中图分类号：TP391　　论文编号：1028716　22-SZ002
学科分类号：0852

硕士学位论文

基于强化学习的人脸属性编辑方法研究

研究生姓名　　任国伟
专 业 类 别　　工程硕士
专 业 领 域　　计算机技术
指 导 教 师　　谭晓阳 教授

南京航空航天大学

研究生院 计算机科学与技术学院

二〇二一年十二月

Nanjing University of Aeronautics and Astronautics
The Graduate School
College of Computer Science and Technology

A Reinforcement Learning Approach for Face Attributes Editing

A Thesis in
Computer Science and Technology

by

Ren Guowei

Advised by

Prof. Tan Xiaoyang

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Engineering

December, 2021

承诺书

本人声明所呈交的博/硕士学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京航空航天大学或其他教育机构的学位或证书而使用过的材料。

本人授权南京航空航天大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本承诺书）

作者签名：_____

日　　期：_____

摘 要

人脸属性编辑作为计算机视觉发展到一定阶段的产物，在人脸图像增强和相关娱乐应用方面有不可替代的作用,其衍生出的面部老化和面部妆容在破获易容犯罪方面有重要的社会意义。人脸属性编辑旨在通过修改人面部的某一属性来获得修改后的期望图像，现有的人脸编辑方法主要分为基于模型的方法和基于附加条件的方法两大类。随着生成对抗网络的出现，人脸属性编辑算法取得了很大进步，但仍存在一些不足。首先，现有的方法大多只能实现二值化的编辑，即具有或不具有某项属性，属性的编辑缺乏连续性。其次，现有方法在保持身份特质方面尚有不足，容易造成其它无关属性的变化。针对上述问题，本文从生成对抗网络的潜在空间出发，提出一种探索式的人脸属性编辑方法，本文主要工作如下：

（1）生成模型是人脸属性编辑任务中的重要组成部分，现有方法大多通过在潜在向量上串联附加条件来实现潜在向量在空间中的跳跃，造成了编辑的不连续。针对这一问题，本文提出通过在潜在向量上叠加增量的方式对潜在空间进行连续性探索，并对探索过程建立了完整的马尔可夫决策过程模型。

（2）针对上述的马尔科夫决策过程模型，本文引入强化学习算法对其进行求解，提出了一种基于深度确定性策略梯度的人脸老化算法。智能体根据人脸初始潜在向量做出决策，并将决策向量叠加到初始向量上构成下一状态向量，同时获得一定的奖励。在交互过程中，智能体通过最大化收益的方式寻找一条最短的人脸老化路径。实验结果表明该算法不仅能够寻找到最佳的人脸老化路径，并且该路径还具有通用性，可以用于其他人脸的老化编辑过程。

（3）在人脸老化实验过程中，需要用到大量带标签的人脸数据集训练年龄预测器，针对这一弊端，本文利用已有的属性分类器构建了网络交互模型，不仅避免了多标签数据集的使用，还能够减少本地运算负载。面对纷杂的属性，本文提出了基于双延迟深度确定性策略梯度的任意属性编辑算法，采取更加细致探索，并通过添加引导奖励加快学习速度。实验结果表明双延迟深度确定性策略梯度算法能够学习到更加稳定的策略，基于该策略的决策向量能够对属性进行连续的编辑。此外，多种属性的编辑实验表明只要能够对属性做出评价，就可以实现任意属性的编辑。

关键词：人脸属性编辑，面部老化，强化学习，生成对抗网络，年龄预测，属性分类

ABSTRACT

As a product of the development of computer vision to a certain stage, face attribute editing plays an irreplaceable role in face image enhancement and entertainment applications. Face aging and facial makeup derived from it have important social significance in cracking down on transfiguration crimes. Face attribute editing aims to obtain the desired image by modifying a certain attribute of human face. The existing face editing methods are mainly divided into two categories: model-based methods and extra condition-based methods. With the emergence of generative adversarial network, face attribute editing algorithms have made great progress, but there are still some shortcomings. Firstly, most of the existing methods can only achieve binary editing, that is, they have or do not have a certain attribute, and the editing of attributes lacks continuity. Secondly, the existing methods are still insufficient in maintaining identity traits, which is easy to cause changes in other irrelevant attributes. Aiming at the above problems, this paper proposes an exploratory face editing method from the perspective of the latent space of generative adversarial network. The main work of this paper is as follows:

(1) Generative model is an important part of face attribute editing task. Most of the existing methods achieve the jump of latent vector in space by concatenating additional conditions on latent vector, resulting in discontinuity of editing. To solve this problem, this paper proposes to continuously explore the latent space by superimposing increment on latent vector, and establishes a complete Markov decision process model for the exploration process.

(2) For the above Markov decision process model, this paper introduces reinforcement learning algorithm to solve it, and proposes a face aging algorithm based on deep deterministic policy gradient. The agent makes a decision according to the initial latent vector of the face, and superimposes the decision vector on the initial vector to form the next state vector, and obtains a certain reward at the same time. During the interaction process, the agent seeks the shortest face aging path by maximizing the total reward. Experimental results show that the algorithm can not only find the best face aging path, but also has universality, and can be used in the aging editing process of other faces.

In the process of face aging experiment, a large number of labeled datasets are needed to train the age predictor. In view of this disadvantage, this paper uses the existing attribute classifier to construct the network interaction model, which can not only avoid the use of multi-labeled datasets, but also reduce the local computing load. Facing the complex attributes, this paper proposes an arbitrary attribute editing algorithm based on twin delayed deep deterministic policy gradient, which takes more detailed exploration, and speeds up the learning speed by adding guidance rewards. The

experimental results show that the twin delayed deep deterministic policy gradient algorithm can learn a more stable policy, and the decision vector based on the policy can edit the attribute continuously. In addition, the editing experiments of various attributes show that the editing of any attribute can be achieved as long as the attribute can be evaluated.

Keywords: face attributes editing, face aging, reinforcement learning, generative adversarial network, age estimation, attributes classification

目录

图表清单

注释表

z	潜在向量	φ'	Critic 目标网络参数
w	中间潜在向量	τ	软更新系数
z_0	人脸初始潜在向量表示	σ	动作噪声方差
w_0	初始人脸状态表示	π	策略
w_n	老化后人脸状态表示	r , r_i	奖励
d	老化路径	r_{ref}	引导奖励
α	系数因子	$\boldsymbol{l}_{\boldsymbol{ref}}$	引导向量
a , a_1, a_i	动作	φ_1	Critic 当前网络 1 参数
θ	Actor 当前网络参数	φ_1'	Critic 目标网络 1 参数
φ	Critic 当前网络参数	φ_2	Critic 当前网络 2 参数
m	数据样本个数	φ_2'	Critic 目标网络 2 参数
θ'	Actor 目标网络参数	ε	截断噪声

缩略词

缩略词	英文全称
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
DIAT	Deep Identity-aware Transfer
UNIT	Unsupervised image-to-image translation networks
CVAE	Conditional Variational Auto-Encoder
CAAE	Conditional Adversarial Auto-Encoder
DBM	Deep Boltzmann Machines
DBNs	Deep Belief Networks
PCA	Principal Component Analysis
VAE	Variational Auto-Encoder
KL	Kullback-Leibler
BN	Batch Normalization
DCGAN	Deep Convolutional Generative Adversarial Nets
ReLU	Rectified Linear Unit
WGAN	Wasserstein Generative Adversarial Nets
CGAN	Conditional Generative Adversarial Nets
ProGAN	Progressive Generative Adversarial Nets
JS	Jensen-Shannon
AdaIN	Adaptive Instance Normalization
MDP	Markov Decision Process
DQN	Deep Q Network
SARSA	State-Action-Reward-State-Action
DPG	Deterministic Policy Gradient
DDPG	Deep Deterministic Policy Gradient
TD3	Twin Delayed Deep Deterministic Policy Gradient
ResNet	Residual Network
FFHQ	Flickr-Faces-High Quality
SSIM	Structure Similarity

第一章 绪论

1.1 课题研究背景及意义

近年来随着互联网数据的积累和计算机硬件特别是独立显卡性能的不断提高，人工智能迎来了新的爆发。深度网络的诞生改变了人们对传统神经网络的认知，极大推动了机器学习和人工智能的发展，深度学习模型与卷积神经网络（Convolutional Neural Network, CNN）[1]的结合在计算机视觉领域取得了重大的突破，图像识别的准确率得到大幅度提高。其中，人脸作为人际交往和各种社会活动中出现频次最高的图像，成为计算机视觉领域研究的重点内容。目前，人脸识别技术在银行、火车站等公共场所得到了广泛的应用，给人们的生活带来了极大的便利。生成对抗网络（Generative Adversarial Network, GAN）[2]的诞生为人脸属性编辑提供了基础，人脸属性编辑是一项通过修改面部相应特征来获得预期人脸图像的技术，目前多应用在移动端娱乐软件中，其衍生出的面部妆容和面部衰老等应用，在追查易容罪犯的过程中也发挥着重要的作用。除此之外，人脸属性编辑在图像数据增强方面也存在着很大的潜力，一些人脸数据集可以通过人脸属性编辑的方式进行扩充，间接提高模型的鲁棒性。人脸属性编辑作为人脸相关研究中极富挑战性的任务之一，无论是从商业娱乐的角度还是从社会科研角度来看，均具有重要的研究价值。

人脸属性编辑可以看成人脸图像伪造方式的一种，即面部表情操作，与之互补的是面部身份操作，二者共同组成了人脸图像伪造。人脸图像伪造并不是新兴的概念，它最早出现在电影工业中，即电影 CG 技术。利用这种技术可以制作出原本并不存在的虚拟人物，提高电影的观赏性，但其制作过程较为复杂，通常需要借助专业设备进行动作捕捉和面部表情跟踪才能制作出逼真的画面。2016 年 3 月，Justus Thies 等人提出的 Face2Face[3]基于无标记的面部表情捕捉技术，实现了面部表情、肌肉变化从源角色到目标角色的实时复制。虽然该技术备受争议，但相关研究并没有停止，深度造伪（DeepFake）技术的出现造成了互联网上假视频的泛滥，引发了人们对身份盗用、虚假信息传播造成社会恐慌的担忧。从技术角度看，DeepFake 属于深度自动编解码器（Auto Encoder-Decoder）模型，关键步骤为编码器提取面部图像的潜在特征信息，解码器根据这些信息重建面部图像。为了实现源图像和目标图像之间的面部交换，需要用到两组自动编解码器对，其中编码器参数在两个网络中共享，这样可以使编码器学习到两组面部图像之间的相似性。

相比之下，人脸属性编辑则偏向对同一身份的面部图像进行细节方面的伪造，即在其它面部属性基本不发生改变的前提下，对指定的面部属性进行编辑，以保持图像的身份信息。目前主流的人脸属性编辑方法大致分为两类：基于模型的方法和基于附加条件的方法，二者均基于编码器-解码器体系，主要区别在于是否需要给出额外的条件。基于模型的方法通过将图像从源域映射到目标域[4]实现人脸属性的编辑。作为典型的基于模型的方法，

CycleGAN[5]在生成图像的真实性方面具有良好的表现，但是域到域之间的转换使得一次训练过程仅能对一个属性进行编辑，编辑效率相对偏低。基于附加条件的方法通过将人脸图像的潜在向量和附加属性向量串联起来表示结构化的人脸属性信息，相比于直接在难以理解的人脸图像潜在向量空间中进行属性编辑，这种方法更加直观和易于理解。附加属性向量的引入实现了对多个属性的控制，解决了一次训练只能编辑一个属性的难题。但是无论是基于模型的方法还是基于附加条件的方法，都只能实现二值化的编辑，造成了编辑过程的不连续。此外，由于编辑过程中直接对整张人脸图像进行编码，容易造成其它非编辑属性的改变，人脸身份保持效果不理想。

本文为弥补上述不足，在潜在向量与属性向量的组合上进行了思路创新，将属性向量与潜在向量的叠加问题转化为强化学习问题，提出了适用于求解属性向量的马尔可夫决策过程模型，针对人脸老化这一特殊过程，采取逐步探索的方式自动寻找最佳的人脸老化路径。同时为拓展适应任意属性的编辑过程，提出了一种利用现有属性分类器的网络模型，避免了对多标签数据集的依赖。

1.2 国内外研究现状

在人际交往和社会活动，人脸是给他人的第一印象，不但具有特殊的地位，也是计算机视觉研究的重点。上世纪五六十年代，认知科学家就开始了对人脸识别（face recognition）问题的探索，人脸识别[6]的工程化应用研究则在六十年代展开，这一时期人脸识别问题被当成一个模式识别问题来研究，主要通过人脸的几何结构来分析人脸器官特征点之间的拓扑关系，进而达到人脸识别的目的。但是这种方法对人脸姿态及表情的要求较高，因而并未投入实际使用。从 20 世纪 90 年代开始，人脸识别研究进入井喷期，这期间所提出的算法在理想图像采集条件下也取得了不错的性能。进入 21 世纪，随着深度学习和卷积神经网络的引入，人脸识别取得了质的飞跃，识别准确率已经可以与人类媲美。为了应对复杂场景下的识别，人脸检测（face detection）作为人脸识别中的重要一环，逐渐发展成为了独立的研究课题。近年来，随着生对抗网络的提出，人脸生成与人脸属性编辑技术成为研究的热点。

在生成对抗网络出现之前，关于人脸属性编辑的研究较为单一，大多是针对某一属性进行研究。在人脸老化的研究中，Ramanathan 等人基于物理建模推演的方式提出了面部生长模型[7]，该模型依据真实数据计算人脸在不同年龄阶段的变化情况。在人脸正面化和眼镜移除方面的研究中，有学者将人脸属性编辑问题看作回归问题，Zhu 等人[8]将任意角度人脸图像作为输入，通过最小化像素级图像重建损失实现了正面人脸的重构。张志刚等人[9]利用同一对象戴眼镜和不带眼镜时的图像组成训练集，训练了一个多元线性回归模型，该模型可以去除戴眼镜人脸图像中的眼镜。

训练上述回归模型往往需要成对的人脸图像数据集，由于缺乏大规模数据集的支持，这类方法并未取得突破性进展。目前主流的人脸属性编辑方法大都基于生成模型，如变分自编码器和生成对抗网络，通过生成重建的方式达到人脸属性编辑的目的。根据是否需要给出额

外的控制条件，人脸属性编辑方法又分为基于模型的方法和基于附加条件的方法。

基于模型的方法本质上是学习两个图像域的对应关系，通过从源图像域到目标图像域的映射实现属性级别的编辑。作为典型的基于模型的方法，DIAT 模型[10]利用 VGG 网络[11]提取原图像与生成图像的特征图来计算身份损失，并将生成图像输入判别网络计算对抗损失，二者分别约束身份和属性，使生成图像在保持身份信息的前提下实现相应属性的编辑。在此基础上，Zhu 等人引入了循环一致性思想，提出了 CycleGAN[5]，它由两个完整的生成对抗网络组成，分别执行相反的域间转换，呈对偶结构。图像从源域转换到目标域后再经过反向转换，得到的图像应该和原始图像具有相似的特征，因此根据是否具有某一属性将人脸图像分为两个域，通过学习两个图像域之间的映射就可以完成人脸属性的编辑。在 UNIT 模型[12]中，作者假设在不同域转换前后的图像可以用共享潜在空间中的同一向量表示，并且每个图像域与共享潜在空间都存在和 CycleGAN 相似的结构。该模型本质上是通过边缘分布来推断联合分布，共享潜在空间的假设解决了推断过程中的欠定问题，能够在无监督的条件下较好的实现人脸属性的转换。然而图像域之间的相互转换总是针对图像整体而言的，却忽略了面部细节，这使得在编辑某一特定属性时，其他无关的属性也会发生不可控的改变。为解决这一问题，Shen 等人[13]提出了残差学习，该模型由两个生成网络和一个判别网络组成，两个生成网络分别接收某一属性相反的人脸图像并生成对应的属性残差，再与输入部分叠加生成最终修改的人脸图像。残差思想的引入使得属性的修改更加注重图像局部，能够更好地学习到属性之间的差异。

相比于基于模型的人脸属性编辑方法，基于附加条件的人脸属性编辑方法由于可以在一次训练过程中实现多个属性的编辑，受到研究人员的广泛关注。基于变分自编码器模型，条件变分自编码器（Conditional Variational Auto-Encoder, CVAE）[14]在编解码阶段加入标签的 one-hot 编码向量作为条件，训练好的模型可以根据不同的条件向量生成期望的图像。Attribute2Image[15]提出了一种分图层的生成模型，将图像分为前景和背景，其中前景部分采用了 CVAE 模型，将潜在编码向量与属性条件连接对生成图像的属性进行控制，降低了采样的不确定性，提高了生成图像的质量。在人脸年龄编辑方面，Zhang 等人提出了条件对抗自编码器（Conditional Adversarial Auto-Encoder, CAAE）模型[16]，该模型假设人脸图像是高维流形上的点，当从这些点出发延某一方向运动时，能够保证在身份不变的同时实现年龄的编辑，CAAE 将年龄向量和人脸的编码向量连接后一同输入到解码器中，其中编码得到的潜在向量用于控制身份特征，附加的年龄向量负责控制年龄。可逆条件生成对抗网络（invertible Conditional GAN, IcGAN）[17]利用两个编码器分别编码潜在特征向量和条件向量，和 CAAE 不同的是，IcGAN 在条件向量部分引入了多个属性，实现了对多个属性的编辑。AttGAN[18]也采用了类似的多属性条件向量的结构，通过引入属性分类约束保证编辑目标属性时，其他属性不发生改变。几乎同时期，StarGAN[19]利用掩码向量控制不同图像域的标签信息，仅用一个生成器和鉴别器就实现了多个图像域之间的转换。StarGAN 与 AttGAN 均采用了对抗损

失、重构损失和分类损失，但是 StarGAN 仅能实现二值化的属性编辑，而 AttGAN 可以控制属性编辑的程度，在一定程度上实现了属性的连续编辑。Lamlpe 等人提出的渐变神经网络（Fader Network）[20]将生成对抗网络的思想融入到自编码器模型中，通过引入一个鉴别器和编码器构成生成对抗组，对抗的最终结果使编码器产生的编码不包含性别这一属性信息，在解码过程中再将控制性别的向量与编码器生成的向量相结合，实现了性别的渐进编辑。和基于模型的方法一样，基于附加条件的方法也存在无关属性发生改变的情况，为此 Zhang 等人在生成对抗网络中加入了空间注意力机制，提出了能够精确编辑遍面部属性的 SaGAN[21]。在 SaGAN 的生成器中，空间注意力网络通过学习蒙版定位需要编辑的属性，属性编辑网络则根据该蒙版对相应的属性进行编辑，不过该方法仅能对人脸局部特征属性进行编辑，而对复合特征属性如年龄、性别则无能为力。

除了上述两大类方法，基于隐层交换的人脸编辑方法也逐渐成为当下研究的热点，GeneGAN[22]将图像编码为待编辑的属性向量和其他属性向量，通过交换待编辑部分的属性向量实现人脸属性的交换，由于交换的属性都来源于真实图像，解码后生成图像的真实性很高。ELEGANT 模型[23]将编码得到的潜在向量分成多个部分，每个分部代表不同的属性特征，实现了对多个属性的编码，残差学习和多尺度判别器的引入保证了生成图像的细节。InterFaceGAN[24]对生成对抗网络中潜在空间进行了语义上的分析，并提出用一个超平面对人脸同一属性进行二值划分，以其法向量作为该属性的方向向量，并采用正交分解的方式对不同属性进行解耦，实现了人脸属性的自由编辑。

不难看出，作为计算机视觉领域的热点问题，国内外学者在生成对抗网络、属性迁移和人脸属性编辑方面做了大量的研究，取得了丰富的成果，直接或间接地推动了人脸属性编辑的进步，编辑效率得到了大幅提升，生成图像的质量也逐渐提高。

1.3 本文研究思路及内容

人脸属性编辑是一项通过修改面部属性获得期望图像的技术，在图像数据增强方面有很大潜力。随着计算机视觉的发展，人脸属性编辑取得了巨大进步，但依旧存在一些不足之处。从国内外研究现状可以看出，在现有的基于模型和基于附加条件的两类方法中，人脸属性编辑的连续性和编辑过程中身份信息的不变性并不可兼得，大多数基于模型和基于附加条件的方法能够仅能够实现属性的二值化编辑，如发色属性的黑发和白发，并没有中间值灰发，而且黑发和白发并不能完全代表年龄这一复杂的复合属性。一些方法，如 InterFaceGAN 和 AttGAN 虽然能够实现或近似实现属性的连续编辑，但编辑前后图像中人脸的身份信息会发生不可控的改变。针对上述人脸属性编辑连续性和编辑过程中无关属性发生改变两个问题，本文从生成对抗网络的潜在空间入手，通过对人脸潜在表示的周围进行探索，实现人脸属性的编辑。本文的主要研究内容如下：

（1）根据生成对抗网络生成图像的基本原理，潜在向量发生变化时，生成图像的属性会发生变化，本文提出通过在潜在向量上叠加增量的方式逐步探索潜在空间，寻找实现属性

编辑的向量路径，并对这一探索过程进行马尔可夫决策过程建模，引入强化学习，将其看作强化学习问题。

（2）根据上述建模过程，针对人脸老化这一特殊过程，本文提出利用深度确定性策略梯度算法对其进行求解。在奖励函数的设计过程中，本文利用 RexNeXt-50 网络训练了一个年龄预测器，该预测器可以对探索过程中生成图像的年龄进行实时预测，并将计算得到的奖励反馈给智能体，此外，本文还利用交互惩罚对探索路径进行约束，确保最佳路径仅在年龄这一复合属性上发生改变，从而获得良好的身份信息保持效果。实验结果表明深度确定性策略梯度不仅能够正确找到人脸老化路径，实现良好的人脸年龄编辑效果，而且该路径还具有一定的通用性。

