密级：_____

南昌大学
NANCHANG UNIVERSITY

学士学位论文
THESIS OF BACHELOR

（2023 — 2024 年）

题　目：基于图像语义解耦技术的人脸图像检测应用

学　　院：计算机 II 类　系：软件学院

专业班级：软件工程 2011 班

学生姓名：蒋涛　学号：8008120306

指导教师：丁峰　职称：副教授

起讫日期：2023.12.31-2024.3.28

南昌大学

学士学位论文原创性申明

本人郑重申明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式表明。本人完全意识到本申明的法律后果由本人承担。

作者签名：　　日期：

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权南昌大学可以将本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和汇编本学位论文。

本学位论文属于　保密☐，在　　年解密后适用本授权书。
　　不保密☐。

（请在以上相应方框内打“✓”）

作者签名：　　日期：

导师签名：　　日期：

基于图像语义解耦技术的人脸图像检测应用

专　　业：软件工程　　学　　号：8008120306
学生姓名：蒋涛　　指导教师：丁峰

摘　要

在当前的国际形势中，如何保障网络空间是一项极其重要的任务。而随着抖音、快手等短视频平台的兴起，视频信息内容安全受到了国家高度重视。由于直接分析视频信息对计算资源要求极大，往往相关任务是通过视频帧/图像来进行。如何对图像信息进行语义解耦，可以极大地简化图像语义分析、内容识别等。本设计要求落地于人脸图像，设计一种人脸语义信息解耦算法，能够提取出人脸语义信息并初步实现语义替换、合成、修改、删除等功能，并将这个算法封装成具备一个可展示具体分析结果并出具分析报告的应用小程序。

关键词：deepfake 检测；深度学习；网络安全，人工智能

Face Detection of Image Semantic Disentangle Technology

Abstract

In the current international situation, how to ensure cyberspace is an extremely important task. With the rise of short video platforms such as TikTok and Kuaishou, the security of video information content has been highly valued by the country. Since direct analysis of video information requires great computational resources, related tasks are often carried out through video frames / images. How to semantically couple image information can greatly simplify image semantic analysis and content recognition. This design requires landing on the face image, designing a face semantic information disentangle algorithm, which can extract the face semantic information and initially realize the functions of semantic replacement, synthesis, modification, deletion and so on, and encapsuate the algorithm into an application small program that can show the specific analysis results and issue the analysis report.

Keywords: deepfake detection; deep learning; network security; artificial intelligence

目 录

第 1 章 引言

1.1 介绍

近年来，人脸伪造生成方法取得了相当大的进展[1] [2] [3] [4]。由于深度学习的成功，生成超逼真的假人脸图像或视频变得越来越容易。攻击者可以利用这些技术制作假新闻、诽谤名人或破坏身份验证，导致严重的政治、社会和安全后果[5]。为了减轻人脸伪造的恶意滥用，迫切需要开发有效的检测方法。

早期的人脸伪造检测方法[6] [7] [8] [9] [10] [11]通常遵循学习卷积神经网络进行图像分类的经典想法：使用现成的卷积神经网络 (CNN) 主干，这些方法直接将人脸图像作为输入，然后将其分类为真实或虚假。这些普通的 CNN 架构在数据集内评估（也就是使用相同的数据集进行训练和测试）中非常高的性能，但在交叉数据集评估中得到了十分差的结果：这意味着普通 CNN 架构的弱泛化能力[12]。为了解决这个问题，最近的工作求助于特定的伪造模式（如噪声特征[13] [14]、局部纹理[15] [16] [17]和频率信息[18] [19]）或者自动数据增强[20]来微调模型，以更好地检测驻留在假人脸中的伪影。尽管提升效果显著，但它们总是依赖于特定训练集中呈现的某些技术缺陷所造成的伪造模式来判定真假，因此在现实场景中由于伪造技术的迭代升级，具有未知模式的伪造容易导致现有方法失败。

为了解决上述问题，过去有不少工作从两个主要因素来增强人脸伪造检测的学习表示。首先是聚类问题，众多模型通过探索真实人脸和伪造人脸的共同特征并将其聚类的方法来分离出已知的伪造模式，而且通过对真样本的学习来泛化模型检测未知伪造模式的性能，这点已经有过去的研究[21] [22]表明真实样本自身已经具有相对紧凑的易于聚类的分布，如果使用真实图像学习的紧凑表示更有可能将未知的伪造模式与真实人脸区分开来。

其次，正如之前所提及的，随着伪造技术的迭代升级，模型检测伪造的难度越来越高。为了确保学习到的表示捕获真实图像和虚假图像之间的本质差异，增强对伪造线索的查找能力成为了一个主要发展趋势。最近的工作已有此类尝试，通过引入多尺度结构[23]、自注意力机制[24]来识别自顶向下从全局到局部的伪造痕迹。同样地，虽然实验结果确实得到了显著提升，但是

也潜在造成了一些负面效果，比如破坏了检测公平性（过去的研究试图通过分布鲁棒优化技术[25]来缓解由此带来的对性别人种特征过度识别问题），使得模型识别能力事倍功半。

以上的种种现有模型不足都表明，从实际上理解伪造痕迹的产生机制是必要的，因为恰恰是仅从工程角度去思考如何提出新模型的思维教条，给模型带来了一些本可以避免的检测缺陷。基于此，笔者在前人工作上以变分自编码器（VAE）为主干的聚类生成任务为出发点，来总结过去模型的构建思路和由此带来的各种缺点。除此之外，笔者试图在最优传输理论的启发下，以传统拓扑学对生成模型（特别是以 GAN[26]为主的对抗生成模型）如何产生伪造痕迹的问题进行合理性的解释。

基于上述想法，笔者提出以传统深度学习和具有潜在发展前景的几何学习相结合的 ACG 模型（Autoencoder-Classification Geometric model）来优化模型对于伪造检测任务的执行效率。简而言之，本文主要贡献如下:

1. 从一个新的角度分析伪造痕迹的产生原因，并提出用于人脸伪造检测的 ACG 模型①，该模型相对于前人工作更能将原本有限的算力资源充分用于对伪造痕迹的检测上。
2. 在基准数据集（包括 Celeb-DF[27]、WildDeepfake[11]）上进行的大量实验验证了所提出的方法优于最先进的方法，而且证明了 ACG 模型在少样本学习上的优越性。

1.2 相关工作

变分自编码器 (VAE)：根据一系列以经典概率论和信息论为主的数学推导②，VAE[28]在聚类任务中最终的损失函数如下所示：

$$\mathbb{E}_{x\sim p(x)}\Big[-\log q(x|z)+\sum_{y}p(y|z)\log\frac{p(z|x)}{q(z|y)}+KL\big(p(y|z)\big\|q(y)\big)\Big],\quad z\sim p(z|x) \tag{1.1}$$

