG C, TANG X, et al. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 2004-2018. http://dx.doi.org/10.1109/tpami.2020.3034267. DOI:10.1109/tpami.2020.3034267.
3. PARK T, LIU M Y, WANG T C, et al. Semantic Image Synthesis With Spatially-Adaptive Normalization[C/OL]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. 2019. http://dx.doi.org/10.1109/cvpr.2019.00244. DOI:10.1109/cvpr.2019.00244.
4. Kawar B, Zada S, Lang O, et al. Imagic: Text-based real image editing with diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 6007-6017.
5. IGARASHI T, MOSCOVICH T, HUGHES J F. As-rigid-as-possible shape manipulation[J/OL]. ACM Transactions on Graphics, 2005: 1134-1141. http://dx.doi.org/10.1145/1073204.1073323. DOI:10.1145/1073204.1073323.
6. KARRAS T, LAINE S, AILA T. A Style-Based Generator Architecture for Generative Adversarial Networks[C/OL]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. 2019. http://dx.doi.org/10.1109/cvpr.2019.00453. DOI:10.1109/cvpr.2019.00453.
7. HUANG X, BELONGIE S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[C/OL]//2017 IEEE International Conference on Computer Vision (ICCV), Venice. 2017. http://dx.doi.org/10.1109/iccv.2017.167. DOI:10.1109/iccv.2017.167.