（3）在人脸老化的编辑过程中，虽然实现了连续性编辑，并且人脸身份信息也得到较好的保持，但这种连续性并不细致，在其他属性的编辑过程中有可能出现离散二值化的情况。针对这一问题，为实现任意属性连续编辑，本文提出利用现有的分类器搭建网络交互模型，对智能体的决策向量进行严格控制，利用双网络缓解强化学习中 Q 值过估计的问题，通过延迟更新使智能体能够学习到更加稳定的策略。实验结果表明，基于双延迟 DDPG 的网络交互模型能够利用现有分类器实现细致化的人脸属性连续编辑，避免了多标签数据集的使用，理论上可以对任何可以评价的属性进行编辑。

1.4 本文内容安排

本文的章节安排如下：

第一章：本章为绪论，首先介绍人脸属性编辑的研究背景和意义，并简述现有的两类算法的优缺点，随后介绍了国内外的研究现状及取得的进步。

第二章：本章主要介绍人脸属性编辑中的重要组成部分生成模型及本文方法中要用到强化学习的相关理论基础。首先从传统的自编码器入手，逐渐引出生成模型的发展历程，针对变分自编码器和生成对抗网络这两类常见的生成模型从理论层面做了详细的介绍。随后按照时间顺序依次介绍了生成对抗网络在发展过程中比较重要的改进 DCGAN 和 WGAN，生成对抗网络在高分辨率图像生成方面取得的突破 ProGAN 和 StyleGAN。本章最后对强化学习模型和常见的算法进行了简单介绍。

第三章：本章提出了基于深度确定性策略梯度的年龄老化算法。本章首先强调年龄编辑的特殊性和连续性编辑的要求，紧接着在相关研究的基础上提出了人脸老化轨迹的假设，根据该假设详细阐述了建模思路并根据该思路对人脸老化问题的求解建立了强化学习模型，随后对奖励函数的设置做了详细介绍，利用 ResNeXt-50 网络搭建并训练了一个年龄预测器为主线奖励的计算提供支持。接下来对深度确定性策略梯度算法在人脸老化任务中的具体表现形式及相关参数设置进行了简要的分析。实验结果表明先前的假设正确，智能体找到了老化路径，最后对该路径下的人脸老化效果进行相关分析，发现其具有通用性。

第四章：本章首先肯定了强化学习方法在人脸老化编辑中的可行性，接着对其中存在的

编辑不细致的问题进行了分析，考虑到人脸老化实验中需要大量带标签数据集训练年龄预测器，本章摒弃了训练属性分类器的思路，转而采用已有的属性分类器，提出基于双延迟 DDPG 算法的任意人脸属性编辑模型。随后，利用已有的百度智能云人脸属性分析方案搭建了网络交互模型并提出引导奖励加快训练速度。针对深度确定性策略梯度存在的对 Q 值的过估计问题，采用双网络和 Actor 网络延迟更新的方式加以改进，最后介绍了试验过程中用到的训练技巧并对实验结果进行了相关分析。

第五章：对本文的主要工作进行了简要的总结和概括，指出工作中的不足及可能出现的问题，并对未来工作了进行了展望。

第二章 生成模型和强化学习基础

2.1 引言

在人脸属性编辑方法中，除了传统的物理模型[25-28]外，无论是基于模型的方法还是基于附加条件的方法，几乎都以生成模型为基础。生成模型是机器学习和概率统计中一类常见的模型，一般用于随机生成观测数据，如文本、音频和图像等。在人脸属性编辑的过程中，生成模型被用于人脸图像的重建，生成模型的好坏直接决定编辑后图像的质量，因此是人脸属性编辑中的核心组成部分。

2.2 深度生成模型：从自编码器到生成对抗网络

生成模型的主要作用是从给定的训练数据中学习一个概率分布，并利用该分布模拟数据的真实分布，进而生成具有变化的数据样本。样本数据量越多，生成模型学习到的概率分布与数据的真实分布越接近。对于显性密度模型，我们可以利用极大似然法直接估计其概率密度并给出确切的分布函数，但对于高维数据，并不存在显性的密度分布，难以直接对其建模。这种情况下，就需要用到深度生成模型，深度生成模型采用无监督的方式学习数据的复杂分布，利用深度神经网络能够近似任意函数这一特性实现对复杂分布的逼近。常见的深度生成模型有深度玻尔兹曼机（Deep Boltzmann Machines，DBM）[29]、深度信念网络（Deep Belief Networks，DBNs）[30]、变分自编码器和生成对抗网络等。

2.2.1 自编码器

自编码器并不是生成模型，但在其演化和发展过程中出现了很多变体，如正则自编码器、稀疏自编码器[31]、去噪自编码器（Denoising AutoEncoder）[32]及变分自编码器等。自编码器属于无监督学习范畴，能够从输入的数据中学习到这些数据的高效表示，是一种数据压缩算法。与主成分分析（Principal Component Analysis , PCA）[33]类似，二者都是通过特征转换降低数据维数进而去除数据的冗余，达到用少量的特征来尽可能完整的描述复杂数据的目的。与主成分分析不同的是，自编码器是一种神经网络，在编码过程中，既可以表征线性变换，也可以表征非线性变换，而主成分分析只能够进行线性变换，这使得自编码器要比主成分分析更加灵活。拥有多个隐藏层且采用非线性激活函数的自编码器在图像的特征抽取中有着重要的作用。

自编码器通常由编码器（Encoder）和解码器（Decoder）两部分组成，如图 2.1 所示，编码器对输入的图像进行压缩，得到一组向量表示，解码器对该向量表示进行解码重建，将其恢复成图像。当编码器或解码器单独存在时，并不具备学习数据特征或压缩数据的能力，为了使其能够学习到数据的特征表示，必须将二者进行结合训练。

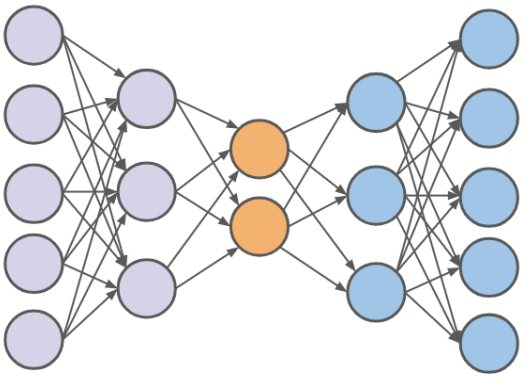

图 2.1 自编码器简易结构示意图

图像 x 经过向量化后由左侧输入（紫色），由神经网络将其压缩为较短的编码向量 z（橙色），该向量也被称之为“瓶颈”，“瓶颈”长度小于输入长度的自编码器也被称作欠完备自编码器，“瓶颈”的存在可以使神经网络对输入数据进行压缩，并从中学习和提取训练数据的特征表示。上述编码过程可以用函数 $z = f\ x$ 表示，解码过程中，编码向量 z 经过解码网络被还原成图像 $\hat{x}$（蓝色），可用函数 $\hat{x} = g(z)$ 表示。通过最小化重建损失 $L(x,\hat{x})$，还原之后的图像应该与输入的图像尽可能的一致，即 $\hat{x} = g(f(x)) = x$。如果自编码器使用线性激活函数并且损失函数是均方差

$$Loss_{mse} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{2.1}$$

则该自编码器可用于实现主成分分析。不完备的自编码器一般没有明确的正则化项，只是根据损失函数来进行训练，容易使模型对输入的数据产生记忆。为了使模型正确学习到数据中的潜在特征，需要添加各种形式的正则化项，以获得较好的泛化性能。以 MNIST 手写数字数据集为例构建的一个堆叠自编码器网络如图 2.2 所示，

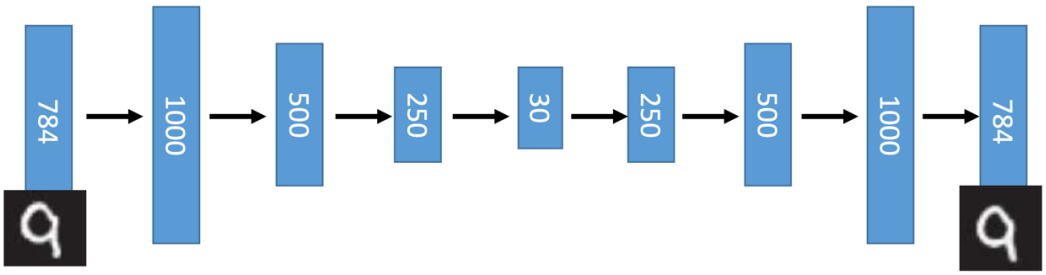

图 2.2 含多个隐藏层的自编码器网络结构图

该网络共有 7 个隐藏层，在编码阶段，图片被拉直成 1×784 的向量作为输入，经过三个隐藏层被输出为 1×30 的编码向量；在解码阶段，用 1×30 的编码向量作为输入，执行相反的操作，对图像进行重建。在图中可以看到数字“9”重建后的效果（右），跟原图（左）相比，两者之间的差异非常小。但是自编码器自身存在着比较大的缺陷，普通的自编码器在处理比较复杂的图像时，效果并不理想，重建后的图像会出现模糊，细节丢失严重。此外，自编码器并不属于生成模型，而更像是数据压缩存储模型，它只能对训练过程中遇到的编码进行解码。当对解码器输入一个训练集中从未出现过的编码时，重构后输出的图像可能是一

幅杂乱的噪声图像，因此并不具备生成新数据的能力。一个真正意义上的生成模型应该能够根据任意的编码向量重建出有意义的图像。

2.2.2 变分自编码器

普通自编码器只能通过神经网络实现对原数据的压缩与重建，并没有学习到原数据的分布情况，因而不能作为生成模型使用。为了能够生成与原数据相近的数据，模型必须能够学习到原数据的分布信息，因此需要在模型中引入随机性，即变分自编码器（Variational Auto-Encoder , VAE）。

变分自编码器最早由 Kingma[34]提出，后来 Doersch[35]又对其进行了详细的阐述。相比普通自编码器，经过改进之后的变分自编码器能够根据任意给出的编码向量重构图像。变分自编码器（VAE）与普通的自编码器的结构类似，仅在编码阶段存在不同，变分自编码器的网络结构如图 2.3 所示，

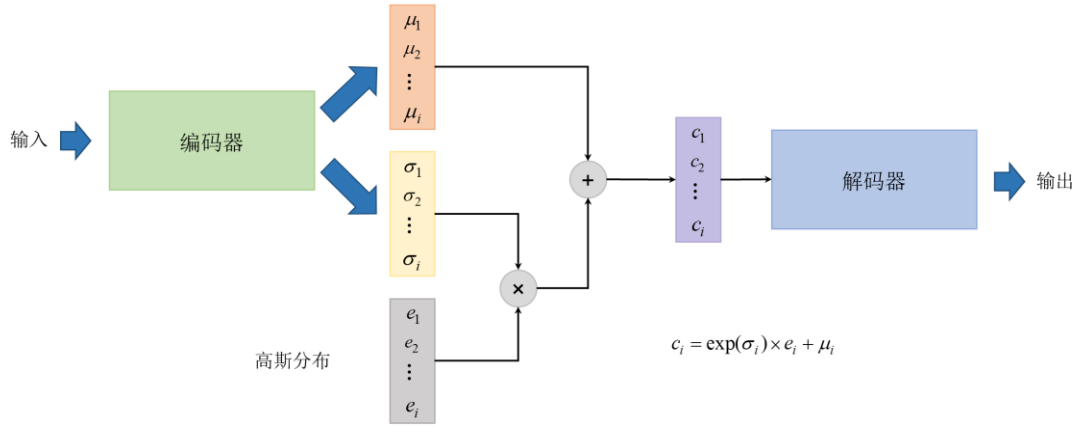

图 2.3 变分自编码器结构示意图

相比普通自编码器只将输入数据编码成一组向量，变分自编码器将输入数据编码成两组向量 μ 和 σ。其中 μ 表示均值，可以理解为普通自编码器的编码后向量，σ 表示添加噪声的方差，由于 σ 是由神经网络编码得到的，可能会出现小于 0 的输出，因此需要对其进行以 e 为底的指数运算，保证方差 $\exp(\sigma)$ 为正数，e 表示从高斯分布中采样出来的高斯噪声，其方差是固定的，但是在与 $\exp(\sigma)$ 相乘之后便产生了不同大小的方差。加入了噪声之后的编码 z 可以表示为 $z_i = \exp(\sigma_i) \times e_i + \mu_i$。加入了高斯噪声后的编码在训练过程中能够使解码器对噪声有较好的鲁棒性。但这并不意味着变分自编码器的损失函数可以简单的继承普通自编码器的损失函数。单一的重构误差约束会使网络学习到的噪声方差逐渐趋于 0，这样一来会造成随机性的下降甚至消失。缺乏随机性的变分自编码器实际上已经退化成了普通自编码器，无法生成新的数据。为了保证模型的随机性，除了需要最小化重构误差外，还需要最小化这一项：

$$\sum_{i=1}^{n}(\exp(\sigma_i) - (1+\sigma_i) + \mu_i^2) \tag{2.1}$$

该项实际上是对编码器生成的方差进行了限制，抛开均值项先不看，当 $\exp(\sigma)-(1+\sigma)$ 取最小值时，$\sigma=0$，但实际的方差 $\exp(\sigma)=1$，因此该项约束保证了噪声的方差不会太小。后面的 μ^2 是 L2 正则项约束，目的是保证模型不会过拟合。

变分自编码器采用无监督的方式学习数据的分布情况，与普通自编码器相比，它继承了普通自编码器的一般结构，随机性的引入使模型对噪声有一定的鲁棒性，能够生成不存在的数据，被广泛应用于基础数据的生成。但是由于分量独立的混合高斯模型并不能拟合任意的概率分布，变分自编码器并没有真正学习到数据的概率分布，而只是对真实分布的一个近似。因此生成的图像特别是复杂图像比较模糊，除英伟达实验室提出的 NVAE[36]方法外，其它基于变分自编码器的图像生成方法在生成图像的质量上均不及基于生成对抗网络的图像生成方法。

2.3 生成对抗网络

生成对抗网络(Generative Adversarial Networks, GAN)[2]是一种不依赖任何先验假设的深度生成模型，它由 Ian Goodfellow 在 2014 年提出。一经提出，生成对抗网络就在机器学习领域掀起了一场史无前例革命，并迅速成为研究的热点。生成对抗网络被誉为近年来在复杂分布上最具前景的无监督学习方法之一，其思想也被广泛应用于机器学习的各个领域。图灵奖得主 Yann LeCun 甚至评价生成对抗网络是“自切片面包出现以来最酷的发明”。

2.3.1 基本结构

生成对抗网络由生成器（Generator）和鉴别器（Discriminator）两个部分组成，其基本思想来源于博弈论中的零和博弈，生成器和鉴别器在互相博弈的过程中逐渐提升自己被赋予的处理问题的能力，其基本的结构如图 2.4 所示，

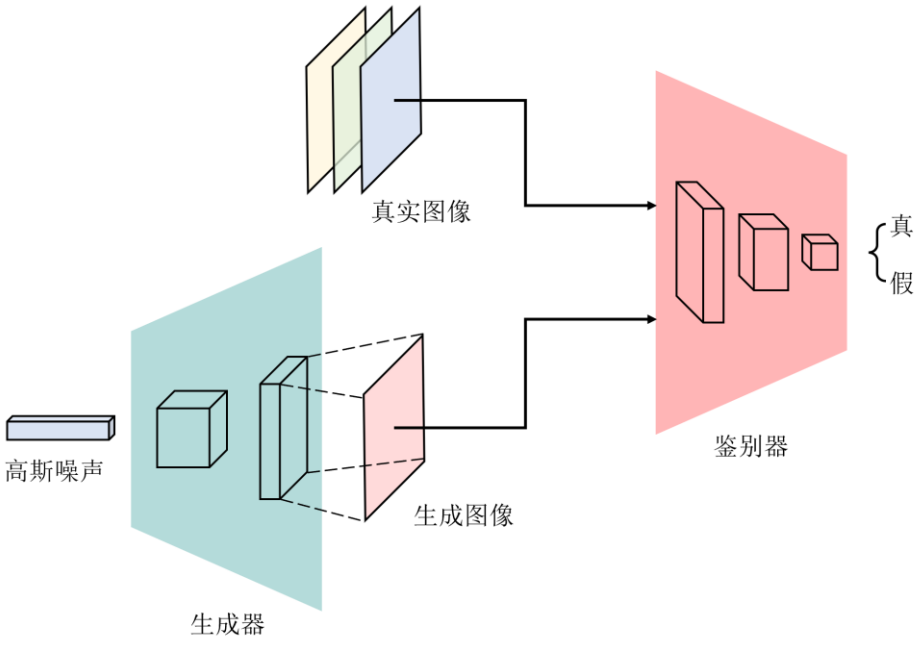

图 2.4 生成对抗网络结构示意图

生成器 G 以从某个分布（一般是高斯分布）中采样得到的 n 维向量 z 作为输入，以生成的图像 $G(z)$作为输出；鉴别器则以真实图像 x 或生成器生成的图像 $G(z)$为输入，通过神经网络计算并输出二者的评分（一般是图片为真实图像的概率），并根据这个评分对输入的图像是否为真实图像进行判定。一方面，生成器需要生成图像尽可能使鉴别器将其判定为真，另一方面，鉴别器要尽可能的鉴别出生成器生成的图像为假。当鉴别器无法分辨出输入的图像是真实图像还是生成器生成的图像[37]时，训练结束，称之为达到了纳什均衡。达到纳什均衡的生成对抗网络能够通过生成器生成以假乱真的图像。最初提出的生成对抗网络中，生成器和鉴别器只要能够拟合函数即可，并不一定是神经网络，但是由于具有两个隐藏层的神经网络几乎可以近似任何函数，实际应用中生成器 G 和鉴别器 D 均由神经网络或是更加复杂的卷积神经网络组合构成，并且二者在结构上具有对称性。可以看出，从 z 到 $G(z)$是一个高度复杂的映射关系，这使得 $G(z)$具备拟合几乎任何复杂分布的能力。强大的拟合能力使得生成器能够生成人类都难以分辨的图像，因此也被用于数据增强，以辅助其他机器学习任务。

2.3.2 模型理论

以图像生成问题为例，在生成对抗网络诞生之前，一般利用极大似然估计这样一个分布，即寻找一组参数 θ，使由参数 θ 决定的分布 $p_g(x;\theta)$ 和真实数据分布 $p_{data}(x)$ 尽可能地接近。从真实数据分布 $p_{data}(x)$ 中采样 m 个样本点 x_1, x_2 ,…, x_m，则生成这样一组数据样本的似然函数为

$$L=\prod_{i=1}^{m} p_g(x^i,\theta) \tag{2.2}$$

因此需要通过调整参数 θ^* 来最大化这个似然函数，即

$$\begin{aligned}\theta^* &= \arg\max_{\theta}\prod_{i=1}^{m} p_g(x^i,\theta)\\ &= \arg\max_{\theta}\log\prod_{i=1}^{m} p_g(x^i,\theta)\\ &= \arg\max_{\theta}\log\sum_{i=1}^{m} p_g(x^i,\theta)\\ &\approx \arg\max_{\theta} E_{x\sim p_{data}(x)}[\log p_g(x,\theta)]\end{aligned} \tag{2.3}$$

上式等价于

$$\arg\max_{\theta}\int_x p_{data}(x)\log p_g(x;\theta)dx-\int_x p_{data}(x)\log p_{data}(x)dx \tag{2.4}$$

后面多减的一项是一个常数，并不影响整体结果，目的是推导出 KL 散度表达式，上式经过化简之后为

$$\begin{aligned}&\arg\max_{\theta}\int_x p_{data}(x)\log\frac{p_g(x;\theta)}{p_{data}(x)}dx\\ &=\arg\min_{\theta} KL(p_{data}(x)\,\|\,p_g(x;\theta))\end{aligned} \tag{2.5}$$

因此可以看出寻找参数θ使似然函数最大化的过程也就是在最小化估计分布和真实分布之间的KL散度[38]。但是对于复杂分布，我们不能单纯地将其假设混合高斯模型或其他概率模型，因此无法给出估计分布的具体形式，也就无法计算似然函数。为解决这一问题，生成对抗网络放弃了估计分布的具体形式，而是通过神经网络去拟合这样一个分布。高斯分布在通过生成器G后，可以被变换成任意一个分布$p_G(x)$，因此希望$p_G(x)$和$p_{data}(x)$越接近越好。

$$G^* = \arg\min_G Div(p_G(x), p_{data}(x)) \tag{2.6}$$

如公式2.6所示，优化目标即为寻找一个生成器G使生成器拟合的分布$p_G(x)$与真实数据分布$p_{data}(x)$的某种散度最小化。然而$p_G(x)$和$p_{data}(x)$的具体形式并未给出，散度也无法通过计算得出，尽管如此，根据$p_G(x)$和$p_{data}(x)$采样却是可以做到的。真实数据样本可以看作从分布$p_{data}(x)$中采样得到的，生成器G生成的图像样本可以看作从$p_G(x)$采样得到的，因此生成对抗网络可以通过训练鉴别器D去“衡量”两种样本之间的差距并加以区分。

定义目标函数

$$V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_G(x)}[\log(1-D(x))] \tag{2.7}$$

训练鉴别器时，生成器是被固定住的，需要调整鉴别器参数使$V(D,G)$最大，当$x \sim p_{data}(x)$时，数据是从真实分布采样得到的，因此需要使其$\log D(x)$尽可能地大；而当$x \sim p_G(x)$时，数据是从生成器拟合的分布中采样得到，需要使其$\log D(x)$尽可能地小，也即使$\log(1-D(x))$尽可能地大。

$$\begin{aligned} V(D,G) &= E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_G(x)}[\log(1-D(x))] \\ &= \int_x p_{data}(x)\log D(x)dx + \int_x p_G(x)\log(1-D(x))dx \\ &= \int_x [p_{data}(x)\log D(x) + p_G(x)\log(1-D(x))]dx \end{aligned} \tag{2.8}$$