其中 z 是一个连续变量，代表编码向量；y 是离散的变量，代表类别；p 代表数据的概率分布；q 代表模型所拟合的概率分布。方括号中的三项损失函数，各有各的含义：

① 该模型开源地址为：https://github.com/j2kevin18/ACG/tree/main

② 具体推导见附录 A.1。

1. $-\log q(x|z)$，可以被叫做重构误差，我们希望重构误差越小越好，也就是 z 尽量保留完整的信息；
2. $\sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)}$，可以被叫做聚类误差，希望 z 能尽量对齐某个类别的“专属”的正态分布，通过这一步起到聚类的作用，所以这个误差经常是工作重点；
3. $KL\big(p(y|z)\big\|q(y)\big)$ 可以被叫做度量误差，我们希望每个类的分布尽量均衡，不会发生两个几乎重合的情况（即坍缩为一个类）。

其中聚类误差经常用交叉熵损失来表示：

$$\mathscr{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + \big(1-y_i\big)\log\big(1-\hat{y}_i\big) \tag{1.2}$$

而重构误差随着对比学习的兴起开始对该误差进行软约束，以期获得一定的泛化效果，比如下面就是一个常用损失函数：

$$D_w = \|\hat{x} - x\|_2 \tag{1.3}$$

$$\mathscr{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N} y_i \big(D_w\big)^2 + \big(1-y_i\big)\max\big(m - D_w, 0\big)^2 \tag{1.4}$$

对于度量误差，大多数工作出于前面两者的聚类效果已经足够优秀，所以并没有将这一项考虑进去。不过即使如此，仍然有工作[29]指出使用度量误差可以使模型不需要通过对伪造样本进行重构也可以带来性能上的提升，更有团队[30]通过变分信息瓶颈理论（VIB）加入 VIB 模块[①]，让模型去挖掘那些对聚类相对关键的信息，而不是冗余信息，也同样取得了满意的结果。

通过复盘 VAE 的成功，我们可以知道是因为标准正态分布相对规整，均有零均值、标准方差等好处，但更重要的是标准正态分布拥有一个非常有价值的特点：它的每个分量是解耦的，用概率的话说，就是相互独立的，满足 $p(x,y) = p(x)p(y)$。经过我们之前的讨论可以知道，如果特征相互独立的话，建模就会容易得多（朴素贝叶斯分类器就是完全精确的模型），而且相互独立的特征可解释性也好很多，因此我们总希望特征相互独立。早在 1992 年 LSTM 之父 Schmidhuber 就提出了 PM 模型（Predictability Minimization）[31]，致力于构建一个特征解耦的自编码器。

所以在模型中下意识的对样本进行解耦是有效的，比如在 deepfake 检测

① 至于度量误差与瓶颈理论的关系，具体推导见附录 A.2。

的相关工作中[32]，作者让模型提取出图片背景特征和人脸特征来排除人脸取证中背景的干扰。可是我们知道人类可理解的范围是有限的，通过语义进行人为的伪造痕迹划分相对于技术迭代来说如同杯水车薪，所以我会在下面的内容详谈如何让模型自主地解耦有关的伪造特征。

模型坍塌与模型混淆：模型坍塌与模型混淆一直是萦绕在深度学习领域的两朵乌云：在计算机视觉领域中，GANs 模型就已经被验证出现了严重的模型坍塌问题[33]。过去的研究已经做出各种试图缓解模型坍塌与模型混淆的工作，比如梯度正则化[34]、多模型混合[35]，也确实取得了一些实质性的成果，然而他们都不能很好地解决相关问题。在自然语言处理领域中，模型坍塌与模型混淆带来的是灾难性遗忘问题[36]和模型幻觉问题[37]，成为了大语言模型发展最大的拦路虎。

模型坍塌与模型混淆迟迟不能解决的原因是大多数工作在设计模型时忽略了深度学习自身的缺陷，经常基于连续能量分布的预设[38]理解模型坍塌与模型混淆现象。如图 1-1（a）所示，我们把真实样本 $x_1, x_2, \ldots, x_n$ 看成一个个在数据空间的坐标，这些真实样本的分布可以用一个能量函数 $E(x)$ 描述，使得真实样本 $x_1, x_2, \ldots, x_n$ 都处于 $E(x)$ 的极小值点，在示意图中可以很形象地看成“坑”，然后我们再把伪造样本 $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ 放到“坑腰”，并把假样本“松开”让其滚入坑底，这就完成了伪造样本对真实样本的拟合。

在能量角度中，一个生成模型的最优解需要它满足两个条件：

1. 需要把假样本 $\hat{x}$ “放”在合适的位置。如果一旦出现初始化不好、优化不够合理等原因，使得 $\hat{x}$ 同时聚在个别坑附近（如图 1-1（b）所示），这样所有伪造样本都往个别坑奔了，所以模型只能生成个别样式的样本，于是便出现了模型坍塌。
2. 假样本 $\hat{x}$ 不能被“困”在局部优化点上，使得模型生成除真实数据分布外的其他不合理样本（如图 1-1（c）所示），于是便出现了模型混淆。

由此可见，基于能量观点的模型坍塌与模型混淆的理论十分清晰直白，之前所提及的工作也是围绕着这个假设去缓解生成模型的模型坍塌与模型混淆问题，甚至可以说由于此类问题的出现给伪造图像检测有了可乘之机。但是这类假设的关键问题在于，将它的理论根基建立在能量分布函数本身是“连续”的前提，忽视了隐空间中可能存在的奇异集合（出现在初等函数里就是“不可导”现象）会导致模型优化的失效，进而缓解模型坍塌与模型混淆的传统方法也会随之失效。先前 Nagarajan & Kolter (2017)[34]；Khayatkhoei 等人

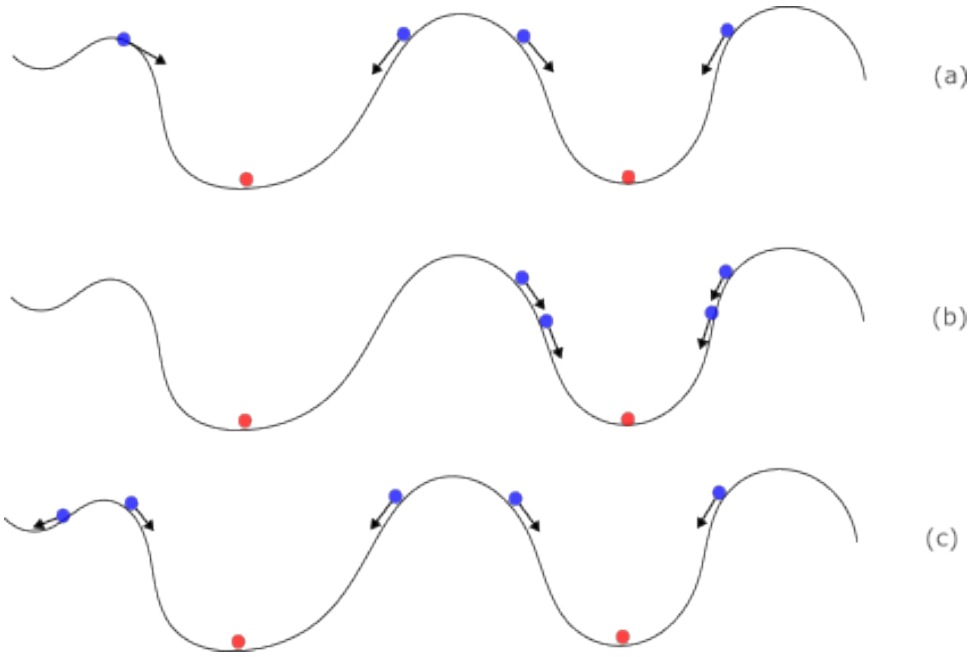


图 1-1: 基于能量观点的模型坍塌与模型混淆示意图，图中曲线表示数据上的能量分布，红圆点表示真实数据，蓝圆点表示生成模型生成出的数据，实心箭头表示原点所在的梯度。(a) 最优模型；(b) 模型坍塌；(c) 模型混淆。

(2018)[35]的工作已经指出，当数据分布中存在多种分布状态（在能量模型中体现为 $E(x)$ 有多个最优的极小值点）时，需要学习的传输映射可能是不连续的；而且即便是只有单个分布状态，顾险峰等人[39]的研究指出仍有可能在生成数据分布中出现奇异点，根据对于这样的问题，笔者会从下面的拓扑学角度重新诠释出现的新情况。

流形分布假设: 流形分布假设在深度学习领域中并不是新近的理论，而是一直伴随着深度学习发展的基石。在深度学习中，流形分布假设[40]被很好地接受: 假设特定类别的自然数据分布集中在嵌入在高维数据空间中的低维流形上。

回到图像生成应用中，如果仅仅讨论 GAN 模型，本质上就是将一种固定的概率分布（例如均匀分布或者高斯分布），变换成训练数据所蕴含的概率分布，例如人脸图像的分布。我们通过一个例子去理解这个问题。

如图 1-2 所示[41]，GAN 的理想数学模型如下: 我们将所有 $n \times n$ 图像构成一个空间，记为图像空间 X，每一张图像看成是空间中的一个点，$x \in X$ 。我们用 $\nu(x)$ 来表示图片 x 是否表达一张人脸的概率，那么 ν 就是 GAN 要学习的

① 具体解释见附录 B.1。

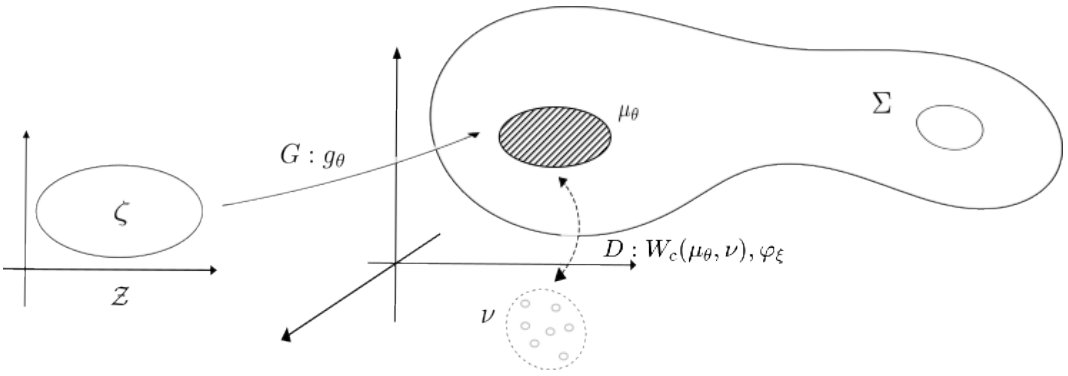


图 1-2: WGAN 的理论框架示意图[①]

目标概率测度。在工程实践中，我们只有一些人脸图像的样本 $\{y_1, y_2, \ldots, y_n\}$，这些样本构成了经验分布作为 ν 的近似。经验分布的公式表达为（其中 δ 为狄拉克分布）：

$$\nu = \frac{1}{n}\sum_{i=1}^{n}\delta\left(x - y_i\right) \tag{1.5}$$

绝大多数图片并不是人脸图像，因此 ν 的支撑集合

$$\Sigma(\nu) := \{x \in X \mid \nu(x) > 0\} \tag{1.6}$$

是图像空间中的一个子流形，Σ 的维数远远小于图像空间 X 的维数。支撑集流形 Σ 的参数空间等价于特征空间，或者隐空间 Z。编码映射 (encoding map) 就是将 Σ 映到特征空间，解码映射 (decoding map) 就是将特征空间映到支撑集流形 Σ。

假设在隐空间有一个固定的概率分布 $\zeta \in P(Z)$，例如高斯分布或者均匀分布。我们用一个深度神经网络 θ 来逼近解码映射 $g_\theta : Z \to X$，g_θ 将 ζ 映成了图像空间中的概率分布 $\mu_\theta := \left(g_\theta\right)_* \zeta$，我们称 μ_θ 为生成分布。

这样就可以知道，以 GANs 为主流的生成模型其实是在做两件事情：

1. 训练判别器 D，核心任务是计算训练数据分布 ν 和生成分布 μ_θ 之间的距离；
2. 训练生成器 G，目的在于调节 g_θ 使得生成分布 μ_θ 尽量接近数据分布 ν。

这里我们讨论的仍然是理想情况，因为最终的生成分布 μ_θ 也存在是非连续集的情况。如图 1-3 所示，假设生成分布 μ_θ 是个非凸集，A、B 是概率分布 ζ 上的两个采样点，而且我们沿着道路 AB 进行取样[②]，得到在奇异集合上的

② 在过去的工作中[42]，Radford 等人发现 GANs 模型生成的人脸样本有着类似在词向量的向量算术

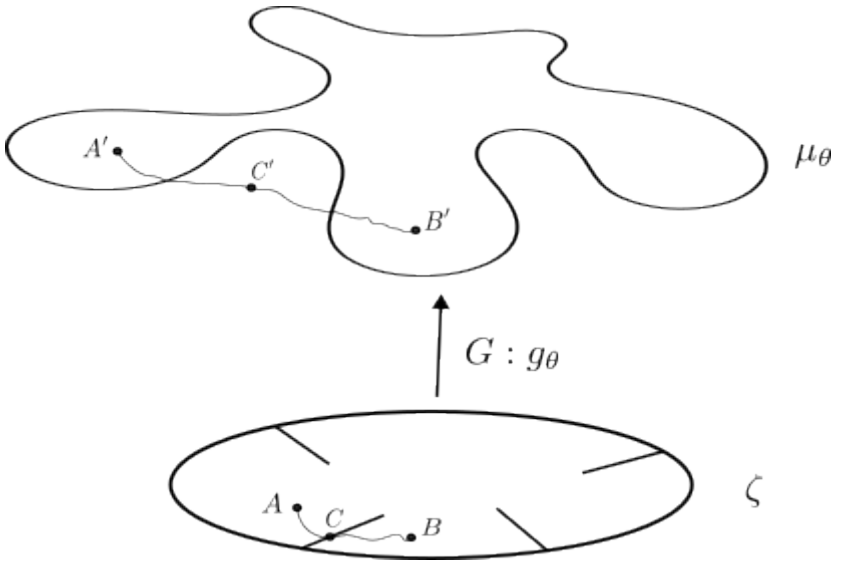


图 1-3: 生成分布 μ_θ 非连续集情况示意图。

点 C。经过 g_θ 映射后，由于 μ_θ 不是连续集，导致解码映射把点映到连续集之中（如点 A'、B'），或者映到连续集间的无意义区域（如点 C'），导致生成图像出现明显的伪造痕迹（如图 1-4）。