假设$D(x)$可以是任意函数，令$p_{data}(x)=m$，$p_G(x)=n$，$D(x)=D$，则被积函数可简化为

$$f(D) = m\log(D) + n\log(1-D) \tag{2.9}$$

对D求导数

$$\frac{df(D)}{dD} = \frac{m}{D} - \frac{n}{1-D} = 0$$

解得

$$D = \frac{m}{m+n} = \frac{p_{data}(x)}{p_{data}(x)+p_G(x)} \tag{2.10}$$

将式2.18代入式2.15中并整理

$$\begin{aligned} \max_D V(D,G) &= -2\log 2 \\ &+ \int_x p_{data}(x)\frac{p_{data}(x)}{(p_{data}(x)+p_G(x))/2}dx + \int_x p_G(x)\frac{p_G(x)}{(p_{data}(x)+p_G(x))/2}dx \\ &= -2\log 2 + KL(p_{data}(x)\,\|\,\frac{p_{data}(x)+p_G(x)}{2}) + KL(p_G(x)\,\|\,\frac{p_{data}(x)+p_G(x)}{2}) \end{aligned} \tag{2.11}$$

后两项的和实际上是 $p_{data}(x)$ 和 $p_G(x)$ 的 Jensen-Shannon 散度。由式(2.14)，生成器的优化目标是寻找一个生成器 G 使 $p_{data}(x)$ 和 $p_G(x)$ 之间的某种散度最小，因此生成对抗网络最终的目标函数为

$$G^* = \arg\min_G \max_D V(G,D) \tag{2.12}$$

这实际上是两个优化问题，首先固定生成器 G，优化内层的鉴别器 D，使真实样本的评分越大越好，假样本的评分越小越好，即尽可能的将真假样本区分开来。优化完成后将鉴别器 D 固定，再对生成器 G 进行优化，使生成的假样本评分越大越好，即尽可能的生成更好地图像使鉴别其难以区分，在实际训练中可以采用梯度下降法对其优化。生成对抗网络的强大之处在于无论真实数据的分布情况有多么复杂，它都能够自动学习其分布情况。因此它比变分自编码器有更好的生成效果。

2.3.3 生成对抗网络的改进：DCGAN 和 WGAN

Ian Goodfellow 提出的 GAN 在指导思想层面上是革命性的，但实际应用中却面临着训练稳定性差、生成样本缺乏多样等问题。Martin Arjovsky 在论文[39]中对 Ian Goodfellow 提出的两种损失函数做了详细的论证，从数学的角度解释了 GAN 难以训练的原因。简单来说，如果鉴别器的鉴别能力过于出众，会使生成器的梯度减弱甚至消失，造成无法继续训练；如果鉴别器的鉴别能力不足，则难以起到有效的对抗作用，因此在训练时需要严格控制生成器和鉴别器的训练程度，防止因生成器和鉴别器不同步造成的训练终止。

为解决生成对抗网络难以训练的问题，DCGAN[40]将在监督学习中常见的卷积神经网络（Convolutional Neural Network, CNN）应用到生成对抗网络，实现了较为稳定的训练。相比普通的生成对抗网络，DCGAN 在鉴别器部分使用带步长的卷积层代替池化层，在生成器部分用反卷积层（Deconvolution）代替上池化层，解决了特征图的上、下采样的问题。其次，在网络构建过程中加入了批次归一化（Batch Normalization, BN）[41]，除生成器的输出层和鉴别器的输入层外，均采用批次归一化对数据进行规范，这样做可以避免梯度消失问题并加快学习和收敛速度。最后，DCGAN 在鉴别器中采用 LeakyReLU[42,43]激活函数，在生成器中最后一层采用 tanh 激活函数，其他层采用 ReLU[44]激活函数。

DCGAN 的网络结构示意如图 2.5 所示，其输入为一个 100 维的向量，经过一个全连接层后被转化为 1024 个 4×4 的特征图，随后经过 4 个反卷积网络，生成一张 64×64 的三通道图像。鉴别器的结构与之相反，鉴别器从生成器或真实样本中接受一张图像，经过四层卷积神经网络和一个全连接层，最终输出为一个标量值，该标量值代表了鉴别器对输入图像的真伪判断。

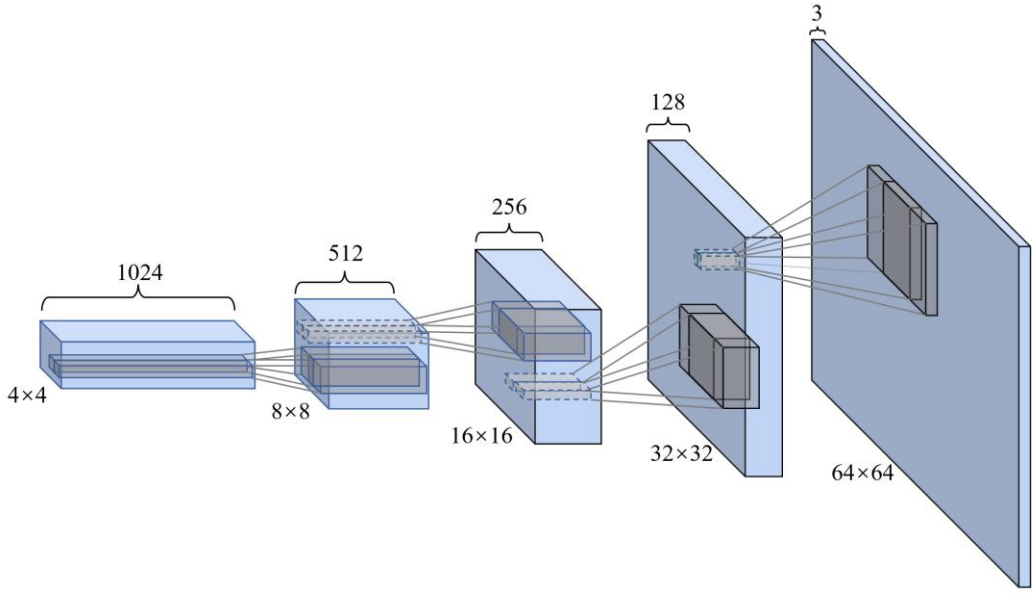

图 2.5 DCGAN 生成器结构示意图

DCGAN 生成器的损失函数为

$$L_G = \frac{1}{n}\sum_{i=1}^{n}[\ln(1 - D(G(z^i)))] \tag{2.13}$$

鉴别器的损失函数为

$$L_D = \frac{1}{n}\sum_{i=1}^{n}[\ln D(x^i) + \ln(1 - D(G(z^i)))] \tag{2.14}$$

DCGAN 主要通过网络结构的优化和激活函数的调整初步解决了传统生成对抗网络训练不稳定的难题，实际上并未从根本上解决问题，网络结构的改动仍有可能造成训练的崩溃。Martin Arjovsky 等人在数学理论分析的基础上，从生成对抗网络的损失函数入手，对生成对抗网络进行改，进提出了新的 WGAN[45]，经过改进后的生成对抗网络即使面对不使用卷积神经网络搭建的生成对抗网络也有良好的效果。

WGAN 中指出生成分布和数据的真实分布均处在高维空间，因此两个分布出现完全不重合的概率非大，而 JS 散度在衡量不重合的分布时存在着致命的缺陷，两个完全不重合的分布之间的 JS 散度时一个常数，这意味着梯度消失，参数无法得到正常更新，这也就解释了 GAN 训练稳定性差的原因。因此 WGAN 提出了用 Wasserstein 距离代替 JS 散度来衡量两个分布之间的距离，其定义如下：

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} E_{(x,y)\sim\gamma}[\|x - y\|] \tag{2.15}$$

其中 $\Pi(P,Q)$ 表示两个分布 P、Q 的所有可能的联合分布，$E_{(x,y)\sim\gamma}[\|x-y\|]$ 表示在联合分布 γ 下，使 P、Q 两个分布相等所付出的代价。在所有可能的联合分布中，该代价的下限值被定义为 Wasserstein 距离。Wasserstein 距离不仅能在两个分布没有任何交集的情况下正确反映

出两个分布之间的距离，还通过求解得到的联合概率分布给出了两个分布互相转化的具体形式。

WGAN 并没有对原始生成对抗网络的结构进行大的改变，而是根据理论推导的结果优化了损失函数，去掉了其中的对数项，优化后生成器的损失函数为

$$L_G = \frac{1}{n}\sum_{i=1}^{n} -D(G(z^i))) \tag{2.16}$$

鉴别器的损失函数为

$$L_D = \frac{1}{n}\sum_{i=1}^{n} [D(x^i) - D(G(z^i))] \tag{2.17}$$

2.4 高清图像生成模型

传统的生成对抗网络和一些经过改进的变体，如 DCGAN、LSGAN[46]、WGAN、WGAN-GP[47]等，经过训练能够较好的生成一些低分辨率的图像，但却难以应用在高分辨率图像的生成上。这主要是以下三点原因造成的，首先，在面对高分辨率的图像时，鉴别器很容易就能将真实图像和生成的图像区分开来，这使得梯度问题得以放大，并不能正确的指示优化方向。其次，生成器和鉴别器中卷积神经网络的参数随图像分辨率的提高，呈指数级增长，这对 GPU 显存的要求较高，在硬件的限制下，会使训练时的批次大小（batch size）下降，进而影响训练的稳定性。最后，生成图像的质量和多样性存在互斥关系，追求较高的分辨率可能会使图像的多样性降低。

2.4.1 ProGAN

为提高生成对抗网络生成图像的分辨率，StackGAN[48]将两个 CGAN[49]连接在一起，分为两个阶段进行训练。利用第一阶段训练生成的 64×64 的图像作为第二阶段训练的输入，成功将生成图像的分辨率提高到了 256×256。同年，英伟达实验室（NVLab）也提出了与之类似训练方法，被称作 Progressive GAN（ProGAN）[50]，其主要思想是在训练过程中逐渐加深生成器和鉴别器网络层数，以达到提高分辨率的目的。新增加的网络结构能比之前网络生成更精细的细节，使得生成图像最终的分辨率可以达到惊人的 1024×1024。ProGAN 的网络结构如图 2.6 所示，训练由 4×4 的分辨率开始，最初生成器和鉴别器只有一层结构，当训练达到纳什均衡时，鉴别器不能分辨出生成器生成的图片和真实图片之间的差距。此时将生成器和鉴别器同时增加一层，再对 8×8 的分辨率结构进行训练，以此类推直到分辨率达到 1024×1024。

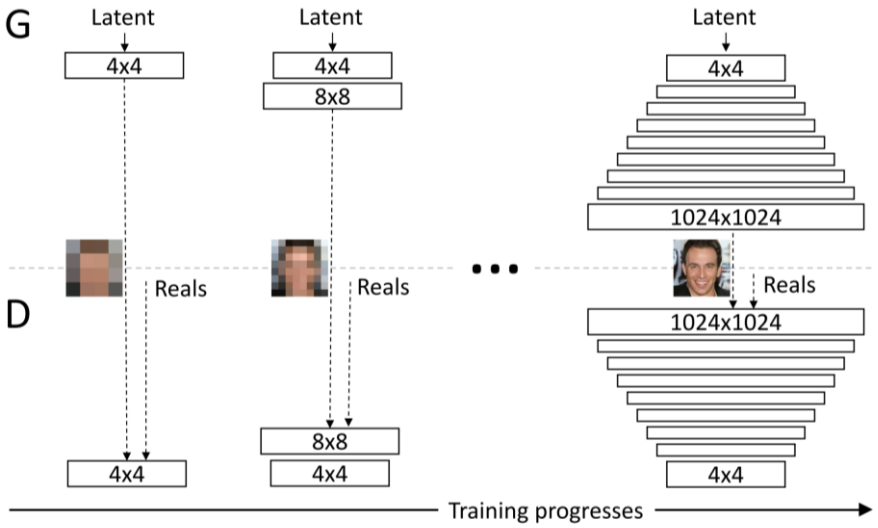

图 2.6 Progressive GAN 训练过程示意图[50]

新加入的网络层虽然使生成图像的分辨率提高了一倍，但由于新层的参数尚处在初始化阶段，直接训练会使得之前已经训练好的部分失效。为解决这一问题，ProGAN 提出了一种过渡模式，渐进的引入新层，使之成为完整的网络。

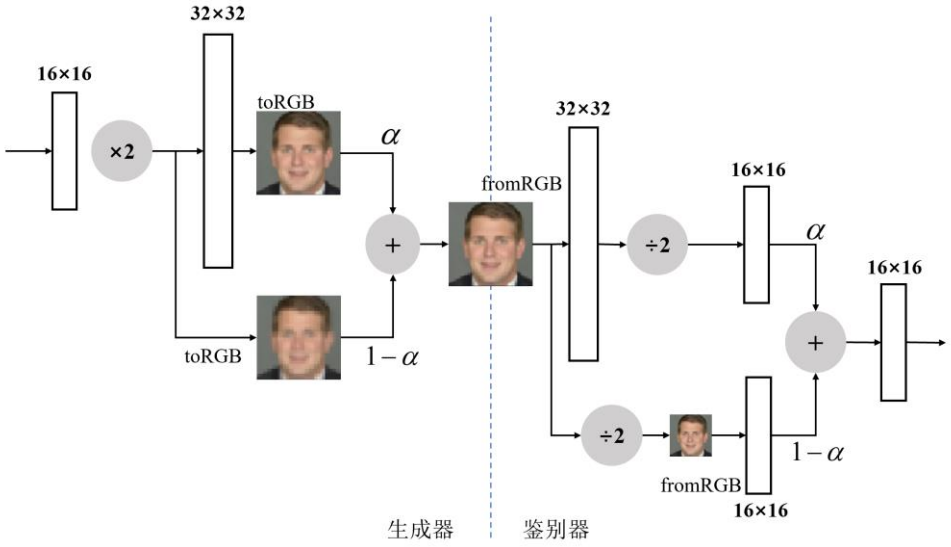

图 2.7 ProGAN 训练过程中逐渐引入新层示意图

ProGAN 提出的过渡模式采用了残差网络的思想，以分辨率由 16×16 增加到 32×32 的过程为例，如图 2.7 所示，在生成器部分，若干个(不同分辨率的特征图个数不同，16×16 分辨率的特征图为 512 个)16×16 的特征图经过插值上采样变为 32×32 的特征图，一方面，该特征图经由 toRGB 卷积层变为一张 32×32 的三通道图片，其本质上是将上一层网络结构生成 16×16 图像的简单放大，称作 residual；另一方面，该特征图被输入新层进行卷积运算，经过新层中的 toRGB 卷积层后也变为一张 32×32 的三通道图片（两个 toRGB 并不相同，实际上每个特定的分辨率都有自己的 toRGB 层），称作 straight。最终的输出结果为

$$\alpha \times straight + (1-\alpha) \times residual \tag{2.18}$$

其中α为系数，由 0 逐渐线性增长到 1，这是因为网络结构增长之初，增长的部分仅处于初始化状态，straight 和真实图片之间的差异很大，直接训练会影响已经训练好的网络，因此需要通过系数将新增的网络逐渐融入，这一想法不但使得之前训练模型得到了利用，也使得训练过程更加平稳。鉴别器的结构与生成器的结构对称，其主要将生成图片或真实图片转换为特征图，再逐层降维，最终得出一个标量，作为生成图片或真实图片的评分。随着训练的进行，鉴别器对生成图片和真实图片的评分逐渐接近，直到区分不出二者的差异，此时在该分辨率上，生成器生成的图片就足以达到以假乱真的效果。由于训练过程是循序渐进的，低分辨率训练阶段非常快，大大节省了整体的训练时间，效果也比直接生成 1024×1024 分辨率的图片要稳定。

2.4.2 StyleGAN

ProGAN 的诞生可以说是一个里程碑式的跨越，理论上只要训练样本满足条件，可以训练出任意分辨率的图像生成器模型，为生成对抗网络之后的发展奠定了基础。ProGAN 的输入是一个服从高斯分布的潜在向量（latent vector）z，大小为 512×1×1，虽然当对潜在向量 z 进行平滑插值操作时，生成器生成的图像的变化也比较平滑，但是 ProGAN 对生成图像特定特征的控制能力非常有限，这些属性相互纠缠，即使略微调整输入，也会同时影响生成图像的多个属性。因此想从潜在向量中获得控制人脸高级属性的向量仍然比较困难，生成器继续充当黑匣子的作用，缺乏对图像合成过程各个方面的理解。在这个背景下，StyleGAN[51] 应运而生，比较完美的解决了这一问题。

和 ProGAN 一样，StyleGAN 也是英伟达实验室的作品，StyleGAN 继承了 ProGAN 的大部分思想，又在其基础上做了进一步的改进和提升。

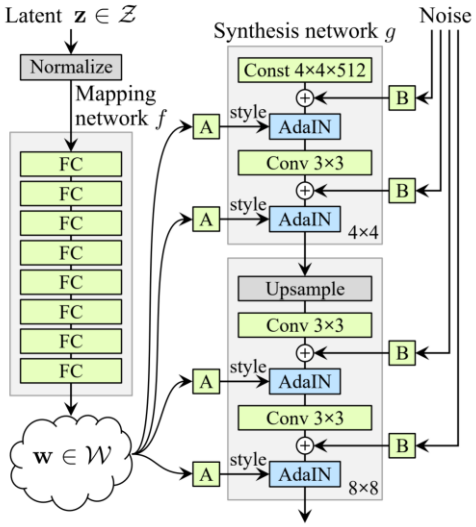

图 2.8 StyleGAN 生成器网络结构示意图

如图 2.8 所示，StyleGAN 的生成器部分主要由映射网络（Mapping Network）和合成网络（Synthesis Network）构成。映射网络由 8 个全连接层构成，主要负责进行分布变换，将服

从正态分布的潜在向量 z 变换为中间潜在向量 w，分布的变换模式在训练过程中自动学习，无需人为干涉。此举的目的是得到一个由线性子空间构成的中间潜在空间，即向量 w 的分布无需跟随训练数据分布，并且可以减少特征之间的相关性，消除部分属性的纠缠。在合成网络部分，相比于 ProGAN 直接将潜在向量 z 作为合成网络的输入，StyleGAN 采用一个可学习的常量作为合成网络的输入，其大小为 512×4×4，而将潜在向量 z 作为映射网络的输入。经过解纠缠的中间潜在向量 w 被当作控制图像高级特征的参数输入到合成网络的每一层中。由于每一层的分辨率不同，因此所控制的视觉特征也不尽相同，一个较为合理的解释是，在低分辨率的层中，中间潜在向量 w 影响粗糙的特征，如面部形状、朝向、五官位置等。在分辨率中等的层，中间潜在向量 w 影响一些较为细致的特征，如发型、面部特征，五官的形状等。在高分辨的层，中间潜在向量 w 影响一些微观的细节特征，如表情（五官细节）、肤质毛发细节等。在中间潜在向量 w 输入合成网络之前，还需要对其进行一次可学习的仿射变换，如图 2.9 所示。

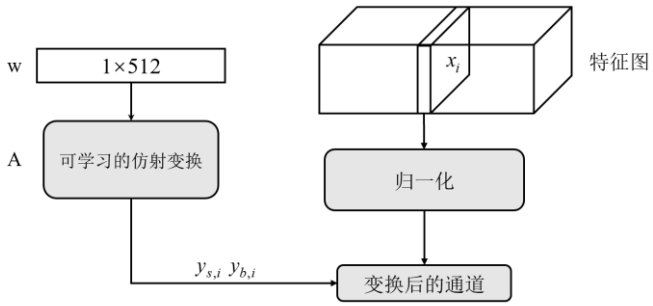

图 2.9 自适应实例归一化示意图

StyleGAN 在此处借鉴了风格迁移中自适应实例归一化[52]思想，将不同层次的样式（style）嵌入不同分辨率的特征图上。首先，经过仿射变换（本质是一个全连接层）将 1×512 中间潜在向量 w 转换为 $1\times 2n$ 的样式 y，其中 n 为特征图的个数，$y=(y_{s,i},y_{b,i})$，$y_{s,i}$ 为缩放因子，$y_{b,i}$ 为平移因子。然后将卷积后的特征图进行自适应实例归一化

$$\frac{x_i-\mu(x_i)}{\sigma(x_i)} \tag{2.19}$$

自适应实例归一化在风格迁移和生成对抗网络任务中的作用被认为是远好于批次标准化（BatchNormalization），在基于样式生成单张图片任务的过程中也有良好的表现。AdaIN 的公式为

$$AdaIN(x_i,y)=y_{s,i}\frac{x_i-\mu(x_i)}{\sigma(x_i)}+y_{b,i} \tag{2.20}$$

利用 AdaIN 对每个特征图进行缩放平移变换，就将中间潜在向量 w 学习到的样式信息融入到了图像生成过程的中间层。生成图像从 4×4 分辨率增长到 1024×1024 分辨率，需要经过 9 层卷积神经网络结构，每层网络进行两次 AdaIN 操作，可以很好地对所生成图像的高级特征进行控制。除此之外，StyleGAN 还在每层网络中加入了随机噪声，其位置在卷积之后

AdaIN 操作之前。引入随机噪声可以为生成器生成的图片添加一些随机细节，例如脸上的雀斑、头发的准确位置等，这些随机细节并不改变人脸的辨识度，但是增加了输出的多样性，使图像更加真实自然。通常情况下，控制噪声的效果并不是一件简单的事情，由于特征纠缠现象的存在，噪声的加入会对图像的其他特征造成干扰。StyleGAN 通过在合成网络中每个 AdaIN 操作前增加一个尺度变换模块 B，巧妙地避免了这一问题。原始噪声是一个由高斯噪声组成的单通道图像，尺度变换模块 B 使用可学习的缩放参数对输入的高斯噪声进行变换，然后将变换后的噪声图像广播到所有的特征图中（每个特征图分别对应一个可学习的参数，因此每个特征图对应的噪声并不相同）。由于噪声经过了尺度变换，且随后又经过了 AdaIN 的自适应实例归一化变换，因此对图像的影响远不如样式对图像的影响，即实现了图像随机多样性，又不影响图像整体特征。