基于此，我们得出人脸伪造痕迹的经验性定则:

性质。要满足这个性质，意味着生成分布 μ_θ 不能是非连续集，否则会导致无法满足向量运算的封闭性质。

图 1-4: 模型学习 CelebA 数据集后通过插值法所生成的图片（出自[43]），第三行的人脸一只眼睛为蓝色，另外一只眼睛为棕色——这在生物学上是极其罕见的事情。

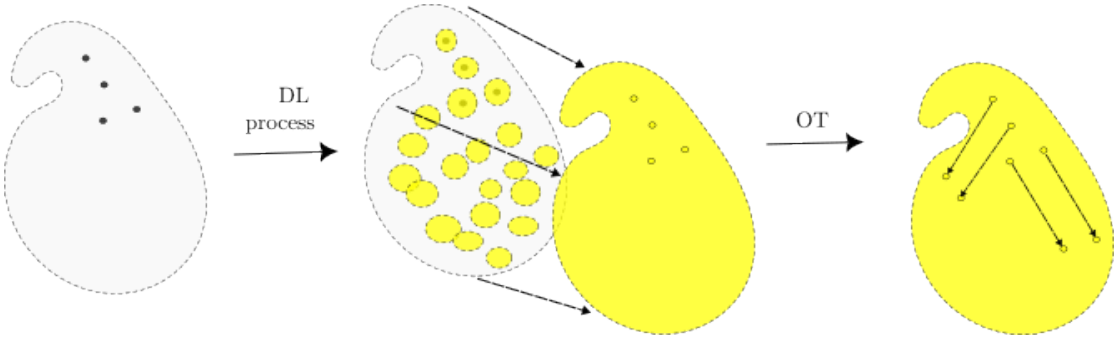


图 1-5: ACG 模型训练示意图。

定律 1.1 人脸伪造痕迹是隐空间映射到生成分布出现的奇异集合导致的。

定律 1.2 隐空间映射到生成分布出现的奇异点越少，人脸伪造痕迹越不明显。

定律 1.3 人脸伪造检测就是减少样本中存在的奇异点个数，让基于真实分布生成的重构图片更为真实，从而增加识别准确度。

基于上述定则，我们可以知道 ACG 模型需要尽可能减少两类奇异点（如图 1-5）：

1. 在训练数据流形到低维隐空间的编码映射时，经过 DNN 拟合的生成分布可能并没有与真实隐空间的分布拟合好，导致出现一系列奇异点（在拓扑数据挖掘领域里[44]早已经发现，并称之为噪声）。所以这一步需要深度学习来进行拟合。
2. 在训练从数据白噪声到隐空间的映射（一般被称为最优传输映射）时，由于深度学习自身的缺陷，导致非凸的真实数据分布映射后的隐空间是不连续的状态，从而产生奇异点，这一步需要以最优传输理论为基础的几何学习来进行训练。

最优传输理论（OT）[45]：最优传输理论是一个由 Gaspard Monge 和经济学家 Leonid Kantorovich 的工作奠基的数学领域。自从它的诞生以来，这一理论在数学、物理学和计算机科学方面做出了重大贡献，在逐渐成为机器学习社区中一个不断发展的研究课题。比如，OTFusion[46]提出了一种新的方法，利用最优传输理论来对齐并融合两个或多个 Transformer 模型；AE-OT[39]、AE-OT-GAN[47]等工作也在利用最优传输理论来生成更高质量的图片。随着新算法的不断提出，在可见的未来，我们可以见到 OT 在大模型的广泛应用。

OT 的核心是研究如何使用最经济的方式，将一种分布变换到另一种分布。举个简单的例子，假设有两个沙堆，一个沙堆如何通过最小的代价，变成和另一个沙堆一样的形状? 这就是最优传输问题，而在深度学习的领域中我们需要讨论的是如何将连续的白噪声分布映射到非连续的隐空间中。

假设隐空间中的向量是 z，且白噪声空间中的向量为 x，我们就可以定义一个衡量 x 和 z 相邻程度的度量:

$$F(x) = max(< x, z_i > + h_i) \tag{1.7}$$

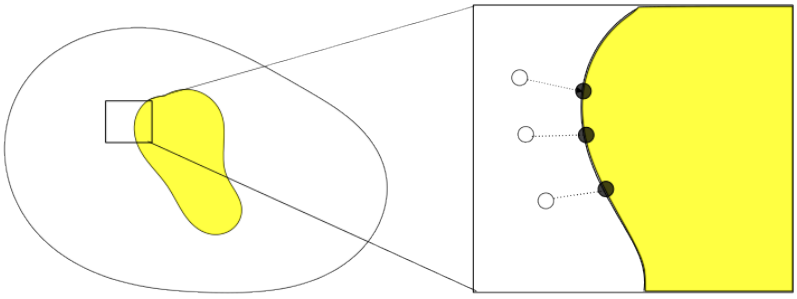
其中 h_i 是可以训练的偏置，且 $\sum_{i=1}^{n} h_i = 0$。从几何意义上讲，就是建立了一个从 z 在白噪声空间的邻域 $U(z)$ 到 z 的双射。

同时我们希望 z 在白噪声空间的邻域 $U(z)$ 的体积 $vol(h)$ 可以被人为均衡，不妨设 $(v_1, v_2, \ldots, v_n)$ 为人为均衡后的体积，我们就可以得到以下的损失函数:

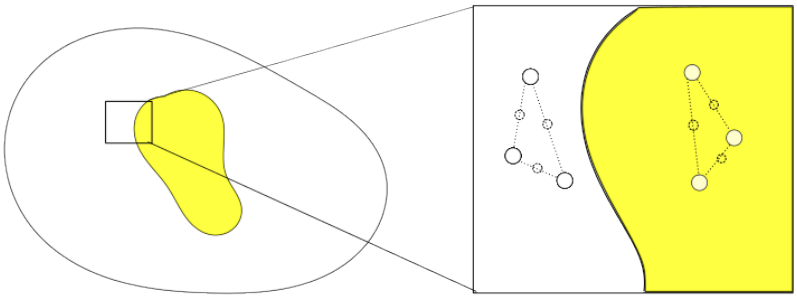
$$\sum_{i=1}^{n} \int_{0}^{h} (vol(h) - v_i) d h_i \tag{1.8}$$

在 ACG 模型中，我们设计的 OT 模块可以通过优化这个损失函数，让模型得到的特征解耦相对于语义解耦的方法更充分。以人脸提取的关于肤色的特征举例，传统的语义解耦作用于整个数据流形的情况如图 1-6(a) 所示。一般认为，一个良好的伪造检测器应该不受各种图片的额外特征影响，但是实验中训练的各种检测器大都会因此发生一些模型坍塌或者模型混淆。如果我们把这些特征看作是数据的背景噪声，其实不难看出这是一个数据去噪的过程。如图所示，真实数据由于各种外部特征的扰动而与真实数据流形产生了一定距离，那么经过外部解耦的特征剔除后，这些数据又会在一定程度上回归真实数据分布，从而缓解一部分的模型失真问题。但是这类方法的缺点也显然是随机性过强，缺乏对于数据分布自身的特征分析。

我们提出的从 OT 中衍生出的内部解耦算法就一定程度上解决了这个问题，情况如图 1-6(b) 所示。在内部解耦算法中，会通过计算出的各个隐空间中的向量 z 中的邻域 $U(z)$ 得到特征之间的相关度，然后同类样本之间以线性插值的方法来生成新的特征。而且这个工作不是本文首次提出，比如原型网络[48]就试图通过类似的方法来实现少样本学习，从侧面也应证了该方法强大的泛化性。



(a) 外部解耦



(b) 内部解耦

图 1-6: 外部解耦和内部解耦示意图，其中黄色区域代表真实数据流形，外面的白色区域代表伪造数据流形。

第 2 章　具体方法

为了捕捉真实人脸和虚假人脸之间的本质差异，笔者设计了一个名为 ACG 的新框架，该框架以 Xception[49] 为骨干网络，由自编码器、几何学习和归类卷积网络三个主要部分组成，如图 2-1 所示。自编码器不仅对真实人脸图像和伪造人脸图像的分布进行充分编码外，还对真实人脸图像的样本点进行度量上的汇聚，以完成一定程度上的聚类。此外，为了避免隐空间上产生奇异点，几何学习方案以最优传输理论为基础探测奇异点并对其消除重构。同时，归类卷积网络完成最后的分类任务。以下小节详细描述了三个部分是如何训练的。

2.1　自编码器训练和微调

由于人脸伪造方法总是多种多样的，所以探索真实人脸的共同特征比过拟合训练集中呈现的特定伪造模式更适合。因此，本文建议先进行全样本的重构学习，通过恢复人脸图像来促进编码器对输入图片的充分编码表示。

具体来说，给定一个输入图像 $X \in \mathbb{R}^{h \times w \times 3}$，我们基于编码器-解码器结构训练自编码器 F。由于先前的研究[50]已经证明，对原始输入的样本重建不会

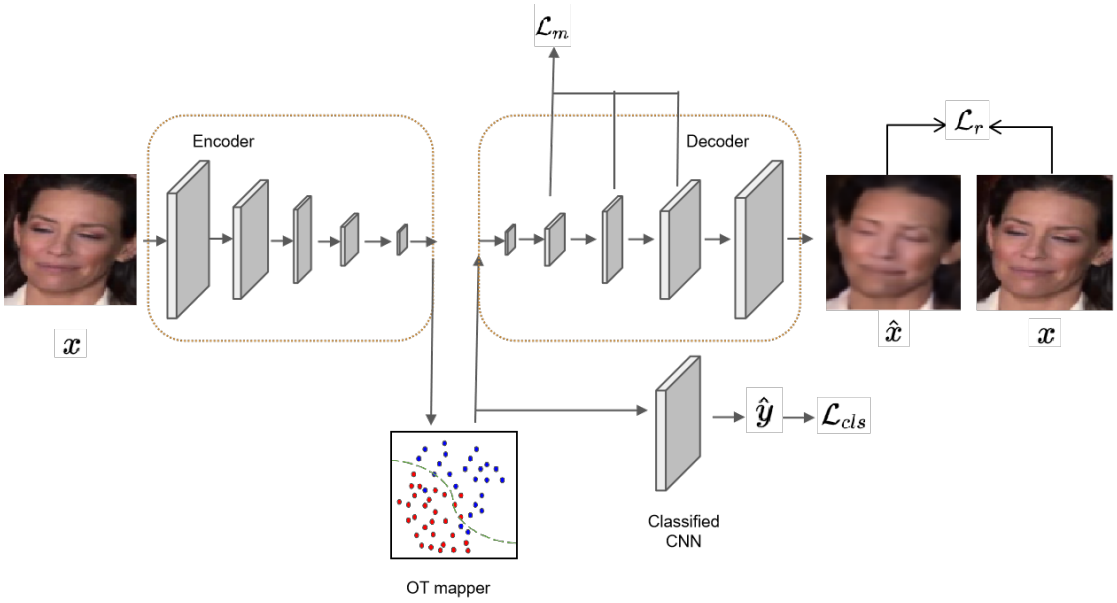


图 2-1：ACG 模型架构示意图。

显著改善学习的编码能力；而且关于伪造检测模型性能可解释性的工作 [] 也证明，对图像像素点之间的低阶交互的过度拟合会极大影响模型识别能力，所以应该在训练期间向输入样本 X 添加一些白噪声以获得 $\widetilde{X}$，旨在学习人脸的稳定表示。因此，图像重建过程可以表述为:

$$\hat{X} = F(\widetilde{X}) \tag{2.1}$$

。在重构过程中，我们计算输入图像与其重构图像之间的重构损失 L_r 为:

$$\mathscr{L}_{\mathrm{r}} = \left\| \hat{\mathbf{X}} - \mathbf{X} \right\|_1 \tag{2.2}$$

。由于对人脸图像点对点的重建，自编码器 AE 的学习到的生成分布得以拟合有意义的人脸图像分布，不过真实图像分布和伪造图像分布之间仍有可能混杂在一起，从而导致真实图像和虚假图像无法区分。

由此，我们需要单独微调解码器来对两个不同模式的样本进行聚类。受到该工作[29]中重构学习架构的启发，令 $F \in \mathbb{R}^{h' \times w' \times c}$ 表示编码器或解码器块的输出特征，并将全局平均池化操作应用于 F 并获得每个输入样本的特征向量 $\overline{F} \in \mathbb{R}^c$，则有度量误差 L_m:

$$\mathscr{L}_{\mathrm{m}} = \frac{1}{N_{RR}} \sum_{i \in R, j \in R} d\left(\overline{\mathbf{F}}_i, \overline{\mathbf{F}}_j\right) - \frac{1}{N_{RF}} \sum_{i \in R, j \in F} d\left(\overline{\mathbf{F}}_i, \overline{\mathbf{F}}_j\right) \tag{2.3}$$

其中 R, F 表示真假样本集。N_{RR} 和 N_{RF} 分别是（真, 真）对和（真, 假）对的总数。$d(\cdot,\cdot)$ 是基于余弦距离的成对距离函数[①]:

$$d(\mathbf{a}, \mathbf{b}) = \frac{1 - \frac{\mathbf{a}}{\|\mathbf{a}\|_2} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|_2}}{2} \tag{2.4}$$