StyleGAN 的鉴别器部分的结构大致和 ProGAN 类似，只是在每个分辨率层内部略有不同，是为了和生成器的结构保持对称，以便达到较好的鉴别效果，完整的网络结构如图 2.10 所示

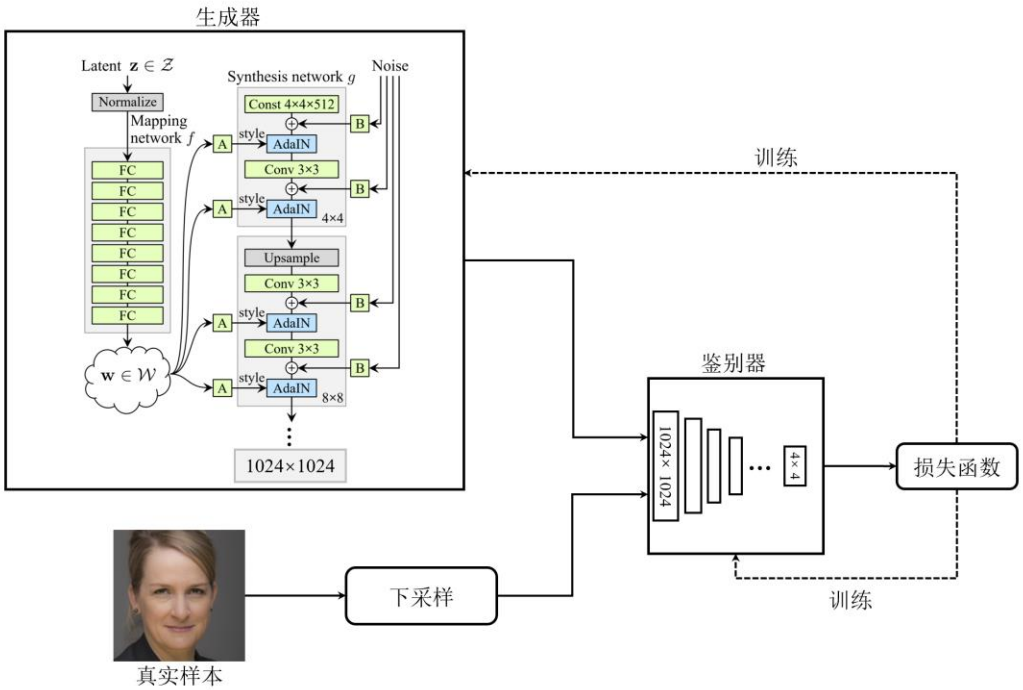

图 2.10 StyleGAN 完整结构示意图

真实样本在输入鉴别器之前要进行分辨率缩放，以适应生成器和鉴别器的层数变化。训练过程中，鉴别器尽可能的区分生成图片和真实图片，生成器则生成更好的图片来欺骗鉴别器，二者对抗训练。StyleGAN 的损失函数并没有采用 ProGAN 中使用到的 WGAN-GP 损失函数，而是采用了含简单梯度惩罚的 Logistic 损失，形式如下：

$$Loss_G = \log(e^{-D(G(z))} + 1) \tag{2.21}$$

$$Loss_D = \log(e^{-D(x)} + 1) + \log(e^{D(G(z))} + 1) + \gamma_1 \times 0.5 \times \sum \nabla^2_{T_{real}} + \gamma_2 \times 0.5 \times \sum \nabla^2_{T_{fake}} \tag{2.22}$$

使用梯度惩罚时，无论是来自真实分布的样本还是来自生成分布的样本，梯度绝对值过高时均会受到惩罚，在 StyleGAN 默认保留了对真实样本的梯度惩罚($\gamma_1=10$)，而舍弃了对生成样本的梯度惩罚($\gamma_2=10$)，这样做可以加快生成器的收敛速度。但是使用该损失函数的一个缺点在于它并不保证一定能够达到纳什均衡，因此对超参数的设置比较严格。

训练好之后的模型可以生成不同分辨率的人脸图像，如图 2.11 所示，可以看出，当分辨率达到 128×128 时，图像的细节已经较为完善，人眼几乎无法识别出这是一张由生成对抗网络生成的图片。尽管如此，StyleGAN 也并不是无可挑剔，某些情况下生成的图像仍然会出现无法预估的滴液状伪影。

图 2.11 利用 StyleGAN 训练生成的不同分辨率的人脸图像

经过研究发现，滴液状伪影的出现是生成器为了规避网络结构中的设计缺陷，在随后英伟达推出的 StyleGAN2[53]中重新设计了生成器的网络结构，修复了这一问题。相比之前 StyleGAN 的网络结构，修改后的网络移除了一些冗余的操作，将偏置和噪声尺度变换操作放到了样式操作区域之外，并只调节每个特征图的标准差。经过修改后的网络可以通过一个名叫“解调”的操作代替原先的自适应实例归一化操作。样式调制由原来的通过对特征图的缩放、平移实现变为通过对卷积权重的缩放实现：

$$w_{ijk}' = s_i \cdot w_{ijk} \tag{2.23}$$

其中 w_{ijk} 为原始权重，w_{ijk}' 为调制后的权重，s_i 为第 i 个特征图对应的缩放系数。

StyleGAN2 中另一个比较大的改变是其放弃了渐进增长的网络结构，虽然渐进增长式的网络结构在合成高分辨率图像任务中被认为非常成功，但这种结构也是伪影出现主要原因。渐进增长的生成器对细节方面有着强烈的位置偏好，例如当人脸在平滑变化时，眼睛或牙齿等细节并不会随着人脸的变化而进行相应的变换，而是停留在原来的位置，当变化累积到一定程度时，这些细节会突然出现在下一个位置。英伟达实验室认为，在渐进增长网络结构中，每一层网络结构都会暂时充当输出分辨率的角色，这会迫使其生成最大频率细节，从而导致训练后的网络在中间层具有过高的频率，从而损害了位移不变性。

2.5 强化学习简介

强化学习（Reinforcement Learning），根据翻译的差异，有时也被称作增强学习，是和监督学习、无监督学习并列的机器学习方法，也是最基本的学习方法。与传统的监督学习和无监督学习不同，强化学习特别强调环境的重要性，主张从与环境的交互过程中学习，因此它不需要带标签的输入，同时也不需要过于追求精确的最优解，强化学习的关注点在于对未知

领域的探索及已有知识利用之间的平衡[54]。

强化学习并不是新出现的机器学习方法，而是历经了几十年的发展，直到 AlphaGo 战胜围棋世界冠军李世石九段，才逐渐被人们所熟知。关于强化学习的研究，最早可以追溯到经典条件反射实验，随后演化出优化控制和动物行为研究两个分支，其中优化控制分支被 Bellman 抽象为马尔可夫决策过程（Markov Decision Process，MDP）。强化学习是所有机器学习中和动物学习甚至人类学习最接近的一种，即生物在外界环境对其行为的奖励或惩罚下逐渐适应并学习，做出能获得最大利益的习惯性行为。强化学习由于其通用性，被广泛应用于控制论、博弈论、仿真优化、多智能体系统等领域。

2.5.1 一般模型

一个完整的强化学习模型主要由智能体（Agent），环境（Environment）、动作（Action）和奖励（Reward）四部分构成，如图 2.12 所示

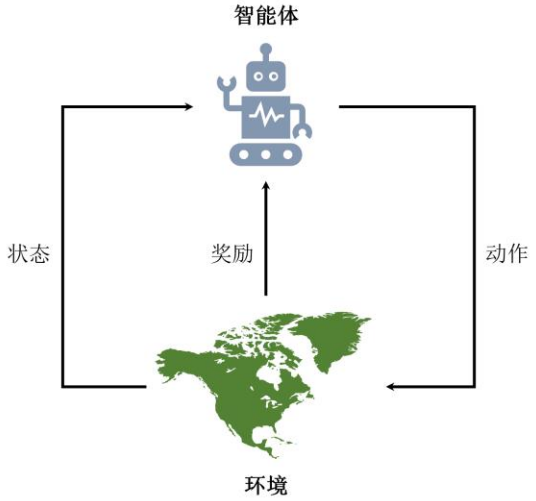

图 2.12 强化学习一般模型框架示意图

智能体在零时刻根据观察到的环境状态（State）S_0 做出一个动作 A_0，环境接受该动作后状态由 S_0 变为 S_1，同时给智能体一个反馈，也就是奖励 R_1，至此智能体与环境之间完成了一轮交互，然后智能体会继续根据新观察到的环境状态及获得的奖励情况进行后续的交互，得到一系列 $\{S_0, A_0, R_1, S_1, A_1, R_2, ..., S_t, A_t, R_t, S_{t+1}\}$ 集合，直到完成某种任务或达到交互上限。注意到，奖励并不是在动作做出后就立刻获得的，而是延时奖励，即 t 时刻做出动作的奖励在 $t+1$ 时刻获得。

将上述强化学习的简单建模进一步展开，引入策略（policy）这一概念，策略是强化学习智能体做出何种动作的依据，也是智能体需要优化的目标，只有好的策略才能做出收益最大的动作选择，策略一般由一个条件概率分布表示

$$\pi(a \mid s) = P(A_t = a \mid S_t = s) \tag{2.24}$$

即在状态 s 下采取动作 a 的概率，除此之外，环境在接收到智能体执行的动作后也要发生相

应的改变，为简化环境模型表示，假设环境状态的转化具有马尔可夫性，即环境转化到下一状态的概率仅与当前环境状态有关，

$$P_{ss'}^{a}=P(S_{t+1}=s'|S_t=s,A_t=a) \tag{2.25}$$

智能体在执行动作后会获得一个延时奖励，但该奖励仅能作为智能体上一个动作好坏的评价。强化学习并不一味的追求当下的奖励，例如在下棋时，吃掉对方一个棋子可能会获得一个较大的奖励，但这有可能是对方设下的陷阱，会导致最终输掉整盘棋。因此更有价值的是状态价值，它表示智能体从某一状态开始所能获得的期望奖励，智能体在状态 s，遵循策略 π 做出动作选择，其状态价值为

$$v_\pi(s)=E_\pi(R_{t+1}+\gamma R_2+\gamma^2 R_{t+3}+...|S_t=s) \tag{2.26}$$

其中 γ 为奖励衰减因子，当 $\gamma=1$ 时代表当前延时奖励与后续延时奖励同等重要；当 $0<\gamma<1$ 时代表当前延时奖励比后续延时奖励更重要；当 $\gamma=0$ 时则完全忽视后续延时奖励，此时状态价值与当前延时奖励相等。通过推导可以发现，状态价值函数满足递推关系，即

$$v_\pi(s)=E_\pi(R_{t+1}+\gamma v_\pi(S_{t+1})|S_t=s) \tag{2.27}$$

上式也被称作贝尔曼方程，除此之外，定义动作价值函数 $q_\pi(s,a)$，它表示在状态 s 开始，选择动作 a，并根据策略 π 进行后续交互所能获得的期望奖励，

$$q_\pi(s,a)=E_\pi(R_{t+1}+\gamma R_2+\gamma^2 R_{t+3}+...|S_t=s,A_t=a) \tag{2.28}$$

同理，动作价值函数也满足递推关系，并且状态价值和动作价值之间存在关系

$$v_\pi(s)=\sum_{a\in A}\pi(a|s)q_\pi(s,a) \tag{2.29}$$

$$q_\pi(s,a)=R_s^a+\gamma\sum_{s'\in S}P_{ss'}^{a}v_\pi(s') \tag{2.30}$$

对强化学习完成建模后，其求解目标就是寻找一个最佳策略使智能体在与环境交互过程中获得最大的收益。

2.5.2 常见方法

根据智能体是否对环境进行建模，强化学习问题可以分为基于模型的（Model based）和无模型的（Model-free）。对于基于模型的强化学习问题，马尔可夫决策过程的转移概率是已知的，可以用动态规划进行求解，常见的方法有值函数迭代方法和策略迭代方法。值函数迭代方法在当前状态 s 对所有可能的动作 a 都计算下一状态的期望价值，并选择最大的期望价值更新当前状态的价值函数，直到达到价值函数收敛。策略迭代则相对复杂一些，分为策略评估和策略提升两部分，即从随即策略开始，计算该策略下的状态价值函数，再根据得到的状态价值函数更新策略，直至策略收敛。策略迭代方法在收敛速度上更快，但是每次更新策略时都需要对策略进行评估，需要进行大量的计算，因此不适合用于状态空间小的问题。

但是多数情况下，智能体对环境模型并不清楚。这种情况下，可以采用蒙特卡洛方法和时间差分学习的方法进行求解。蒙特卡洛方法直接从完整的交互过程中进行学习，并将状态的平均收益作为该状态值函数的估计，但一般蒙特卡洛方法把动作价值函数作为优化目标，而非状态价值函数。和蒙特卡洛法相比，时间差分学习则不需要等到交互回合结束再去更新，它可以从不完整的交互序列中进行学习，利用延时奖励和下一状态的值函数近似当前状态的收益，只需要两个状态就可以对状态值函数进行更新。Q-learning[55]和 SARSA[56][57]是时间差分学习方法中两个经典的算法，二者均基于 Q-table，不同的是，Q-learning 再进入下一状态后并不真正执行下一个动作，而是在所有可能的动作中选取动作价值函数最大的 Q 值进行更新。

实际问题中，状态空间往往很大，难以用 Q-Table 存储所有动作价值函数，DeepMind 将 Q-learning 与神经网络相结合，提出了 DQN[58]算法。DQN 利用深度神经网络估计动作价值函数，并基于 Q-learning 构造损失函数，利用梯度下降法对网络参数更新，实现最佳策略的求解。

上述算法均是通过状态价值函数或动作价值函数对策略进行调整优化的，也叫基于值函数（Value-Based）的方法，除此之外，强化学习中还有一类直接对策略进行调整优化的方法，基于策略的（Policy-Based）方法[59]。相比基于值函数的方法，基于策略的强化学习方法能够处理连续动作空间下的强化学习问题，并且能够学习随机策略。策略梯度的主要思想是调整策略使智能体收益最大化，因此策略被表示为带有参数的函数，学习最优策略的过程就是调整参数使收益最大化的过程，该过程可以用梯度上升法进行求解。在基于值函数和基于策略梯度方法的基础上，Actor-Critic 架构对二者的优点进行了整合，形成了目前主流的深度强化学习算法。

2.6 本章小结

本章主要对生成模型和强化学习相关的基础知识进行了介绍。首先从简单的自编码器切入，简单介绍了变分自编码器的原理及它作为真正意义上的生成模型和普通自编码器的差异。其次对生成对抗网络作了全面详细的介绍，包括基础模型和发展过程中比较重要的改进模型，体现了它们在图像生成领域的重要性。最后针对本文用到的高分辨率图像生成模型进行了深入的探讨和分析，详细阐述了模型的演化过程和设计思想。最后对强化学习的基础知识和常见的算法进行了简要的概述，为本文三四章利用强化学习解决人脸属性编辑问题做了理论上的铺垫。

第三章 基于深度确定性策略梯度的人脸老化编辑

人脸属性编辑一直以来都是计算机视觉领域研究热点，年龄作为一种具有现实意义的复合属性，在跨时间案件侦破中可以发挥巨大的作用，因此面部衰老[60-62]和具有商业价值的面部妆容[63-65]一同成为独立的研究分支。与其他的单一属性不同，年龄的编辑必须具有连续性，一些基于模型和基于附加条件的人脸属性编辑方法在年龄编辑上具有很大的弱点。除此之外，年龄是一种复合属性，由头发颜色、面部皱纹、皮肤光泽等属性共同构成，现有的方法在保持身份特征信息上效果不佳。

为解决上述缺点，本文提出了一种基于深度确定性策略梯度的年龄编辑方法，能够在实现年龄连续编辑的情况下较好的保持被编辑人的身份特征。在本章中，我们首先对人脸年龄编辑的建模思路进行了详细的阐述，针对要解决的问题进行了反复推敲并最终确定和方法的选择，最后对实验结果进行了详细的分析。

3.1 建模思路及流程

生成对抗网络的诞生使得生成清晰图像的成为可能，也是几乎所有人脸属性编辑模型的基础。由生成对抗网络生成图像的过程可以看作由一个低维分布到高维分布的变换。在低维分布中取一点 z，将其变换到高维空间中，意味着我们得到了一张图像 $G(z)$，根据分布的连续性，当 A 邻域中的点被变换到高维空间中时，也应该与 $G(z)$离得很近。此外有大量研究[40,51,66]表明，当对潜在空间中的两点 z_1、z_2 进行线性插值时，其生成的图像也是连续变化的。在 CAAE[16]中，作者假设人脸图像处在高维流形上，通过在该流形滑动，可以实现某种属性的连续变化。受此启发，结合生成对抗网络在图像生成上的连续性，我们可以假设人脸图像在高维分布中运动时可以实现某种属性的连续变化。因此，只要找出代表年龄的运动轨迹，就可以完成对年龄的编辑，如图 3.1 所示。

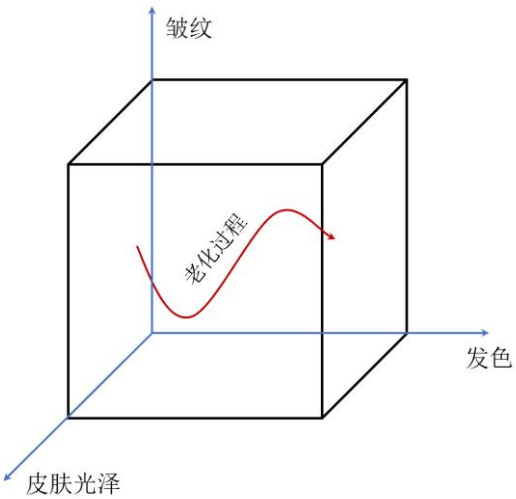

图 3.1 三维空间中人脸老化过程示意图

然而高维空间对人类来说是不可想象的，直接在像素点上操作不但容易损坏图像，而且操作难度极大，甚至还不如 Photoshop 来得直接，显然是不可取的。但这是一个特殊的高维空间，是生成对抗网络的生成空间，它和潜在空间在某种程度上是对应的。潜在空间中的点一般是一个 $n\times1$ 的向量，维度比图像低几个数量级，因此我们可以在潜在空间中寻找这样一条路径，实现人脸年龄的编辑。

寻找路径的过程，实际上是修改潜在向量的过程，换言之就是在原始潜在向量上添加一个微小增量形成新的潜在向量，通过观察两个潜在向量生成的图像，确定是否实现了正确的编辑。这里存在一个问题，这一微小增量根据什么去产生以及如何修改这一微小增量使生成图像达到正确的编辑。根据传统思路，可以利用神经网络去生成这样一个微小增量，并设计一个关于年龄的损失函数对其求解，通过反向传播逐渐优化生成的微小增量。但是这可能会导致生成网络参数的改变，即年龄的变化是生成器网络被优化的结果，与生成微小增量的网络无关，或者与二者皆有关，这两种情况实际上都破坏了生成器的结构，会导致生成对抗网络崩溃。即便将生成器部分固定，损失函数也只能保证在年龄上有梯度，随着微小增量的累积，或许能够实现年龄属性的编辑，但人脸的样貌必然发生大幅度的改变。理论上，设计一个除了年龄之外其他属性不变的损失函数几乎不可能。

3.1.1 强化学习模型的建立

通过观察可以发现，寻找衰老路径的过程和强化学习中的经典问题悬崖寻路问题比较类似。悬崖寻路问题，如图 3.2 所示，是在二维网格寻找一条能以最短的步数从起点到终点的问题。智能体在网格中可以朝上、下、左、右四个方向进行移动，每次只能移动一步，但不能超出网格边界。智能体每走一步都要受到一个较小的惩罚，如果进入红色部分（悬崖）则受到很大的惩罚并立即回到出发点，当到达终点时，则回合结束。

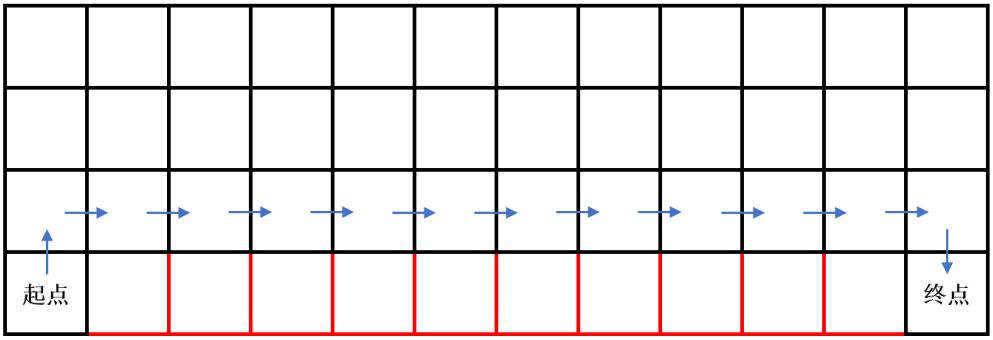

图 3.2 悬崖寻路问题示意图

根据类比，我们可以将起始人脸图像看成寻路问题的起点，将完成老化后的图像看成寻路问题的终点。很显然，人脸图像的变化过程就是强化学习中环境的变化过程，人脸的生成分布构成了强化学习的状态集合。完成上述类比后，我们需要引入一个智能体，智能体会根据观察到的环境状态做出相应的动作。很自然的，智能体的动作即是前面提到过的微小增量，