L_m 的第一部分鼓励从真实样本中学习紧凑的表示，而第二部分强调真实样本和虚假样本之间的差异。而且我们在这一过程中只对真实样本进行重构，则有重构损失 L_r 为:

$$\mathscr{L}_{\mathrm{r}} = \frac{1}{|R|} \sum_{i \in R} \left\| \hat{\mathbf{X}}_i - \mathbf{X}_i \right\|_1 \tag{2.5}$$

其中 R 表示真实样本集，$|R|$ 是真实样本集个数。

① 在 A.3 中会有对该公式的详细推导。

2.2 针对特征的几何学习

回想前文内容，正是由于隐空间中奇异集合的存在导致人脸出现伪造痕迹，结合具体工作[33]来说，由于奇异集合随着模型各层的传播，导致诸如 GANs 的生成器整体能生成图片的范围不如其中单个层能生成图片的范围。可以预见，不进行几何学习修正的编码器生成的图片由此也会带有严重的伪造痕迹，从而造成严重误判，损害模型泛化性能。

基于顾险峰团队在 AE-OT 上[39]提出的算法，几何学习在自编码器训练完成后分为三个阶段：

1. 提取特征。从编码器生成的隐空间 Z 中按批次提取被压缩的图像向量，作为图像经编码器提取到的特征嵌入分布。
2. 最优传输计算。采用半离散最优传输方法，构建出从标准噪声分布 ζ 到图像嵌入分布 Z 之间的映射 T，实现从噪声到具体嵌入的变换过程。
3. 生成特征。经过上述两个过程，我们可以通过以下映射实现图像生成

$$F_{geo} : T \cdot f(\zeta) \tag{2.6}$$

其中 f 为编码器，但这一过程无法生成图像样本之外的其他数据。所以为了拓展这个最优传输，该过程使用类似插值的思想实现从噪声分布到生成图像的映射，并避免奇点在生成特征过程中产生。

2.3 归类卷积网络

过去各种研究者已经展示了标准的 CNN 架构如何有效的检测伪造痕迹，比如使用真假图片对训练利用 CNN 构建的孪生网络[6]等等，表明经典的 CNN 模型可以有效的检测到这些伪影的存在，但是 CNN 只能检测出在训练集中出现过的伪造方式，不具有泛化性。我们结合上述的自编码器、几何学习的特点，采用经典 CNN 模型 xeception 网络来完成归类任务，并以及采用二元分类的交叉熵损失 L_{cls}。

2.4 训练损失函数

所提出框架的总损失函数 L 包括重建损失和用于重建学习的度量学习损失，以及用于二元分类的交叉熵损失 L_{cls}：

$$L = L_{cls} + \lambda_1 L_r + \lambda_2 L_m, \tag{2.7}$$

其中 λ_1 和 λ_2 是平衡不同损失的权重参数 L_{cls}。

第 3 章 实验方法

3.1 实验设置

数据集。由于算力条件限制，本文仅在 Celeb-DF[27]、WildDeepfake (WDF)[11] 上评估了上文提出的方法和现有方法。Celeb-DF 包括 590 个真实视频和 5,639 个高质量的假视频，由改进的 DeepFake 算法制作。WildDeepfake 是一个来源于真实世界的数据集，包含 3,805 个真实序列和 3,509 个假序列。它中的所有视频都是从互联网上获得的，具有更多的身份出现在各种场景中，具有现实应用意义。

评估指标。为了评估新方法，本文选择性报告了相关工作中最常用的指标，包括准确度 (Acc)、接收器工作特征曲线下面积 (AUC) 和错误率 (EER)。

实施细节。我们所提出的框架以 32 的批大小对其进行训练，Adam[51] 优化器初始学习率为 2e-4，权重衰减为 1e-5。步骤学习率调度器用于调整学习率。本文只使用随机水平翻转进行数据增强。对于训练集，由于原 Celeb-DF 数据集有 9GB 左右，为了节约内存，我们对视频分别抽 5 帧，这样训练集仅含有 2590 张图片，内存占用总共不到 100MB。

3.2 实验结果

数据内评估。在本节中，我们将提出的方法与当前最先进的方法进行了比较。如表 3-1 所示，对于 Celeb-DF 数据集，我们的方法在现有方法上取得了很大的改进。ACG 在具有挑战性的 Celeb-DF 数据集判别任务下，与基准模型 RECCE[29] 相比，本文方法的 AUC 得分超过了 0.05%，这可以认为这是基准模型在传统深度学习方法下解耦不足的表现，而我们的方法通过几何学习产生更稳健的表示，使其作为伪造分类的有效指导，因此我们的方法明显优于同类方法。与其他方法相比，在 Celeb-DF 上也可以观察到明显的性能提升，比如我们的方法将 Acc 提高了 1.29%，上述结果初步证明了所提出的 ACG 框架的有效性。

数据外评估。为了评估我们的方法在未知伪造上的泛化能力，我们通过对不同的数据集进行训练和测试来进行跨数据集实验。具体来说，我们在 Celeb-DF 上训练模型，然后在 WildeDeepfake 上测试它们。结果如表 3-2 所示。从表

表 3-1: 数据内评估，可以验证所提出的方法优于当前最先进的方法。

检测方法	Celeb-DF	
	ACC(%)	AUC(%)
Xception[49]	97.90	99.73
RFM[12]	97.96	99.94
Add-Net[11]	96.93	99.55
F^3-Net[19]	95.95	98.93
MultiAtt[17]	97.92	99.94
RECCE[29]	98.59	99.94
ACG(ours)	**99.88**	**99.99**

表 3-2: 在 AUC(%) 和 EER(%) 指标上的交叉测试。

检测方法	训练集	测试集（WDF）	
		AUC(%)↑	Err(%)↓
Xception[49]	FF++	62.72	40.65
RFM[12]	FF++	57.75	45.45
Add-Net[11]	FF++	62.35	41.42
F^3-Net[19]	FF++	57.10	45.12
MultiAtt[17]	FF++	59.74	43.73
RECCE[29]	FF++	64.31	40.53
ACG(ours)	Celeb-DF	**77.58**	**28.85**

中，我们观察到 ACG 在不可见的测试数据上通常优于所有列出的方法，通常有很大的优势。例如，在 WildDeepfake 数据集上进行测试时，大多数先前方法的 AUC 分数下降到大约 60%。不同的是，ACG 的 AUC 为 77.58%，超过 RECCE[29] 13.27%。性能主要得益于所提出的 ACG 框架不像现有方法那样无法准确拟合伪造模式，而该框架只对真实人脸的分布进行建模，且 OT 的内部解耦算法会引导模型学习真实人脸和虚假人脸之间的本质差异，以实现更好的泛化性。以上结果表明，探索真实人脸的共同特征来区分具有未知模式的伪造是可行的，而且验证了检测模型在少样本学习上的可行性。

重构可视化。为了直观地理解重构学习，我们可视化了重构网络的输出和原始输入之间的差异，如图 3-1 所示。我们可以看到，真实人脸可以用很

图 3-1: 在 Celeb-DF[27] 数据集上重建所提方法的可视化。第一行和第二行分别显示了输入图像和重建结果。

少的模糊很好地重建，而假人脸的伪造区域无法恢复，表明即使我们的方法只在图像级监督下进行训练，伪造区域的痕迹依旧是可能的。这说明对于高质量的 Celeb-DF 数据集，虽然源和操作方法仍然未知，但我们的方法仍然可以指示可能的伪造区域。可视化验证了所提出的框架可以有效地捕捉真实人脸和虚假人脸之间的本质差异。

分类决策分析。为了更好地理解我们方法的决策机制，我们在图 3-2 中提供了 Celeb-DF 上的 Grad-CAM[52] 可视化。我们观察到，无论面部真实性如何，基线方法主要集中在图像的中心区域进行分类，缺乏对不同伪造的全面理解。不同的是，我们的方法为真假人脸生成可区分的热图，其中突出的区域在伪造技术上有所不同，即使它只使用二进制标签进行训练。例如，假样本的热图都集中在主要面部区域，而对于真样本的热图则没有集中的迹象，结果从决策的角度解释了 ACG 的有效性。

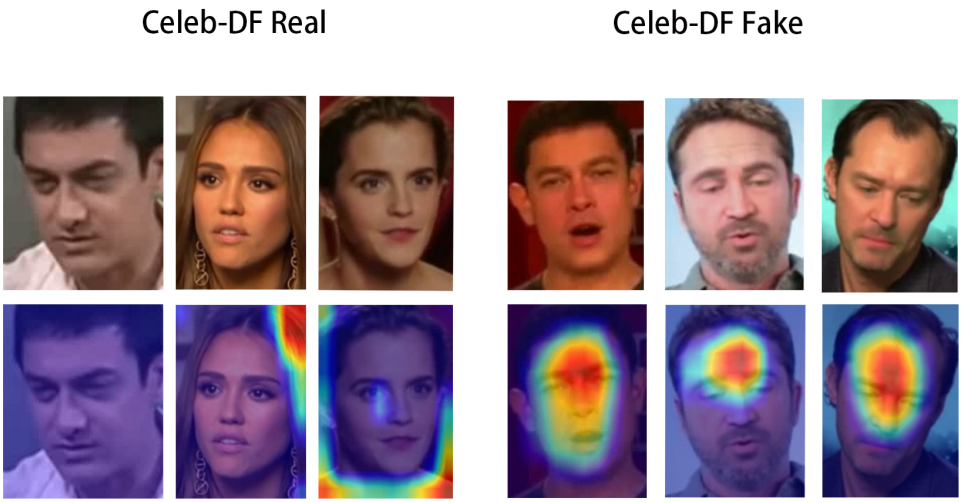


图 3-2: Grad-CAM[52] 可视化

第 4 章 结论

这项工作给出了传统拓扑理论对模式崩溃/混合的理论解释：OT 映射在数据集上不连续但 DNN 只能表示连续函数的冲突会导致以 GANs 为主流的传统生成模型上生成的图片存在明显伪造痕迹。基于这个问题，通过 AE 与几何学习的融合，提出了 ACG 模型来检测伪造人脸，且该模型训练时同时使用自编码器和用 OT 算法来增强其解耦性能。该模型在相对较少的合成数据集和真实数据集上进行广泛并合理地测试，并且实验结果验证并展示了与最先进的技术相比的优势。

附录A 公式推导

A.1 VAE聚类任务

关于详细证明可以参考苏剑林的工作[53]，里面指出VAE的变分判断其实是基于下面的不等式(这里延续之前的记号规定):

$$\begin{aligned} KL(p(x,z)\|q(x,z)) &= KL(p(x)\|q(x)) + \int p(x)KL(p(z|x)\|q(z|x))dx \\ &\geqslant KL(p(x)\|q(x)) \end{aligned} \tag{A.1}$$

这里的思路是，原本我们希望计算以下的等式:

$$KL(p(x)\|q(x)) = \int p(x)\log\frac{p(x)}{q(x)}dx \tag{A.2}$$

但是等式（A.2）难以计算，所以我们将联合分布作为等式（A.2）的上界，并最小化这个上界。如果我们将隐变量定为(z,y)，则有:

$$KL\Big(p(x,z,y)\Big\|q(x,z,y)\Big) = \sum_y \iint p(z,y|x)p(x)\ln\frac{p(z,y|x)p(x)}{q(x|z,y)q(z,y)}dzdx \tag{A.3}$$

我们在这里假设:

$$p(z,y|x) = p(y|z)p(z|x), \quad q(x|z,y) = q(x|z), \quad q(z,y) = q(z|y)q(y) \tag{A.4}$$

代入等式（A.3），则有:

$$KL\Big(p(x,z,y)\Big\|q(x,z,y)\Big) = \sum_y \iint p(y|z)p(z|x)p(x)\ln\frac{p(y|z)p(z|x)p(x)}{q(x|z)q(z|y)q(y)}dzdx \tag{A.5}$$

这里实际上已经得到了聚类任务的损失函数，描述了编码和生成过程:

1. 从原始数据中采样到x，然后通过$p(z|x)$可以得到编码特征z，然后通过分类器$p(y|z)$对编码特征进行分类，从而得到类别;
2. 从分布$q(y)$中选取一个类别y，然后从分布$q(z|y)$中选取一个随机隐变量z，然后通过生成器$q(x|z)$解码为原始样本。

A.2 信息瓶颈理论

信息瓶颈理论[54]出于一个基本假设：更低的成本、更少的信息可以得到更好的泛化能力，同时也意味着模型能找到一些普适的规律和特性。

比如在分类任务中，标注数据对是$(x_1, y_1), \ldots, (x_N, y_N)$。我们把这个任务分为两步来理解，第一步是编码，第二步就是分类。第一步是把x编码为一个隐变量z，然后分类器把z识别为类别y。然后我们试想在z处加一个“瓶颈”β，作用是不允许流过z的信息量多于β来完成分类任务，所以模型训练时会倾向于让最重要的信息通过瓶颈。

其中最重要的是变分信息瓶颈[55]，变分信息瓶颈最后的损失函数为：

$$\mathbb{E}_{x\sim p(x)}\Big[\mathbb{E}_{z\sim p(z|x)}\big[-\log p(y|z)\big] + \lambda \cdot KL\big(p(z|x)\big\|q(z)\big)\Big] \tag{A.6}$$

这里可以注意到，变分信息瓶颈的损失函数跟等式（1.1）在形式上是十分相似的，特别是$\lambda \cdot KL\big(p(z|x)\big\|q(z)\big)$与度量误差更是如此。

在我们之前所提到的能量模型中，自然希望把假样本$\hat{x}$“放”的位置变得随机化，因为随机化可以使假样本在能量分布的位置变得更均匀，出现样本扎堆的情况也会大大减少（当然也会带来模型混淆的隐患）。因此，我们需要减少类别y与编码特征z之间的互信息，让类别y与编码特征z之间是独立的。