环境在接受智能体的动作后，环境状态会做出相应的改变，并给出延时奖励。至此，针对人脸年龄编辑的强化学习模型初步完成，其流程图如图 3.3 所示。

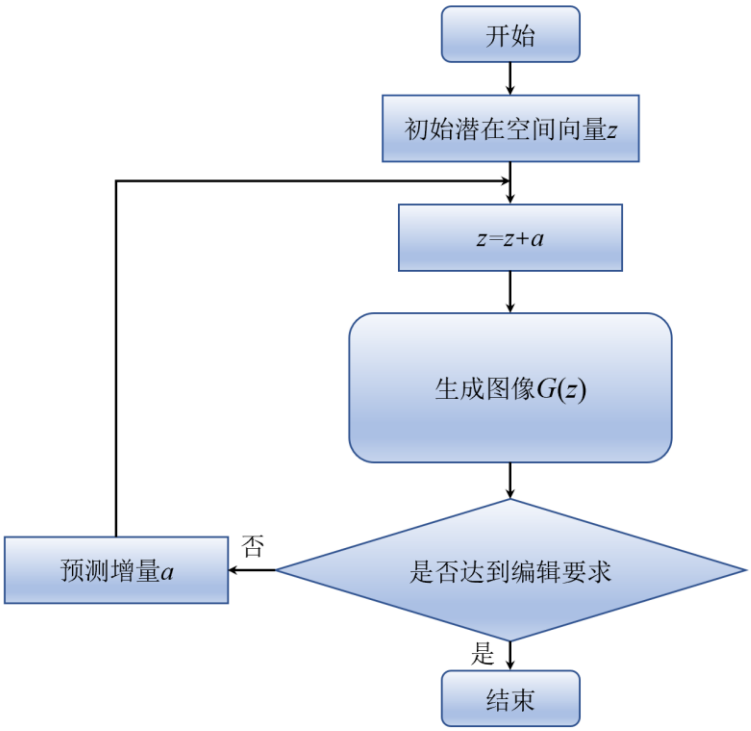

图 3.3 人脸年龄编辑建模流程图

在图像生成器部分，普通的生成对抗网络模型会有图像分辨率的限制，为了突破分辨率的限制，实现高清人脸图像的编辑，本文选用了 StyleGAN2 模型。该模型以 512×1 维的高斯噪声向量 z 作为输入，与传统生成对抗网络不同，向量 z 需要经过一个可学习的映射网络转换为中间潜在向量 w。映射网络由八个全连接层组成，目的是得到一个由线性子空间构成的中间潜在空间 W。因此可以认为，中间潜在向量的每一个分量都能单独控制图像的一个特征，但这并不代表年龄属性是由其中的一个分量控制的，因为年龄是复合属性，由多个特征共同控制，即使特征之间的纠缠消除了，属性之间的纠缠仍然可能存在。因此中间潜在向量和人脸属性之间并不是简单的线性关系。

潜在向量 z 经过映射变换，最终由中间潜在向量 w 控制人脸图像的生成。但由于 z，w，$G(w)$是依次映射的关系，三者中的任意一个都可以代表强化学习中的状态，因此为了便于智能体对年龄路径的探索，由线性子空间构成的中间潜在向量更适合状态的表示。对于给定的初始潜在向量 z_0，其中间潜在向量为 w_0，用表示 w_0 初始人脸状态，w_n 表示年龄编辑完成后的人脸状态，则

$$w_n = w_0 + d \tag{3.1}$$

其中 d 为年龄路径向量，因此年龄编辑的求解目标是求解一个最佳的 d，使生成图像

$G(w_0+d)$ 和 $G(w_0)$ 仅在年龄上发生变化，而其他属性基本不变，以保持良好的身份特质。引入属性评价器 b，目标函数可表示为

$$f(d)=\left|b_{i=Age}[G(w_0+d)]-b_{i=Age}[G(w_0)]\right|-\alpha\sum_{i\neq Age}\left|b_i[G(w_0+d)]-b_i[G(w_0)]\right| \tag{3.2}$$

其中 α 为系数因子，用来均衡年龄和其他属性之间的权重。因此最优年龄路径向量

$$d=\arg\max_{d} f(d) \tag{3.3}$$

在 InterFaceGAN 中，并没有路径的概念，而是以年龄方向向量代替，并假设年龄是随方向向量线性变化的，在潜在空间中进行大量采样，通过求解二分类面的法向量的方式寻找年龄的方向向量。这种方法可以基本实现年龄的编辑，但是受生成模型和人脸采样的影响，人脸图像在老化后逐渐偏向欧美人种，身份特质保持性比较差。

根据本文提出的强化学习模型，年龄路径 d 实际上由很多个被视作动作的微小增量组合而成，为防止微小增量的大小不一致，需要将其单位化，用 a 表示，a 的欧式距离可以取 1，也可以是指定的长度。

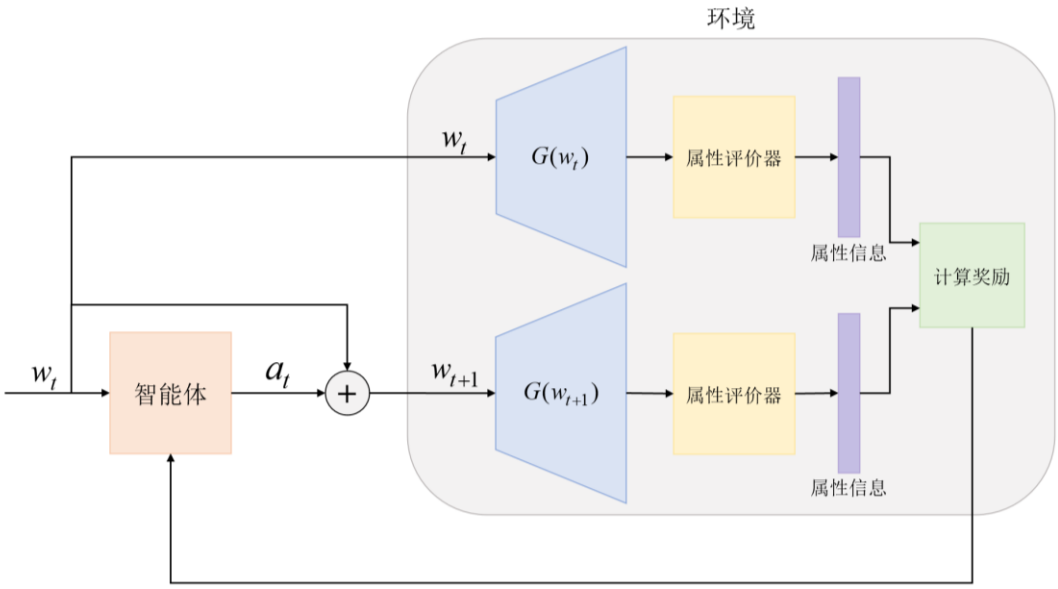

图 3.4 智能体与环境交互过程示意图

智能体与环境的交互过程如图所示，在 t 时刻，智能体根据某种策略和当前状态 w_t 选择一个动作

$$a_t=\pi(w_t\mid\theta) \tag{3.4}$$

其中 θ 为智能体的参数，环境接收动作后，将动作与上一状态叠加生成新的状态

$$w_{t+1}=w_t+a_t \tag{3.5}$$

随后状态 w_t 和 w_{t+1} 被输入同一个经过预训练的生成器得到两张人脸图像，再根据两张图像在年龄上的差异计算奖励函数并反馈给智能体。通过合理的设置奖励函数，智能体就可以在

与环境的交互过程中学习到一系列的动作集合，该集合既是人脸图像老化的路径。

3.1.2 奖励函数的设置

利用强化学习解决实际问题的核心是如何合理的设置奖励函数。在强化学习中，智能体有且仅有一个目标，那就是最大化所有收到的奖励总和。智能体做出的每一个决策都是为了实现“利益”的最大化，如果不能把任务目标融入到奖励函数中，则意味着该问题不适合用强化学习求解。对于奖励函数的设置，本文遵从了一般强化学习问题中奖励函数的设置习惯，主要分为主线奖励、目标达成奖励、交互惩罚和溢出惩罚。

对于年龄编辑来说，无论是人脸老化还是人脸年轻化，都是具有方向性的。以人脸老化为例，假设初始人脸的年龄为 25 岁，目标人脸的年龄为 65 岁，智能体的任务目标就是从 20 岁的状态逐步移动到 60 岁的状态。因此主线奖励是年龄差，即当智能体移动一步后，如果新的人脸图像年龄比上一步的人脸图像年龄大，就得到正的奖励，反之，则收到负的奖励，奖励数值为年龄差的大小。目标达成奖励是智能体在达成目标时给出的额外奖励，一般情况下可以不设置，但为了避免智能体在前期盲目探索，便于智能体确定大致的探索方向，本文将其设置为 50，即 0 岁到 100 岁的一半。

交互惩罚是强化学习中比较常见的一种机制，即在每次与环境交互后受到的固定惩罚，该项存在的意义是促使智能体提高效率，防止其做无用功。例如，当仅存在主线奖励时，由于最终获得的累计奖励是相同的，原本一步可以完成的目标，有的智能体可能要走十步才能完成。此外，交互惩罚的加入会促使智能体以最少的步数到达目标，有利于保持身份特质不变。如图 3-5 所示，在二维空间中，假设某一时刻红色箭头所指向的方向代表年龄的方向，与之垂直的方向为其他属性的方向。智能体在该时刻做出的任何其他方向的动作选择均无法以最短的步数获得相同的主线奖励，反映在图像上则表现为人脸的其他属性发生了改变。

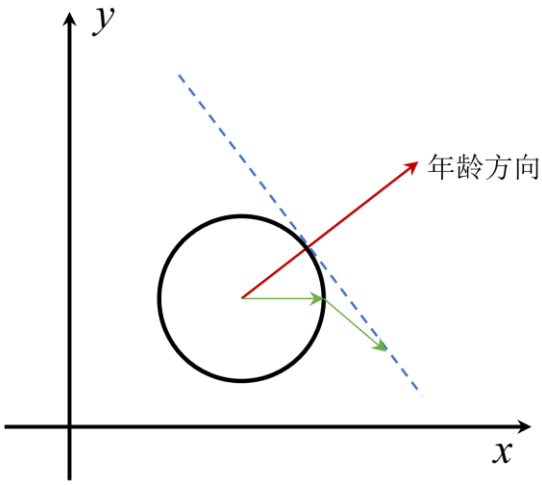

图 3.5 获得相同奖励不同动作所需的步数

由于智能体是根据选择的动作在中间潜在空间进行探索，会存在探索出界的情况，即经

过修改的中间潜在向量已经不属于生成分布，此时生成器生成的图像会出现模糊变形，可以用鉴别器进行鉴别。这种情况下应当给予智能体较大的惩罚，并立即结束此次探索。

3.2 年龄预测网络

主线奖励的计算过程中，需要利用到两张生成图像种人脸的年龄。由于生成数据并没有真实标签，因此需要对其进行预测。本文利用 ResNeXt-50 网络搭建了一个年龄预测模型，为主线奖励的计算提供支持。

3.2.1 网络结构

ResNeXt[67]同时融合了 ResNet[68]和 Inception[69]的思想，能够在参数复杂度不变的情况下提高分类问题的准确性，其网络单元结构如图 3.6 所示，从外部看，ResNeXt 仍是残差结构，但是内部被分成了 32 条相同结构的计算通道，每条计算通道由三层卷积神经网络构成，卷积核大小分别为 1×1、3×3、1×1。

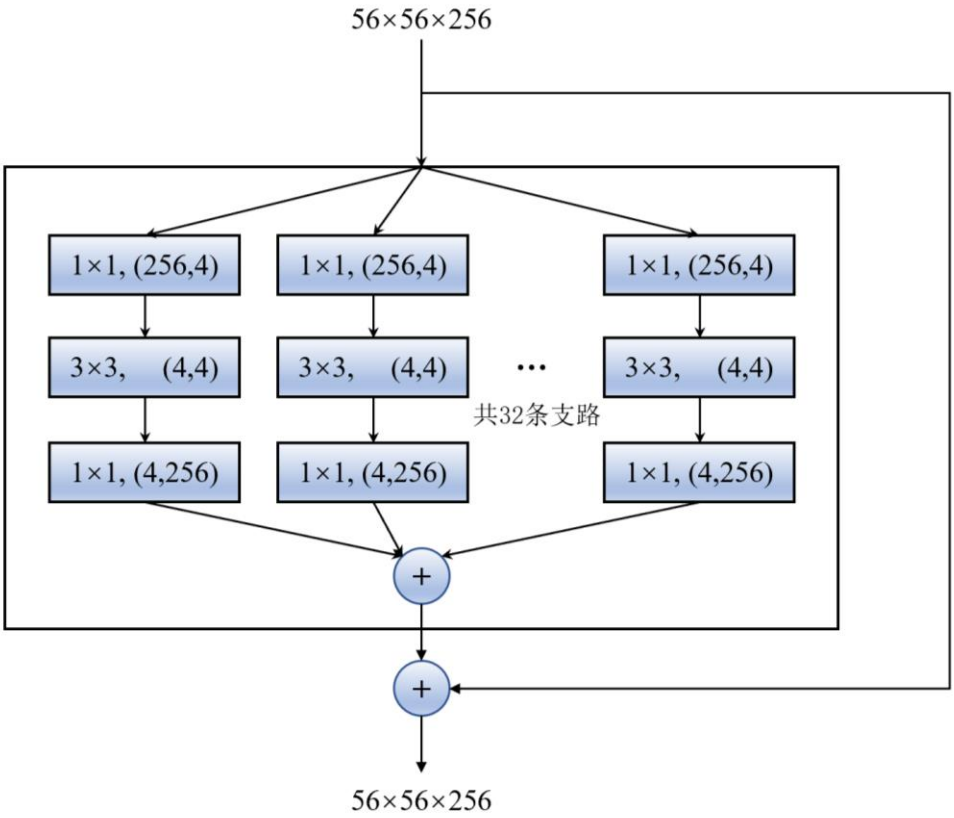

图 3.6 ResNeXt-50(32×4d)残差单元示意图

相比 ResNet 和 Inception 网络分别通过加深、加宽网络来提高网络的分类性能，ResNeXt 通过增加通道数可以在保证参数基本不发生改变的情况下，更有效地提高网络性能。ResNeXt-50（32×4d）网络参数如图 3.7 所示，

ResNeXt-50 (32×4d)
7×7, 64, stride 2
3×3 max pool, stride 2
[1×1, 128 3×3, 128, C=32 1×1, 256] ×3
[1×1, 256 3×3, 256, C=32 1×1, 512] ×4
[1×1, 512 3×3, 512, C=32 1×1, 1024] ×6
[1×1, 1024 3×3, 1024, C=32 1×1, 2048] ×3

图 3.7 ResNeXt-50(32×4d)完整网络参数[67]

一张 224×224 的图像经过该网络后最终输出为 2048×1 的向量，在本文中，为实现年龄预测，在其后面添加了一层全连接层，将其输出为 101×1 的向量，即把年龄预测当成分类问题处理，从 0 岁到 100 岁共 101 类。

3.2.2 数据集

APPA-REAL 数据集共包含 7591 张包含真实年龄标签和表征年龄标签的人脸图片，其中真实年龄范围为 0-95 岁，外观年龄则是根据问卷投票产生，每张图片平均外观年龄票数为 38 票，是人们心中普遍认为的年龄，和实际年龄有一定差异。APPA-REAL 数据集被划分为三部分，其中训练集有 4113 张图片，验证集有 1500 张图片，测试集有 1978 张图片。

3.2.3 训练

训练过程中，选用交叉熵损失函数（Cross Entropy Loss），使用 Adam 优化器对其进行优化。经过 50 个回合的训练，损失函数曲线逐渐下降并趋于稳定，如图 3.8 所示，此时年龄预测准确率却一直在 10%附近摇摆，不再上升。这是因为此处的准确率指的是绝对准确率，一个人的外观年龄和真实年龄之间存在很大的差异，预测器预测的年龄与真实年龄只要不一致就被判定为预测错误。训练过程中这么设置可以使预测器更好的学习年龄特征，实际测试时，可以选用平均预测年龄作为最终的预测结果。人脸图像经过神经网络后被输出为 101 维的向量，在最后一层经过 softmax 运算，转换为预测年龄的概率，利用该概率对年龄求加权平均值即为平均预测年龄。这种预测方式更加贴合人类的判断形式，对年龄区间的把握更加出色。

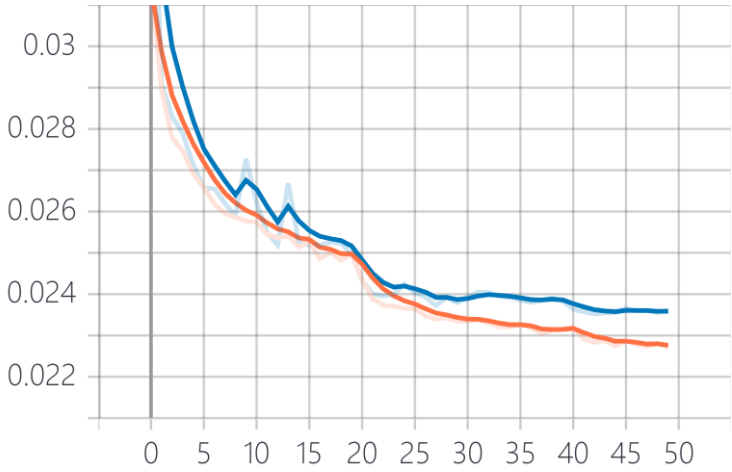

图 3.8 损失函数变化情况（训练（橙色）、验证（蓝色））

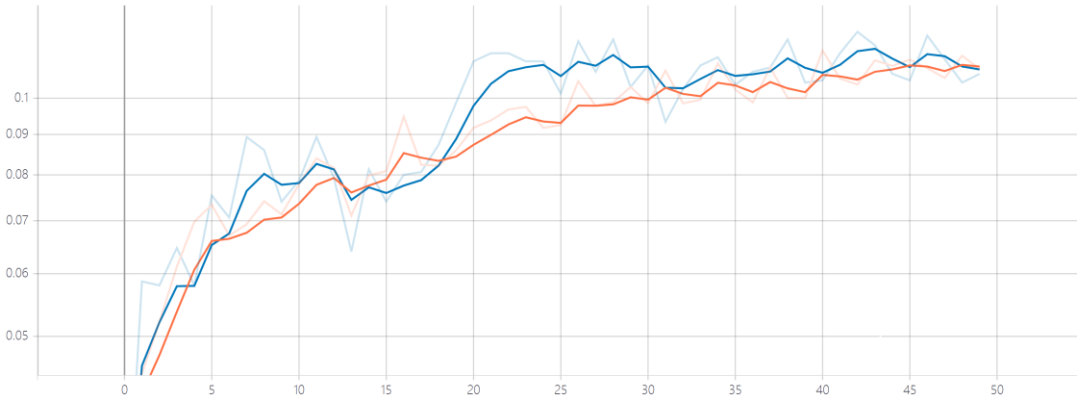

图 3.9 年龄预测准确率变化情况（训练（橙色）、验证（蓝色））

为测试年龄预测器的准确程度，本文选取比较有代表性图像对其中的人物进行年龄预测，如图 3.10 所示

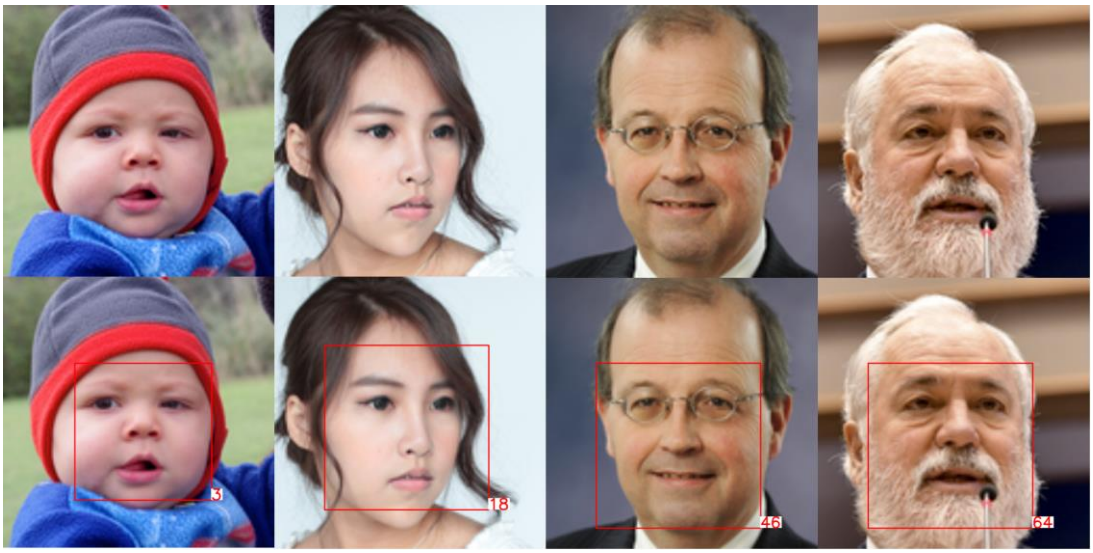

图 3.10 不同年龄段人物年龄预测结果

可以看出，预测器预测的结果和人类的主观判断基本一致，为保证年龄预测器的可靠性，针对 1978 张测试集图像进行预测，并计算平均预测误差。结果显示平均预测误差在五岁以内，并没有出现跨年龄段的预测错误，因此可以认为该年龄预测器基本满足本文强化学习模型的要求。

3.3 深度确定性策略梯度算法

3.3.1 算法选择

在一些经典强化学习问题中，状态空间和动作空间都是有限的，因此可以计算所有可能状态、所有可能动作的 Q 值 $Q(s, a)$，并根据最大的 Q 值选择要执行的下一动作。当遇到状态空间连续的问题时，如 Atari 游戏，$Q(s, a)$的计算及存储是难以进行的，DQN 算法将神经网络与 Q-Learning 结合，利用神经网络去估计某一状态下所有动作的值函数。根据本文提出的人脸编辑模型，不但人脸图像及其中间潜在向量表示是连续的，每一步决策的动作也是连续的，虽然动作的欧式距离是一个定值，但其方向可以是 512 维的空间中超球体的任意方向。基于 DQN 改进的 NAF 算法[70]可以实现连续动作空间上的控制，但实现起来却很复杂。

基于值函数的强化学习算法在处理连续动作空间问题上具有先天性的劣势，因此本文在强化学习算法的选择过程中更加倾向基于策略的算法。和基于值函数的算法相比，基于策略的算法更加高效，也更容易收敛。基于策略的强化学习算法将策略用包含参数 θ 的连续函数表示

$$\pi_\theta(s,a) = P(a \mid s,\theta) \tag{3.6}$$

该方法既能用于离散动作空间，用 softmax 函数计算动作发生的概率，也能用于连续动作空间，从高斯分布中产生动作的概率分布。利用梯度上升法，就可以对策略函数的参数进行优化，使其最大化初始状态收益期望或状态平均价值。在无模型的强化学习分类中，基于策略的算法又分为随机策略和确定性策略，随机策略较为成熟，但需要大量的数据采样，学习过程较慢；确定性策略节省了采样成本，但面对同一状态只能输出唯一动作，不具备自学习的能力。结合本文提出的人脸年龄编辑强化学习模型，考虑到采样成本较大，综合考虑之下采用了深度确定性策略梯度算法对年龄向量求解。

3.3.2 深度确定性策略梯度算法描述

深度确定性策略梯度（Deep Deterministic Policy Gradient, DDPG）算法，在确定性策略梯度（Deterministic Policy Gradient, DPG）算法[72]的基础上融入了 DQN[58]的思想，利用深度神经网络实现连续空间中动作输出及 Q 值的计算。

深度确定性策略梯度算法采用演员评论家（Actor-Critic）架构，共包含四个网络，分别为 Actor 当前网络、Critic 当前网络、Actor 目标网络和 Critic 目标网络。其中 Actor 当前网络负责与环境的交互，根据环境状态 s 选择要执行的动作 a，Critic 当前网络负责根据网络

参数 φ 计算在状态 s 执行动作 a 的 Q 值 $Q(s,a|\varphi)$，Actor 目标网络负责根据下一状态 s' 选择最优的下一动作 a'，Critic 目标网络负责根据网络参数 φ' 计算在下一状态 s' 执行 Actor 目标网络选择的动作 a' 时的 Q 值 $Q'(s',a'|\varphi')$。此外，DDPG 算法还采用了经验回放，如图 3.11 所示，在交互阶段，Actor 当前网络根据参数 θ 及 512 维的中间潜在向量 w_t 选择一个年龄方向向量 a_t，由于本文采用的是确定性策略算法，缺乏探索性，因此需要通过添加高斯噪声增强探索，添加了噪声后的年龄向量为

$$a_t = \pi(w_t|\theta) + N(0,\sigma) \tag{3.7}$$

其中 σ 为方差，在训练的过程会逐渐减小。为保证年龄向量的欧氏距离不发生改变，添加噪声后需要对其归一化处理。

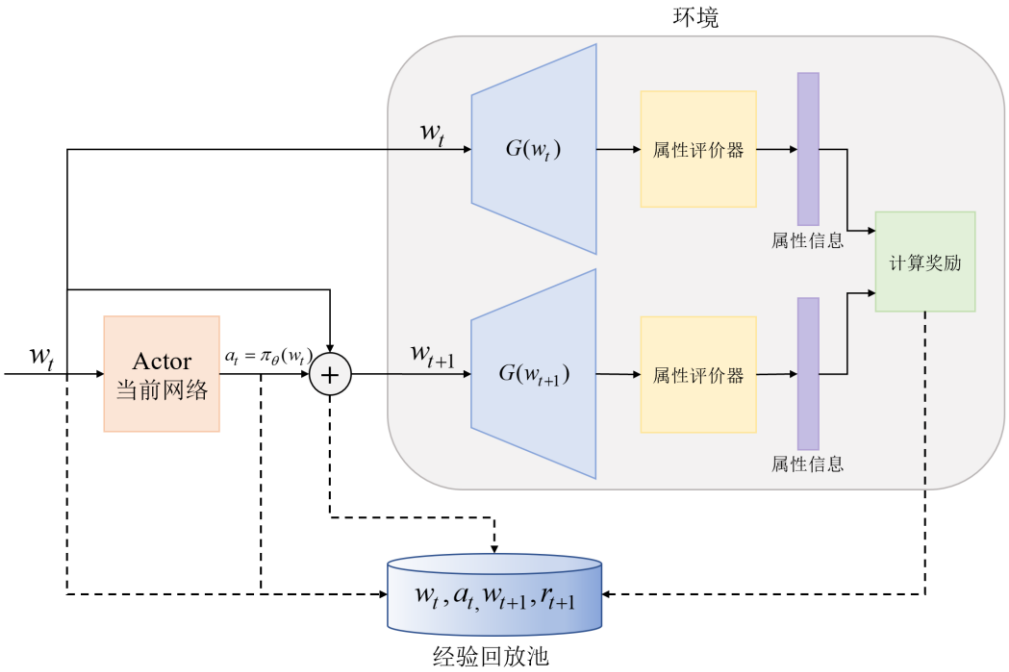

图 3.11 交互过程的经验回放模型

当经验回放池中的数据积累到一定数量的时候，开始进行学习，更新网络参数。首先，从经验回放集合中采样 m 个样本 $\{w_i, a_i, w_{i+1}, R_i\}, i = 1,2,...m$，利用 Actor 目标网络估计最佳动作 $a_{i+1} = \pi'(w_{i+1}|\theta')$，计算当前目标的 Q 值 $y_i = r_i + \gamma Q'(w_{i+1}, a_{i+1}|\varphi')$，通过最小化损失函数

$$L = \frac{1}{m}\sum_{i=1}^{m}(y_i - Q(w_i, a_i|\varphi))^2 \tag{3.8}$$

更新 Critic 当前网络的参数，随后根据策略梯度更新 Actor 当前网络的参数。根据策略梯度定理，在任何马尔可夫决策过程中，目标函数对策略的梯度都具有以下形式

$$\nabla_\theta J(\theta) = E_\pi[\nabla_\theta \log \pi(a|s;\theta) Q_\pi(s,a)] \tag{3.9}$$

由于确定性策略中 Actor 输出的是一个确定动作，无需在整个动作空间采样，因此上式 $\nabla_\theta \log \pi(a|s;\theta)$ 需改为 $\nabla_\theta \pi(s;\theta)$，结合本文实际，策略梯度可改写为

$$\nabla_\theta J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\nabla_a Q(s,a\mid\varphi)|_{s=w_i,a=\pi(w_i)}\cdot\nabla_\theta\pi(s\mid\theta)|_{s=w_i} \tag{3.10}$$

完整的算法伪代码表述如表 3.1 所示

表 3.1 DDPG 算法伪代码

深度确定性策略算法（Deep Deterministic Policy Gradient, DDPG）

清空经验回放集合，设定软更新频率 τ

随机初始化 Actor 当前网络参数 θ，Critic 当前网络参数 φ

更新 Actor 目标网络参数 $\theta'=\theta$，Critic 目标网络参数 $\varphi'=\varphi$

for episode=1，T **do**

获取初始状态（人脸中间潜在向量） w_1

for k=1，step **do**

根据 Actor 当前网络得到动作（年龄向量） $a_k=\pi(w_k\mid\theta)+N(0,\sigma)$

动作归一化 $a_k=\frac{a_k}{|a_k|}$

执行动作 a_k，得到奖励 r_{k+1} 并进入下一状态 w_{k+1}

将 $\{s_k,a_k,r_{k+1},s_{k+1}\}$ 存入经验回放集合

从经验回放集合中取出 m 个样本 $\{s_i,a_i,r_{i+1},s_{i+1}\}$

计算目标 Q 值 $y_i=r_{i+1}+\gamma Q'(w_{i+1},\pi'(w_{i+1}\mid\theta')\mid\varphi')$

使用损失函数 $L=\frac{1}{m}\sum_{i=1}^{m}(y_i-Q(w_i,a_i\mid\varphi))^2$ 更新 Critic 当前网络参数

用策略梯度 $\nabla_\theta J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\nabla_a Q(s,a\mid\varphi)|_{s=w_i,a=\pi(w_i\mid\theta)}\cdot\nabla_\theta\pi(s\mid\theta)|_{s=w_i}$ 更新 Actor 当前网络参数

更新目标网络参数：

$\theta'\leftarrow\tau\theta+(1-\tau)\theta'$

$\varphi'\leftarrow\tau\varphi+(1-\tau)\varphi'$

end for

end for

3.4 实验过程及结果分析

3.4.1 实验参数设置

本文采用在 FFHQ 高清人脸数据集上预训练的 StyleGAN2 模型作为人脸生成器，并利用

一个训练好的年龄预测器对生成的人脸图像的年龄进行预测，环境根据采取行动前后生成的两张人脸图像的年龄预测计算奖励函数，奖励衰减因子为 0.9。

强化学习模型方面，Actor 网络的学习率被设置为 0.001，Critic 网络的学习率被设置为 0.002，当前网络和目标网络之间的软更新率被设置为 0.01。为保证充足的经验样本，经验回放池大小为 1536，经验回放池未存满之前，网络不进行更新，一直处于交互过程，每个轮次，交互步数上限为 100，无论是否成目标均要回到初始状态重新开始。经验回放池存满之后，交互的同时从经验回放池采样对网络进行更新，采样批次大小为 32。此外，当前 Actor 网络与环境交互时所添加的噪声初始方差设置为 3，并在经验回放池存满后以 0.9996 的系数进行衰减。其他方面，实验环境为 tensorflow1.x，显卡为 RTX2080Ti。

3.4.2 直观效果分析

为探索强化学习模型的寻路过程，针对 Actor 做出的每一个动作，对状态的变化情况进行观测，智能体未学到任何策略时，随机探索生成的图像结果如图 3.12 所示

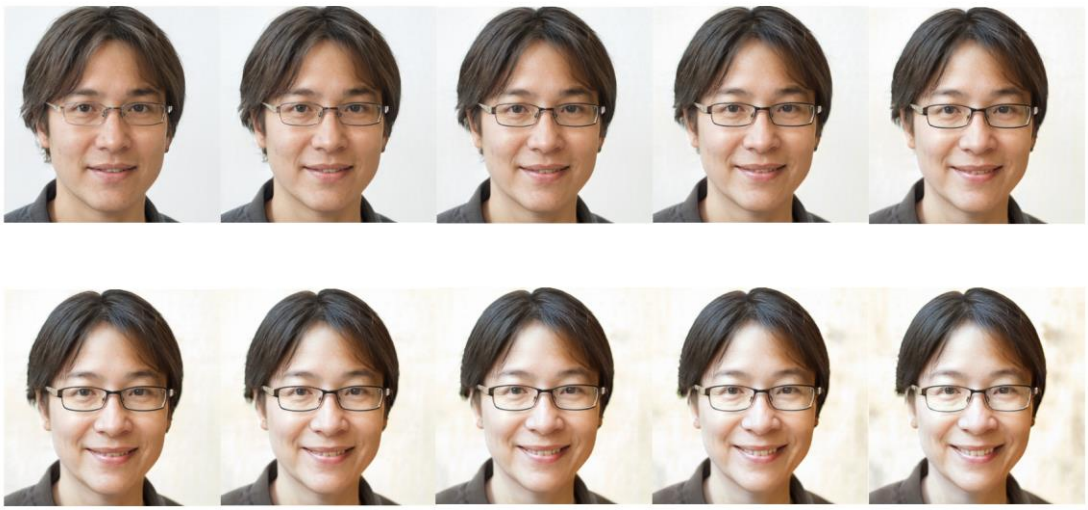

图 3.12 年龄编辑过程中人脸状态变化（随机）

虽然相邻两张图像的差异不大，但经过动作的不断累积，生成人脸的各个属性都在发生变化。在交互过程初期，网络参数均未得到更新，在面对同一状态，输出的动作也是确定的，由于执行动作前后两个状态的差异并不大，这会导致 Actor 的动作方向受限，样本缺乏多样性。在交互过程中加入高斯噪声可以很好的避免这一问题。

训练初期，Actor 会在状态空间进行各种各样的探索，生成的图像也是随机变化的，随着网络的不断更新，智能体的决策向量会逐渐趋向于总奖励大的方向。经过 1000 个轮次的训练，智能体每轮次获得的总奖励逐渐增大并趋于稳定，如图 3.13 所示，此时智能体已经达成了年龄编辑的目标。

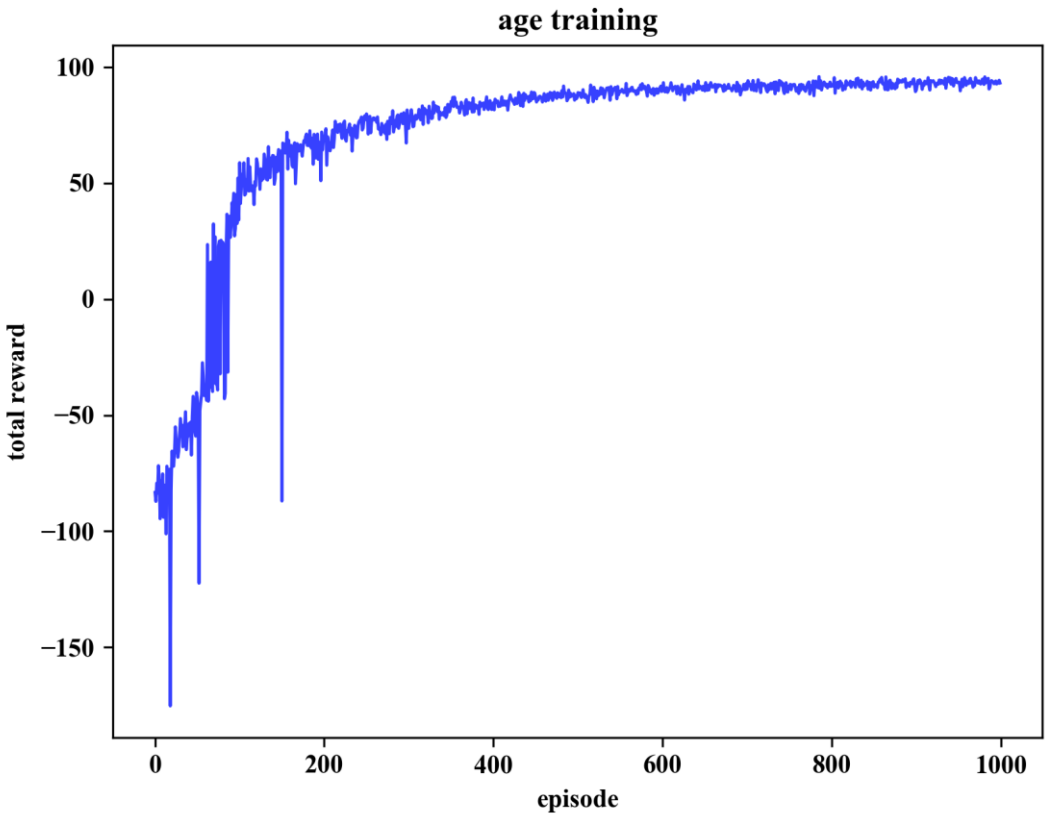

图 3.13 年龄编辑中每回合总奖励随训练回合数变化曲线

保存最优策略，在该策略下从初始状态出发，根据智能体每次做出的动作记录相应状态的转移过程，并将所有状态转换成人脸图像，结果如图 3.14 所示，

初始状态S1　　状态S2　　状态S3

图 3.14 年龄编辑效果图

具有最优策略的智能体仅执行了两个动作就达成了年龄编辑目标，从图中可以看出人脸在老化的过程中，头发逐渐变白，眉毛逐渐变淡，眼部、耳部皱纹逐渐增多，法令纹也逐渐加深，是多种因素共同作用的结果，这也符合年龄的复合属性的假设。在人脸身份特质方面，从直观上来看状态 S1 和 S2 对应的人脸差距并不大，基本可以认为是同一个人，但 S1 和 S3 对应的人脸图像存在较大差距，这是两张图像之间的年龄差距较大所导致的。由于年龄属性的特殊性，在现实生活中，一个人年轻时的样貌和年老时的样貌也存在着较大差异，不能单纯

的因为两张人脸差距较大就认为模型出现了错误。事实上，S1 和 S3 两张人脸图像无论是在发型还是五官位置，均未发生明显变化。

关于年龄编辑引起的样貌差异，一种合理的解释是 Actor 输出动作的欧氏距离过大，反应在图像上就是每次移动会产生较大的年龄跨度。为此，本文对年龄编辑过程进行了插值输出，如图 3.15 所示，经过插值化输出的人脸在年龄过渡上更加自然，这说明年龄这一属性在潜在空间中是近似线性变化的，同时也说明减小 Actor 输出动作的欧氏距离可能会获得更加细致的编辑效果。

图 3.15 人脸年龄平滑过渡效果图

3.4.3 通用性分析

虽然年龄老化向量是从固定的初始状态出发训练得到的，但 Actor 网络输出的动作实际代表的老化的方向，在最优动作表示的方向上仅年龄发生改变，而其他的属性基本不发生改变。结合年龄属性在潜在空间可能是线性变化的，因此可以推测这样的方向也能够对其他的人脸进行年龄编辑。对此，我们将 Actor 网络根据初始状态做出的最优动作进行保存，并将其作为年龄老化方向向量对潜在空间中其他的人脸表示进行年龄编辑，实验结果显示该向量对大多数图片具有良好的年龄编辑效果，图 3.16 给出了一组人脸的年龄老化编辑效果

图 3.16 年龄老化向量通用测试效果图

从中可以看出，不同人脸的老化过程基本一致，均体现在头发颜色逐渐变白，眉毛逐渐变淡，面部皱纹逐渐增多。不同人脸由于在潜在空间分布中的位置不同，细节方面会有些许差异，但总体效果均满足对人脸年龄老化的要求。上述通用测试结果表明，利用深度确定性策略梯度方法求解出的年龄老化向量不只单独适用于特定的人脸，而是具有一定的通用性，实现了一次训练，多次应用的效果，在进行真实人脸的年龄编辑时，只需要将人脸编码到该潜在空间，就可以利用老化向量对其进行老化编辑。

3.4.4 对比分析

为验证利用深度确定性策略梯度方法进行年龄编辑时，保持身份特质方面的效果，本文选取目前最先进的同类方法 InterFaceGAN 与本文提出的方法进行年龄编辑的对比。InterFaceGAN 假设在潜在空间中，任何二元属性都可以被一个线性超平面分为两类，超平面同侧具有相同的属性，对于年龄编辑，InterFaceGAN 选取某一年龄中间值将样本分为两类，并利用该超平面的法向量作为年龄编辑的方向向量。然而该假设忽略了年龄编辑的轨迹性，超平面的法向量方向只能代表大致的年龄老化方向，由于其他属性的干扰，势必会引起身份特质的改变。在同一生成模型下，选取同一张人脸作进行年龄编辑实验对比。图 3.17 中给出了一组人脸年龄编辑的对比效果

本文方法

InterFaceGAN

图 3.17 年龄编辑效果对比图

从图中可以看出，InterFaceGAN 确实实现了人脸年龄的编辑，但是人脸的其他属性却发生了较大的改变。这是因为在构建二分类超平面的时候需要在潜在空间进行大量的采样，如果样本中欧美人种分布的比例较大，则会使超平面的法向量偏向欧美人种。此外 InterFaceGAN 将超平面法向量进行归一化后作为年龄的方向向量，这意味着方向向量始终是固定的，而潜在空间中年龄属性只是近似的线性子空间，会存在一定误差。虽然 InterFaceGAN 对姿势、微笑、性别和是否戴眼镜等四个属性进行了解耦，但事实上人脸属性远不止这些，不可能一一解耦。因此从直观上看，本文基于深度确定性策略梯度的强化学

习方法在进行年龄编辑时，无论是在编辑目标还是保持身份特质上，均优于 InterFaceGAN。为了在客观数据上看出二者之间的差异，本文采用结构相似性（Structure Similarity, SSIM）和 128 维 FaceNet 距离[73]衡量编辑前后的两张图像的相似度。在潜在空间中随机选取 1000 张图片，分别用两种方法进行年龄编辑，计算编辑前后图像的 SSIM 和 FaceNet 距离的平均值，结果如表 3.2 所示

表 3.2 年龄编辑量化对比（SSIM/FaceNet 距离）

	InterFaceGAN	本文方法
SSIM ↑	0.670	0.856
FaceNet 距离 ↓	1.011	0.761

结果显示本文的方法在人脸年龄编辑前后，两张人脸的相似度更高，结合直观感受，可以说明本文基于深度确定性策略梯度的年龄编辑方法能够在实现年龄编辑目标的同时保持较好的身份信息。

3.5 本章小结

本章针对一般人脸属性编辑方法中存在的编辑连续性的问题，提出了新的编辑思路，并给出了基于深度确定性策略梯度的人脸老化方案。利用强化学习对人脸的潜在空间进行逐步探索，通过最大化收益的方式，可以使智能体找到人脸老化的方向路径，实现年龄的连续性编辑，奖励函数对路径的约束可以进一步提升编辑质量，避免编辑前后人脸身份信息变化过大。

本章首先对建模思路和流程做了详细的分析和阐述，利用悬崖寻路问题进行类比，将在潜在向量叠加增量的过程看作马尔可夫决策过程，并建立了相应的强化学习模型。随后，针对人脸年龄编辑问题，给出了强化学习奖励函数的设置思路和方法。根据奖励函数中主线奖励的计算方法，利用 ResNeXt-50 模型训练了一个年龄预测器对探索过程中生成的图像进行年龄预测。由于人脸老化建模中强化学习动作空间的连续性，经过多方位比较，最终选择了深度确定性策略梯度算法进行求解，故对 DDPG 算法及其在人脸老化问题中的具体形式进行了介绍。最后我们将本章提出的方法在训练好的高清图像生成模型 StyleGAN2 上进行了实验，并给出了相关实验参数。

实验结果表明，本章提出的人脸老化模型能够正确的对人脸年龄进行连续编辑，直观效果良好，并且利用强化学习方法找到的人脸老化向量路径具有一定的通用性，并不依赖特定的人脸，实现了一次训练，多次编辑的效果。在 SSIM 和 128 维 FaceNet 距离两项指标的对比上，本章提出的方法均优于目前最先进的同类方法 InterFaceGAN，结合直观图像对比，说明本章的方法在保持人脸身份信息上更稳定，也更加符合人类的认知。

第四章 基于双延迟 DDPG 算法的任意人脸属性编辑

经过上一章的论述，充分证实了本文提出的强化学习模型在人脸年龄编辑上的可行性，基于深度确定策略梯度的强化学习算法在人脸年龄编辑的过程中展现了良好的效果，在人脸身份特质保持方面也具有优势。此外，由智能体决策得到的年龄编辑向量并不只针对特定的人脸有效，而是在潜在空间中具有一定通用性。但是决策向量的欧式距离过大导致年龄的编辑并不细致，在年龄这一属性中该问题并不明显，对于其他属性，则会导致编辑过程的离散化。为了实现平滑过渡的细致编辑，本章利用双延迟 DDPG 算法对模型做了进一步优化。利用现有的属性分类器，通过网络交互实时获取和计算奖励函数，避免了多标签数据集的使用。

4.1 网络交互模型

4.1.1 模型架构

在年龄编辑的过程中，主线奖励的计算需要用到年龄预测器，而训练这样一个年龄预测器，需要用到大量带有真实年龄标签的数据。对于人脸其它属性的编辑，这意味着每次在寻找新的属性方向向量时，都要单独训练一个对应的属性分类器，需要用到大量的多标签人脸数据集。考虑到目前人脸识别及人脸属性识别已经较为成熟，本文利用现有的属性分类模型对生成图像进行分析，搭建了一个类似于 C/S 架构的网络模型。如图所示，本地生成器生在成人脸图像之后，将图像由 PNG 格式转换为 Base64 编码格式，并向远程服务器发送 HTTP 请求，服务器收到请求后将图像的属性信息返回。环境模块根据返回的属性信息计算奖励函数，并将其反馈给智能体。

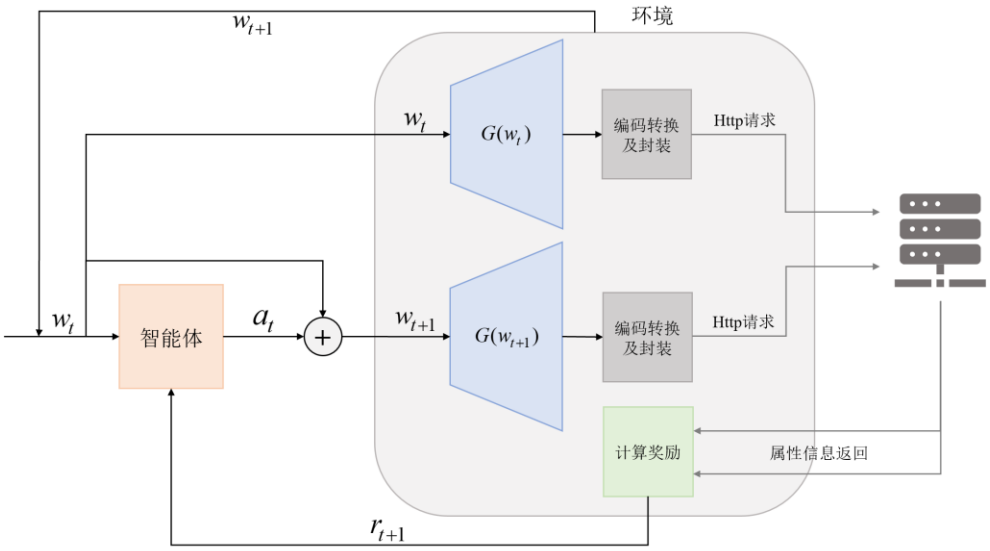

图 4.1 网络交互模型示意图

这种模式下，本地只需负责图像生成及强化学习部分的学习运算，大大减轻了运算负载。此外，现有成熟的人脸属性评价体系一般具有业内较高的准确率，可以为奖励函数的计算提供更为精准的数据，有利于进一步提升人脸图像的编辑质量。本文在服务器端选用百度智能云的人脸检测与属性分析解决方案，该方案在人脸检测部分能够提供人脸位置定位和多达150 个关键点的定位，在人脸属性分析部分能够提供包括年龄、性别在内的多种属性的信息评价，表 4.1 中给出了出本文实验过程中用到的属性及相关描述

表 4.1 百度智能云方案的部分属性分析及描述

属性	描述
性别	男/女，性别置信度（范围 0~1）
是否戴眼镜	无眼镜/普通眼睛/太阳镜，眼镜置信度（范围 0~1）
姿势	左右角度-90（左）、90（右）
眼睛状态	0（闭合）、1（睁开）

4.1.2 奖励函数的优化

在上一章的年龄编辑任务中，环境的奖惩机制主要包括主线奖励、目标达成奖励、交互惩罚和溢出惩罚，在这其中主线奖励最为重要，智能体执行动作前后的两个状态对应图像的年龄差越大代表该决策越好，智能体会通过网络参数更新增大类似决策的概率，使之更加容易朝年龄增大的方向行动。尽管如此，由于潜在空间的高维性，智能体并不是每次行动都会收到主线奖励。主线奖励为零的强化学习数据对网络更新的作用很小，因此智能体仍需要在潜在空间中进行大量无用的探索，降低了搜索最佳路径的效率。

对于本章中要编辑的属性，则更是如此。一个男性在变化为女性的过程中可能长时间收不到主线奖励，在设定的步数内达成编辑目标几乎可以被认为是小概率事件。缺乏有效数据会造成神经网络更新缓慢，智能体不能学到相应的最佳策略，甚至学不到任何有效的策略。为解决这一问题，本文新增了引导奖励，其低维空间的示意图 4.2 如所示

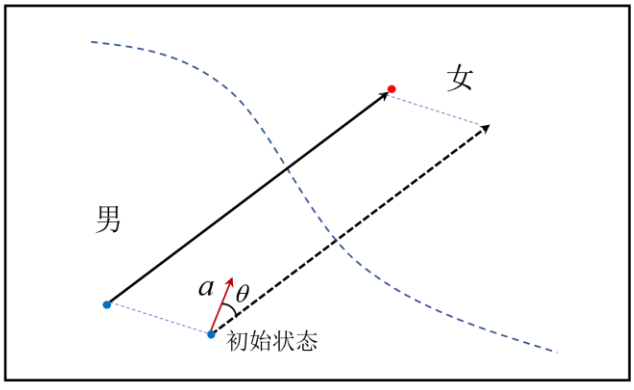

图 4.2 引导向量示意图

假设二值属性（如性别）在潜在空间中通过某一分类面进行划分，在两类样本中分别任取一个样本，其对应的潜在向量在潜在空间中由两个点表示，两点之间形成的向量即定义为引导向量$\boldsymbol{l}_{ref}$。在初始状态下，智能体无论采取任何方向的动作，由于二值属性的类别未发生实质性改变，因而均不能获得主线奖励。引导向量的加入使得决策向量和引导向量之间存在一个夹角，如果该夹角是锐角，则说明智能体的决策方向是正确的，应该获得奖励，反之则要受到惩罚。因此可以根据引导向量和决策向量夹角的余弦值定义引导奖励

$$r_{ref} = \cos < \boldsymbol{a}_t, \boldsymbol{l}_{ref} > \tag{4.1}$$

引导奖励的取值范围是[-1, 1]，这说明决策向量与引导向量的夹角越小则收到的奖励越大，这会导致智能体做出的决策更加贴合引导向量。但引导向量是根据潜在空间中分类面两侧的任意点计算得到的，只能代表一个粗略的指向，并不是编辑该二值属性的最佳向量。过分依赖引导奖励会使编辑前后人脸的身份发生改变，因此引导奖励需要随训练过程逐步衰减，直至完全消失。这样既能使智能体在前期的探索中快速收集有效数据，找到属性编辑的大致方向，又可以在后期寻找到最佳的编辑方向路径，保持身份的不变。

4.2 双延迟 DDPG 算法

4.2.1 深度确定性策略梯度存在的问题

在一些基于值函数的强化学习算法中，如 Q-learning 和 DQN，并不进行下一次的真实交互，而是根据当前策略选取动作价值最大的动作，这会导致动作价值存在过估计的问题，深度确定性策略梯度算法（DDPG）作为 DQN 在连续动作空间上的拓展，也存在该问题。动作价值的过高估计会使智能体选择一个实际上较差动作对网络进行更新，从而导致策略的崩溃。使用 DDPG 算法对年龄进行编辑时，由于每次执行动作都能收到较大的主线奖励，该问题并不明显。当对其他属性进行细致编辑时，需要通过减小输出动作的欧氏距离实现编辑的连续性。这种情况下，每次执行动作收到的主线奖励就会变小，过估计的存在会使目标 Q 值中的延时奖励部分被忽略，而将过估计的动作价值函数当成更新的目标，即

$$y_i = \gamma Q'(w_{i+1}, a_{i+1} \mid \varphi') \tag{4.2}$$

随着时间的累积，策略会变得逐渐不稳定甚至崩溃。为解决这一问题，本文采取了双延迟深度确定性策略梯度算法（Twin Delayed Deep Deterministic Policy Gradient, TD3）[74]对其他属性进行更为细致的编辑。

4.2.2 算法描述

TD3 采用了 Double DQN 中防止 Q 值过估计的设计思路，使用双网络分别对动作价值函数进行估计。如图 4.3 所示，TD3 由 6 个网络组成，分别是 Actor 当前网络，Actor 目标网络，两个 Critic 当前网络和两个 Critic 目标网络。其中 Actor 当前网络负责与环境交互获取

强化学习数据，未在图中体现出来。在训练阶段，每次从经验回放池采样 m 个数据样本 $\{w_i, a_i, w_{i+1}, r_{i+1}\}, i=1,2,...m$ ，Actor 目标网络根据下一状态 w_{i+1} 估计下一动作 $a_{i+1}=\pi'(w_{i+1}|\theta')$，随后两个 Critic 目标网络根据下一状态 w_{i+1} 和下一动作 a_{i+1} 分别估计动作价值函数 Q_1' 和 Q_2'，并选取较小的作为目标 Q 值

$$y_i = r_{i+1} + \gamma \min(Q_1'(w_{i+1}, \pi'(w_{i+1}|\theta')|\varphi_1'), Q_2'(w_{i+1}, \pi'(w_{i+1}|\theta')|\varphi_2')) \tag{4.3}$$

最后，利用该目标 Q 值分别对两个 Critic 当前网络进行更新。

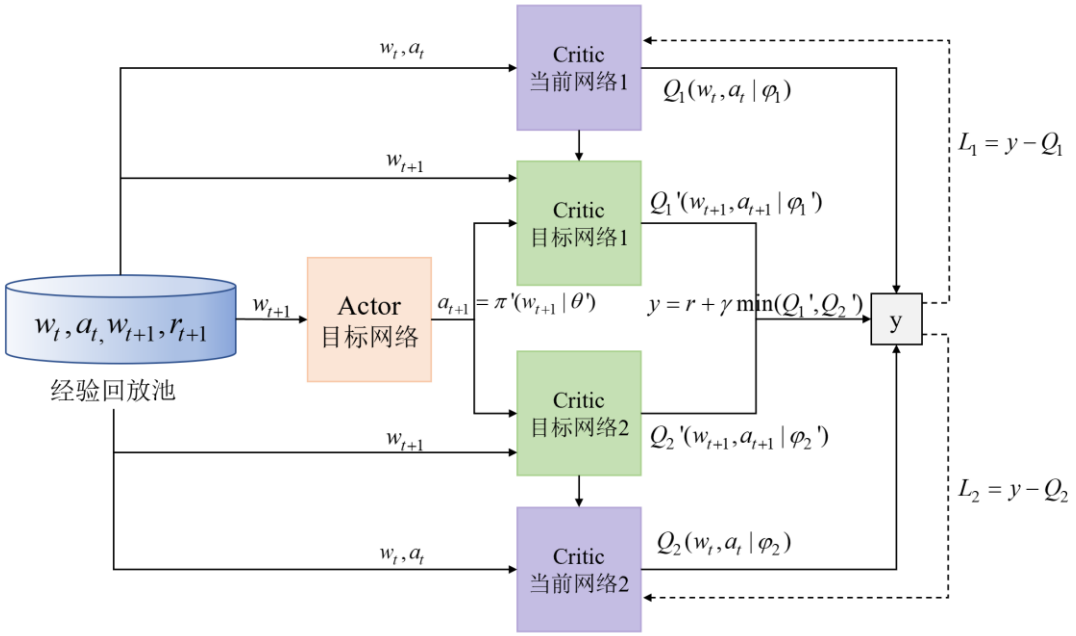

图 4.3 TD3 网络结构示意图

由于两对 Critic 网络是独立的，彼此互不影响，选取较小的 Q 值估计可以有效缓解过估计的问题。此外，TD3 还在对 Actor 目标网络添加了经过裁剪的噪声，使对 Q 值的估计在动作维度更加平滑，并使 Actor 当前网络的更新频率小于 Critic 当前网络的更新频率。结合本文实际的完整 TD3 算法如下表所示

表 4.2 TD3 算法伪代码

双延迟深度确定性策略梯度算法（Twin Delayed Deep Deterministic Policy Gradient, TD3）
清空经验回放集合，设定软更新频率 τ
随机初始化 Actor 当前网络参数 θ，Critic 当前网络参数 φ_1， φ_2
更新 Actor 目标网络参数 $\theta'=\theta$，Critic 目标网络参数 $\varphi_1'=\varphi_1$， $\varphi_2'=\varphi_2$
for episode=1，T **do**
获取初始状态（人脸中间潜在向量） w_1
for k=1，step **do**
根据 Actor 当前网络得到动作（年龄向量） $a_k=\pi(w_k|\theta)+N(0,\sigma_1)$

动作归一化 $a_k = \frac{a_k}{|a_k|}$

执行动作 a_k，得到奖励 r_{k+1} 并进入下一状态 w_{k+1}

将 $\{s_k, a_k, r_{k+1}, s_{k+1}\}$ 存入经验回放集合

从经验回放集合中取出 m 个样本 $\{s_i, a_i, r_{i+1}, s_{i+1}\}$

$a_{i+1} = \pi(w_{i+1} | \theta') + \varepsilon$， $\varepsilon \sim clip(N(0, \sigma_2), -c, c)$

动作归一化 $a_{i+1} = \frac{a_{i+1}}{|a_{i+1}|}$

计算目标 Q 值 $y_i = r_{i+1} + \gamma \min(Q_1'(w_{i+1}, a_{i+1} | \varphi_1'), Q_2'(w_{i+1}, a_{i+1} | \varphi_2'))$

使用损失函数 $L_j = \frac{1}{m}\sum_{i=1}^{m}(y_i - Q_j(w_i, a_i | \varphi_j))^2, j = 1,2$ 分别更新 Critic 当前网络参数

if k mod d **then**

用策略梯度 $\nabla_\theta J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\nabla_a Q(s, a | \varphi_1)|_{s=w_i, a=\pi(w_i|\theta)} \cdot \nabla_\theta \pi(s | \theta)|_{s=w_i}$ 更新 Actor 当前网络参数

更新目标网络参数：

$\theta' \leftarrow \tau\theta + (1-\tau)\theta'$

$\varphi_1' \leftarrow \tau\varphi_1 + (1-\tau)\varphi_1'$

$\varphi_2' \leftarrow \tau\varphi_2 + (1-\tau)\varphi_2'$

end if

end for

end for

4.3 实验过程及结果分析

4.3.1 训练技巧

由于本章实验中用的是现有的属性分类器，无法按照自己的意愿输出便于奖励函数计算的结果，因此需要对属性分类的反馈进行调整，这里给出本文训练中用到的技巧。对于一般的均匀数值评价，可以按照一定比例放大到适合强化学习任务的范围，主线奖励为前后两次数值相减。对于一些二值属性，如性别，一般是概率评价，即以多大的概率认定某张人脸图像具有该属性，此时需要根据所选模型的实际情况构建函数，将概率表示转化为数值表示。以性别编辑为例，我们得到的反馈大多数是以 90%以上的概率认定这是一个男性或女性，而几乎不会得到以 50%的概率认定这是一个男性或女性。假设男性概率为 100%时属性评分为 0，女性概率为 100%时属性评分为 100，则图 4.4 中所示的分段函数表示当男性概率由

100%下降至 90%时（图中用负数表示），属性评分从 0 上升至 50。

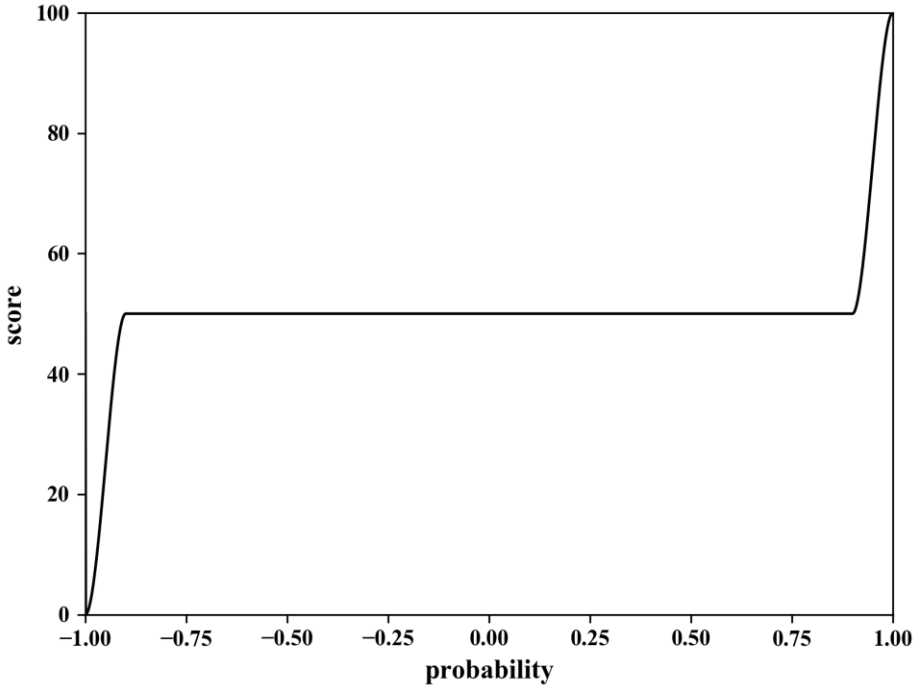

图 4.4 性别编辑中奖励转换函数

因此在人脸样本出现密集的区间，属性评分需要急促变化，人脸样本出现较少的区间则不用去关心，因为在此处的奖励函数计算是没有意义的。

此外，动作函数的欧氏距离和每执行一次动作可获得的奖励几乎是成反比的，为实现人脸的连续性编辑，动作的欧氏距离需要缩短到一个合适的数值，这会导致奖励反馈减弱，造成学习效率的下降，即便添加了引导奖励后，也需要很长的时间才能找到大致的属性向量方向。因此一种更好的办法是先采用欧氏距离较大的动作进行训练，在学得相应策略后，缩短动作向量的长度，再次进行训练。在已有的策略下，智能体会做出同样方向的决策，只是距离缩短了，因而能够快速收集有效数据，找到新的策略。

4.3.2 直观效果分析

跨性别编辑大多用于娱乐应用，由于性别的限制，必然导致身份的改变，因此很难有一种标准去衡量性别编辑的好坏，因此本节仅从直观的编辑效果上对其进行分析。如图所示，随着训练回合的增加，智能体每回合获得的收益逐渐上升并趋于稳定，由于采用了双延迟 DDPG 算法，避免了训练中后期收益波动甚至收益骤降情况的发生，这说明智能体学习到了稳定的策略。1000 回合之后，缩短智能体输出动作的欧氏距离，由于到达编辑目标所需的步数增加，智能体获得收益下降，但智能体能够在当前策略的基础上迅速重新找到最佳策略。

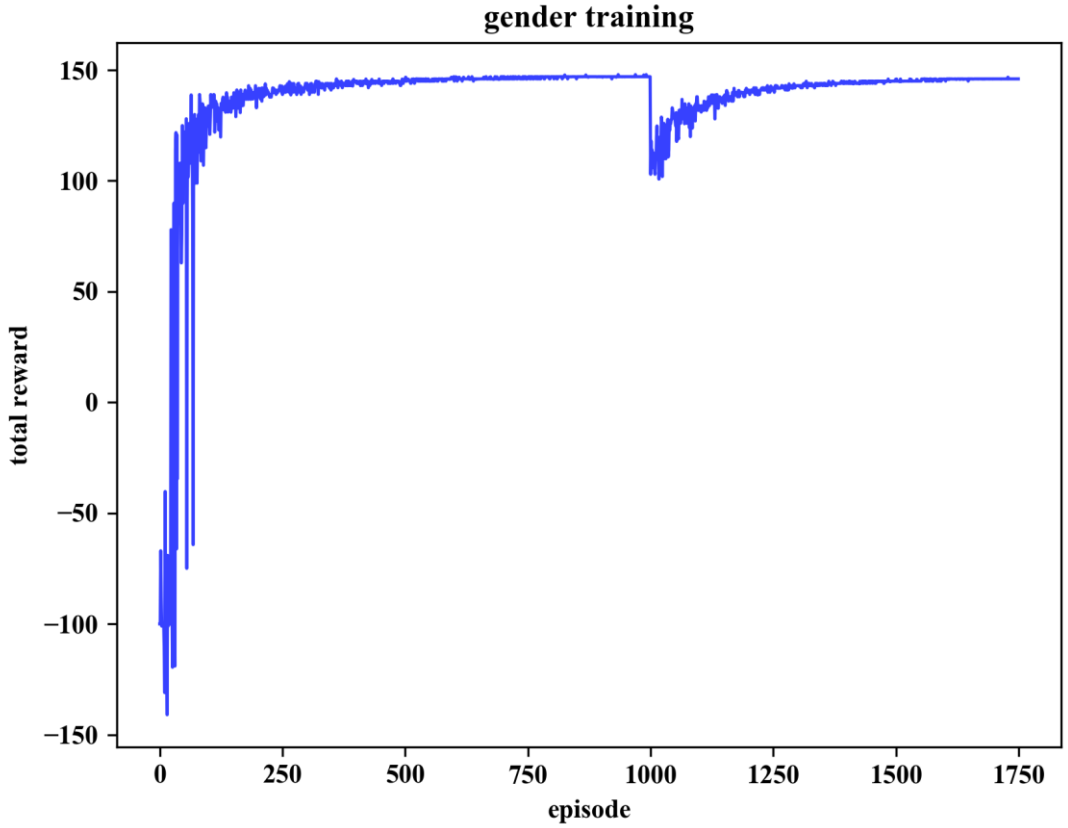

图 4.5 性别编辑中每回合总奖励随训练回合数变化曲线

在策略再次稳定后，将智能体做出的决策动作向量保存，对人脸图像进行编辑。如图 4.6 所示，图像中男性逐渐向女性化的方向进行过渡，具体体现为头发逐渐变长，嘴唇逐渐红润饱满。此外，将编辑前后两张图像进行对比，可以发现两张人脸相似度仍较高，一些生物性状并未发生明显的改变。

图 4.6 性别编辑效果图

缩短决策向量可以使智能体探索到一个完整的决策向量路径，从而可以利用强化学习寻找一条最优的路径，沿该路径移动可以使被编辑属性的变化最大，其他属性基本不发生改变。如果决策向量过大，则会出现一次决策就到达编辑目标的情况，这种情况下虽然能完成指定属性的编辑，但由于不能保证路径最短，会造成编辑前后人脸的差异较大。如图 4.7 所示

图 4.7 决策向量过大时的性别编辑效果

性别编辑的强化目标是从男性出发寻找一条由男变女的路径，由于任务设定的原因，当生成图像被判定为女性时，智能体就停止继续搜索，因此图像在被编辑后大多呈现为短发女性，沿此路径继续进行编辑，可能会出现长发女性的特征。考虑到不同人脸图像的潜在表示在潜在空间中的位置并不相同，一些本来就偏女性化的图像在沿该路径进行编辑后更容易出现明显的女性特征，如图 4.8 所示，

图 4.8 其他人脸性别编辑效果

可以看出，之前用于年龄编辑的人脸在进行性别编辑的过程中，出现了更加明显的女性特征变化，保留基本面部特征的同时，在头发，嘴型和牙齿特征上发生了改变，这说明基于强化学习的人脸编辑方法仅改变属性编辑的关键特征，而几乎不改变其他特征，这也是其他属性编辑中身份信息得以良好保持的原因。此外，性别向量可以对不同人物进行性别编辑进一步说明了属性向量的通用性。

图 4.9 给出了是否戴眼镜眼镜、人脸角度和眼睛开合三个属性的编辑效果图，直观来看，三个属性均达到了设定的编辑要求，且取得良好的直观效果。眼睛的开合程度是一个不常见的属性，通常情况下用于生成对抗网络训练的数据集中很少出现闭着眼睛的人物，强化学习智能体对这项属性的编辑，一方面说明了 StyleGAN2 潜在空间的完备性，另一方面也体现了强化学习方法在人脸属性编辑中的优势，由于强化学习的目标是根据编辑要求设定的，因此理论上只要奖励函数经过合理的设计，并且存在相应的属性评价，就能够实现任意属性的编辑。

图 4.9 眼镜、人脸角度、眼睛开合编辑效果

4.3.3 差分分析

本文提出的方法在人脸图像编辑前后保持身份信息方面具有良好的效果，在人脸结构相似度（SSIM）和 128 维 FaceNet 距离指标上，与同类算法 InterFaceGAN 相比，均存在较大的优势，如表 4.3 所示

表 4.3 不同属性 SSIM/FaceNet 距离对比结果

SSIM ↑	角度	眼睛开合	性别	眼镜
本文方法	0.764	0.889	0.733	0.906
InterFaceGAN	—	—	0.733	0.674
FaceNet 距离 ↓				
本文方法	0.893	0.758	0.907	0.724
InterFaceGAN	—	—	0.997	0.795

为进一步了解人脸编辑过程中属性的变化情况，我们对编辑过程中生成的图像进行差分处理，突出智能体在每次决策后对应图像的变化部分。以眼镜编辑过程为例，如图 4.10 所示，

图 4.10 相邻编辑图像差分效果图

图中黑色代表编辑前后未发生变化的部分，结果显示，当智能体学习到稳定策略时，每次决策几乎仅对相应图像的眼部进行改变，其他部分的改变则很小。从第二次决策开始，图中隐约可见眼镜框的出现，第三、四次决策更加深了对眼镜框部分的生成。这从侧面说明了本文假设的正确性，同时也可以说明在最大化收益的约束下，强化学习智能体产生了类似于注意力机制的效果，专注于某一属性的编辑，减小了对其他属性的改变，从而能够保持编辑前后人脸的身份信息特征。将四幅差分图像进行叠加，得到初始人脸与编辑完成后人脸的差分图像，如图 4.11 所示，

图 4.11 眼镜编辑完成前后总体差分效果

可以发现，眼镜属性编辑前后，主要发生变化的部分仍集中在眼部，两鬓也有些许波及。其他发生改变的部分，如嘴部、鼻孔附近较为轻微，几乎可以忽略不记。性别、眼睛开合和人脸角度编辑完成前后的差分图像如图 4.12 所示，三者均在被编辑属性上发生相应的变化，其他属性则变化很小，由此可以说明强化学习智能体可以根据不同的奖励反馈实现不同属性的特异性编辑，并保持良好的身份不变性。因此只要能够对某一属性进行评价，本文提出的方法就可以根据奖励反馈对其进行编辑。

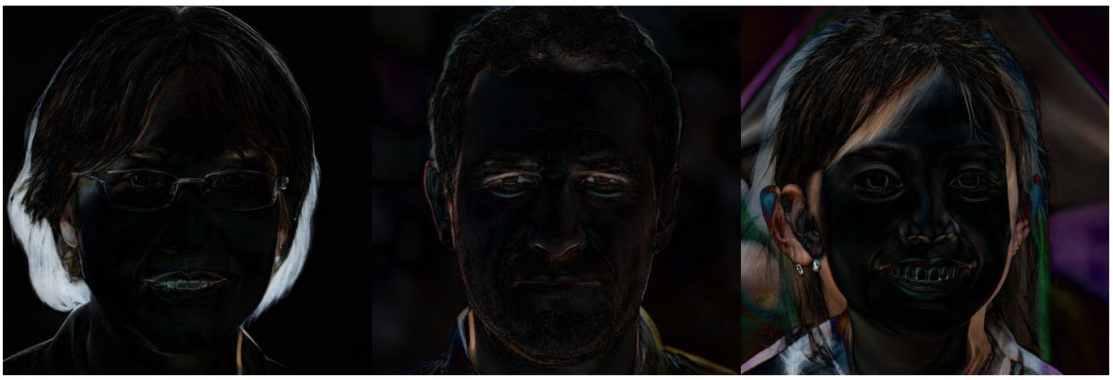

图 4.12 其他三种属性编辑完成前后差分效果图

4.4 本章小结

本章针对年龄老化实验过程中年龄跨度大、编辑不细致的情况作了进一步分析，同时指出在年龄预测器的训练中用到了大量的带标签的数据集。为实现人脸其它属性的细致化编辑，避免对多标签数据集的依赖，本章利用现有的属性分类器计算主线奖励，提出了基于双延迟 DDPG 算法的网络交互模型。

本章首先介绍了网络交互模型的架构，本地只进行潜在空间的探索和中间图像的生成，而将属性评价工作交给远端服务器，为此本章选用了较为成熟的百度智能云人脸属性分析方案为奖励函数的计算提供支持。针对二值化属性长时间得不到奖励反馈的问题，对奖励函数做了相应的优化，提出了引导奖励的概念。随后，针对深度确定性策略梯度算法中存在的 Q 值过估计的问题进行了分析，介绍了双网络的解决方案，并采用 TD3 算法对人脸潜在空间的其他属性向量进行求解。本章最后，对实验过程中的一些训练技巧做了相关说明。

实验结果表明，TD3 算法能有效避免 DDPG 算法在训练过程中出现的策略崩溃问题，智能体能够学习到更加稳定的策略，能够对性别、是否戴眼镜、人脸角度、眼睛开合等属性的进行细致的连续编辑。差分分析表明，基于强化学习的人脸编辑方法会根据奖励函数的设置对人脸的属性进行特异性编辑，而很少对人脸的其他区域有影响，这是人脸身份信息在编辑过程中得以良好保持主要原因，由此也从侧面说明只要能够对属性做出评价，就可以对其进行编辑。

第五章 总结与展望

5.1 本文工作总结

人脸属性编辑作为近几年计算机视觉领域的研究热点，在娱乐、美妆等领域展现了巨大的应用价值，在数据增强甚至安防领域也有巨大的应用潜力，在未来很长一段时间内都是计算机视觉领域研究的重点。本文从实际出发，对人脸属性编辑的意义和背景做了深入的调研，查阅了大量的文献。虽然现有基于模型的方法和基于附加条件的方法已经取得了很大的进步，但在人脸属性编辑的连续性方面仍有欠缺，二值化的编辑无法满足一些属性的变化要求，此外，人脸属性编辑前后容易发生不可控的改变，造成身份信息发生变化。因此本文从这两点出发展开研究，本文主要工作如下：

（1）将强化学习与计算机视觉相结合，利用训练好的生成对抗网络作为载体，通过在潜在向量上叠加属性向量的方式探索生成图像的变化，以此达到人脸属性编辑的目的。本文对这一探索过程进行了马尔科夫决策过程建模，人脸属性编辑问题转换成一个强化学习问题。

（2）针对人脸属性中年龄这一特殊复合属性，本文提出基于深度确定性策略梯度算法的人脸老化模型，对其建模思路和建模过程进行了详细的阐述。奖励函数作为强化学习解决实际问题的核心，决定着实验的成败，本文历经多次实验，提出了一个融合了多种奖惩机制的奖励函数，该奖励函数能够对年龄变化做出正确的评价，最终引导智能体以最短的路径完成人脸老化的编辑任务。实验结果验证了利用强化学习方法实现年龄编辑的正确性，在相关奖惩机制的约束下实现了人脸年龄的连续性编辑，同时具有良好的身份信息保持特性。实验过程中还发现，以强化学习方法寻找得到的属性向量并不依赖于某个对象，而是具有通用性，表征了年龄属性的变化方向。

（3）为使这一强化学习模型能够适应人脸任意属性，并实现更加细致的编辑，本文对其进一步优化，抛弃了属性分类预测模块的训练，而采用现有的属性分类器，提出了网络交互模型，不但减轻了本地的运算压力，还减少了对多标签数据集的依赖。针对 DDPG 算法中 Q 值过估计的问题，本文利用双网络对其进行估计，采用了双延迟的深度确定性策略梯度算法。在训练过程中提出引导奖励以保证智能体快速有效的收集有效数据，并给出了二值化属性连续编辑的训练技巧。实验结果表明基于双延迟的深度确定性策略梯度算法的人脸属性编辑方法能够学习到更加稳定的策略，实现对人脸属性更加细致的编辑。差分分析表明，本文提出的方法能够根据奖励函数的设置自动寻找特异性的属性向量路径，该路径下几乎不会对人脸其它属性造成影响，鉴于奖励函数的重要性，只要能够对属性做出评价，得出奖励函数，就可以对任意的属性进行编辑。

5.2 未来展望

在人脸属性连续性编辑和编辑前后人脸身份信息保持两点上，本文提出的方法取得了较好的效果，验证了强化学习方法在人脸属性编辑应用中的可行性和正确性，为人脸属性编辑这一计算机视觉问题提供了新的参考思路。此外，本文也存在一些劣势和不足之处，需要在未来的工作中进一步解决。

首先，强化学习中的环境是人为定义的，由于强化学习方法的探索性质，每进行一次探索都要对图像进行生成和计算奖励函数，计算开销比较大，致使训练过程较慢。网络交互模型虽然缓解了这一问题，但又引入了网络传输延迟。

其次，本文中固定了智能体的决策向的长度，由于不知道属性向量长度和属性编辑程度的对应关系，并不能一定保证编辑是连续的，如果一次决策就能达到编辑目标就会出现二值化的情况，未来或许可以考虑自适应长度的决策向量实现自动调整。

最后，由于潜在空间中相同长度的不同属性向量对应不同的编辑程度，无法实现不同属性向量的互相叠加，若想实现比较合理的人脸多属性编辑，需要进行手动调整或者联合要编辑的属性重写奖励函数并重新训练。未来可以考虑在一次训练中对多个属性进行联合训练，以获得不同属性之间的长度比例。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. MIT Press, 2014.

[3] Thies J, Zollhfer M, Stamminger M, et al. Face2Face: Real-time Face Capture and Reenactment of RGB Videos[J]. 2020.

[4] Zheng X, Guo Y, Huang H, et al. A Survey of Deep Facial Attribute Analysis[J]. International Journal of Computer Vision, 2020(8).

[5] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J]. IEEE, 2017.

[6] Bledsoe W W. Man-machine face recognition. 1966.

[7] Ramanathan N, Chellappa R. Modeling Age Progression in Young Faces[C]// Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006.

[8] Zhu Z, Luo P, Wang X, et al. Recover Canonical-View Faces in the Wild with Deep Neural Networks[J]. Eprint Arxiv, 2014.

[9] Zhang Z, Yu P. Eyeglasses Removal from Facial Image Based on MVLR[J]. Springer New York, 2013, pp.101-109.

[10] Mu L, Zuo W, D Zhang. Deep Identity-aware Transfer of Facial Attributes[J]. 2016.

[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[12]Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C]//Advances in neural information processing systems. 2017: 700-708.

[13] Wei S, Liu R. Learning Residual Images for Face Attribute Manipulation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[14] Sohn K, Yan X, Lee H. Learning Structured Output Representation using Deep Conditional Generative Models. 2015.

[15] Yan X, Yang J, Sohn K, et al. Attribute2Image: Conditional Image Generation from Visual Attributes[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[16] Zhang Z, Song Y, Qi H. Age Progression/Regression by Conditional Adversarial Autoencoder[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[17] Perarnau G, Weijer J, Raducanu B, et al. Invertible Conditional GANs for image editing[J]. 2016.

[18] He Z, Zuo W, Kan M, et al. AttGAN: Facial Attribute Editing by Only Changing What You Want[J]. 2017.

[19] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[20] Lample G, Zeghidour N, Usunier N, et al. Fader Networks: Manipulating Images by Sliding Attributes[J]. 2017.

[21] Zhang G, Kan M, Shan S, et al. Generative Adversarial Network with Spatial Attention for Face Attribute Editing[C]// European Conference on Computer Vision. Springer, Cham, 2018.

[22] Zhou S, Xiao T, Yi Y, et al. GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data[C]// British Machine Vision Conference 2017. 2017.

[23] Xiao T, Hong J, Ma J. ELEGANT: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes[J]. Springer, Cham, 2018.

[24] Shen Y, Yang C, Tang X, et al. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs[J]. 2020.

[25] Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(4):442–455, 2002. 2, 7

[26] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages 387–394. IEEE, 2006. 2

[27] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3):385–401, 2010. 2, 7, 8

[28] Y. Tazoe, H. Gohara, A. Maejima, and S. Morishima. Facial aging simulator considering geometry and patch-tiled texture. In ACM SIGGRAPH 2012 Posters, page 90. ACM, 2012. 2

[29] Salakhutdinov R, Hinton G. Deep boltzmann machines[C]//Artificial intelligence and statistics. PMLR, 2009: 448-455.

[30] Hinton G E. Deep belief networks[J]. Scholarpedia, 2009, 4(5): 5947.

[31] Ng A. Sparse autoencoder[J]. CS294A Lecture notes, 2011, 72(2011): 1-19.

[32] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders[C]// Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008. 2008.

[33] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.

[34] Kingma D P, Welling M. Auto-Encoding Variational Bayes[J]. arXiv.org, 2014.

[35] Doersch C. Tutorial on Variational Autoencoders[J]. 2016.

[36] Vahdat A, Kautz J. NVAE: A Deep Hierarchical Variational Autoencoder[J]. 2020.

[37] 刘颖, 朱丽, 林庆帆. 基于空间变换网络的图像超分辨率重建[J]. 西安邮电大学学报, 2020.

[38] 陈佛计,朱枫,吴清潇,郝颖明,王恩德,崔芸阁.生成对抗网络及其在图像生成中的应用研究综述[J].计算机学报,2021,44(02):347-369.

[39] Arjovsky M, Bottou L. Towards Principled Methods for Training Generative Adversarial Networks[J]. Stat, 2017, 1050.

[40] Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. Computer ence, 2015.

[41] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. PMLR, 2015: 448-456.

[42] Maas A L, Hannun A Y, Ng A Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. 2013.

[43] Xu B, Wang N, Chen T, et al. Empirical Evaluation of Rectified Activations in Convolutional Network[J]. Computer ence, 2015.

[44] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Icml. 2010.

[45] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[J]. 2017.

[46] Mao X, Li Q, Xie H, et al. Least Squares Generative Adversarial Networks[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[47] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved Training of Wasserstein GANs[J]. 2017.

[48] Zhang H, Xu T, Li H, et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[49] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[50] Karras T, Aila T, Laine S, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation[J]. 2017.

[51] KarrasT, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[52] Huang X, Belongie S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[J]. IEEE, 2017.

[53] Karras T, Laine S, Aittala M, et al. Analyzing and Improving the Image Quality of StyleGAN[J]. 2019.

[54] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of artificial intelligence research, 1996, 4: 237-285.

[55] Kröse, Ben J. A. Learning from delayed rewards.[C]// International Workshop on Network & Operating System Support for Digital Audio & Video. Springer-Verlag, 1991.

[56] Rummery G A, Niranjan M. On-Line Q-Learning Using Connectionist Systems[J]. Technical Report, 1994.

[57] Sutton R S. Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding[J]. Advances in Neural Information Processing Systems, 1996, 8.

[58] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with Deep Reinforcement Learning[J]. Computer Science, 2013.

[59] Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation[C]//Advances in neural information processing systems. 2000: 1057-1063.

[60] Suo, J., Zhu, S.C., Shan, S., Chen, X.: A compositional and dynamic model for face aging. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI) 32(3), 385–401 (2010)

[61] Nhan Duong, C., Luu, K., Gia Quach, K., Nguyen, N., Patterson, E., Bui, T.D., Le, N.: Automatic face aging in videos via deep reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10013–10022 (2019)

[62] Liu, Y., Li, Q., Sun, Z.: Attribute-aware face aging with wavelet-based generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11877–11886 (2019)

[63] Li, Y., Wang, R., Liu, H., Jiang, H., Shan, S., Chen, X.: Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3819–3827. IEEE (2015)

[64] Chang, H., Lu, J., Yu, F., Finkelstein, A.: Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In: Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition (CVPR), pp. 40–48 (2018)

[65] Cao, C., Lu, F., Li, C., Lin, S., Shen, X.: Makeup removal via bidirectional tunable de-makeup network. IEEE Transactions on Multimedia (2019)

[66] Brock A, Donahue J, Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis[J]. 2018.

[67] Xie S, Girshick R, P Dollár, et al. Aggregated Residual Transformations for Deep Neural Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[68] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016.

[69] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. IEEE Computer Society, 2014.

[70] Gu S, Lillicrap T, Sutskever I, et al. Continuous Deep Q-Learning with Model-based Acceleration[J]. JMLR.org, 2016.

[71] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. Computer ence, 2015.

[72] Silver D, Lever G, Heess N, et al. Deterministic Policy Gradient Algorithms. JMLR.org, 2014.

[73] F Schroff, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering[J]. IEEE, 2015.

[74] Fujimoto S, Hoof H V , Meger D. Addressing Function Approximation Error in Actor-Critic Methods[J]. 2018.

致谢

光阴似箭，眨眼间两年半的研究生时光已接近尾声。回首这段时光，从入学前的憧憬到刚入学时的迷茫，从刚开始做实验的好奇到科研过程中的辛酸，我经历了很多，沿途并不平坦，甚至有些坎坷。一路走来，多亏了老师同学和家人的关心和支持，值此论文完成之际，向他们致以最真挚的谢意。

首先，感谢我的导师谭晓阳教授。本论文是在谭老师的悉心指导下完成的，从选题到实验再到论文撰写，谭老师在每个环节都进行了耐心的指导，给出了大量宝贵的意见和建议。谭老师渊博的学识，不断创新的科研思维，严谨的治学态度以及持续学习的精神深深的感染和激励着我，使我能够顺利完成研究生阶段的学习。在此，谨向谭老师致以衷心的感谢和崇高的敬意。

研究生期间的学习和生活还离不开实验室师兄和同学们的帮助，在此特别感谢李尧和张哲两位师兄，他们指引我研究生快速入门，为科研生活走向正轨奠定了基础。感谢同级的蒋珂同学为我答疑解惑，感谢所有师兄师姐、同门和师弟师妹共同营造的和谐的实验室氛围，使我有一个轻松舒适的学习环境。同时还要感谢室友在生活的照应，使枯燥的科研生活有了一丝乐趣。

最后，我要感谢父母和家人，感谢父母二十多年来的养育之恩，感谢你们无时无刻的关怀和疏导，有你们在背后支持，我才能够勇往直前。对于你们无私的爱与付出，我将会用一生去回报。

最后的最后，向参加评阅论文、答辩的专家和老师表示感谢，感谢你们百忙之中抽出时间，感谢你们的宝贵意见。

在学期间的研究成果及发表的学术论文情况

攻读硕士期间发表（录用）论文情况

1. 任国伟, 谭晓阳. 一种基于强化学习的人脸编辑方法. 第十八届中国机器学习会议（录用）

攻读博士学位期间参加科研项目情况

1. 国家自然科学基金（61976115，61732006）
2. 南航人工智能+项目（NZ2020012，56XZA18009）
3. 全军共用信息系统装备预研（50912040302）