A.3 度量误差

关于度量误差个人目前了解到的最早工作来自[56]，实际上就是狄拉克分布积分近似的结果。总的来说，如下的上采样过程：

$$z \sim q(z), \quad x = G(z) \tag{A.7}$$

实际上就是假设x的分布为：

$$q(x) = \int \delta\big(x - G(z)\big) q(z) dz \tag{A.8}$$

其中$\delta(\cdot)$代表着（多元的）狄拉克函数。注意$q(x)$也可以写成：

$$q(x) = \mathbb{E}_{z\sim q(z)}\big[\delta\big(x - G(z)\big)\big] \tag{A.9}$$

而$\delta(\cdot)$实际上就是方差趋于0的高斯分布：

$$\delta(x) = \lim_{\sigma\to 0} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \tag{A.10}$$

这样一来，我们不妨就让 σ 取个有限值，算完之后再让 $\sigma \to 0$，即

$$q(x) = \lim_{\sigma \to 0} \mathbb{E}_{z \sim q(z)} \left[\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x - G(z)\|^2}{2\sigma^2} \right) \right] \tag{A.11}$$

然后，我们做最大似然，即以 $-\int p(x) log q(x) dx$ 为损失函数，$p(x)$ 是真实样本的分布：

$$\begin{aligned} loss =& -\int p(x) \log \left\{ \mathbb{E}_{z \sim q(z)} \left[\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x - G(z)\|^2}{2\sigma^2} \right) \right] \right\} dx \\ =& \mathbb{E}_{x \sim p(x)} \left[-\log \left\{ \mathbb{E}_{z \sim q(z)} \left[\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x - G(z)\|^2}{2\sigma^2} \right) \right] \right\} \right] \\ \sim& \mathbb{E}_{x \sim p(x)} \left[-\log \left\{ \mathbb{E}_{z \sim q(z)} \left[\exp\left(-\frac{\|x - G(z)\|^2}{2\sigma^2} \right) \right] \right\} \right] \end{aligned} \tag{A.12}$$

在最后一个式子中，我们已经省去了与优化无关的常数。现在我们将 $\mathbb{E}$ 转化为采样，即把 $x_1, x_2, \ldots, x_M \sim p(x)$ 和 $z_1, z_2, \ldots, z_N \sim q(z)$ 代入损失函数：

$$\begin{aligned} loss \sim& -\frac{1}{M} \sum_{i=1}^{M} \log \left\{ \frac{1}{N} \sum_{j=1}^{N} \exp\left(-\frac{\|x_i - G(z_j)\|^2}{2\sigma^2} \right) \right\} \\ \sim& -\frac{1}{M} \sum_{i=1}^{M} \log \left\{ \sum_{j=1}^{N} \exp\left(-\frac{\|x_i - G(z_j)\|^2}{2\sigma^2} \right) \right\} \end{aligned} \tag{A.13}$$

我们可以知道，$logsumexp$（指数、求和、然后取对数）实际上是 max 的光滑近似，当 σ→0 时它就是 max，加上了负号就是 min，所以最终 $\sigma \to 0$ 时的最简形式为：

$$loss \sim \frac{1}{M} \sum_{i=1}^{M} \left(\min_{j=1}^{N} \|x_i - G(z_j)\|^2 \right) \tag{A.14}$$

其实这个思路还可以推广到一般的散度优化，比如我们可以用

$$KL(q(x) \| p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx = \mathbb{E}_{x \sim q(x)} \big[\log q(x) - \log p(x) \big] \tag{A.15}$$

做优化目标，然后 $logq(x)$、$logp(x)$ 按照同样的方法进行处理，那么结果是：

$$loss \sim -\frac{1}{M} \sum_{i=1}^{M} \left(\min_{j=1}^{N} \|G(z_i) - x_j\|^2 - \min_{j=1}^{K} \|G(z_i) - G(z_j)\|^2 \right) \tag{A.16}$$

这里便相当于是用了 L_2 距离的度量误差。

附录 B 最优传输理论概要

定律 B.1: **最优传输理论** 给定带有概率测度的空间 (X,μ) 和 (Y,ν)，具有相同的总质量，$\mu(X)=\nu(Y)$。一个映射 $T: X\to Y$ 被称为是保持测度，如果对于一切可测集合 $B\subset Y$，我们都有

$$\int_{T^{-1}(B)} d\mu(x)=\int_B d\nu(y) \tag{B.1}$$

记为 $T_*\mu=\nu$。给定距离函数 $c(x,y)$，代表两点间的某种距离，传输映射的传输代价函数为:

$$\mathscr{C}(T):=\int_X c(x,T(x))d\mu(x) \tag{B.2}$$

定律 B.2: **蒙日问题** 法国数学家蒙日于 18 世纪提出了最优传输映射问题: 如何找到保测度的映射，使得传输代价最小，

$$\text{(MP)}\qquad \min_{T*\mu=\nu}\mathscr{C}(T) \tag{B.3}$$

这种映射被称为是最优传输映射 (Optimal Mass Transportation Map)。最优传输映射对应的传输代价被称为是概率测度之间的 Wasserstein 距离:

$$W_c(\mu,\nu):=\min_{T:X\to Y}\left\{\int_X c(x,T(x))d\mu(x)\mid T_*\mu=\nu\right\} \tag{B.4}$$

。

定律 B.3: Kantorovich **对偶问题** Kantorovich 证明了蒙日问题解的存在性唯一性，并且发明了线性规划 (Linear Programming)，为此于 1975 年获得了诺贝尔经济奖。由线性规划的对偶性，Kantorovich 给出了 Wasserstein 距离的对偶方法:

$$(DP)\quad W_c(\mu,\nu):=\max_{\varphi,\psi}\left\{\int_X\varphi d\mu+\int_Y\psi d\nu\mid\varphi(x)+\psi(y)\leqslant c(x,y)\right\} \tag{B.5}$$

等价的，我们将 ψ 换成 φ 的 c-变换，$\varphi^c(y):=\inf_x\{c(x,y)-\varphi(x)\}$，那么 Wasserstein 距离为:

$$(DP)\quad W_c(\mu,\nu):=\max_{\varphi}\left\{\int_X\varphi d\mu+\int_Y\varphi^c d\nu\right\} \tag{B.6}$$

这里 φ 被称为是 Kantorovich 势能。

定律 B.4：WGAN **模型** 在 WGAN[57] 中，判别器计算测度间的 Wasserstein 距离就是利用上式：这里距离函数为 $c(x,y)=|x-y|$，可以证明如果 Kantorovich 势能为 1 -Lipsitz，那么 $\varphi^c=-\varphi$。这里 Kantorovich 势能由一个深度神经网络 ξ 来计算，记为 $\varphi\xi$。Wasserstein 距离为

$$W_c\left(\mu_\theta,\nu\right)=\max_{\xi}\left\{\int_X\varphi_\xi d\mu_\theta-\int_Y\varphi_\xi d\nu\right\} \tag{B.7}$$

，生成器极小化 Wasserstein 距离，

$$\min_{\theta}W_c\left(\mu_\theta,\nu\right) \tag{B.8}$$

。所以整个 WGAN 进行极小-极大优化:

$$\min_{\theta}\max_{\xi}\left\{\int_{\mathcal{Z}}\varphi_\xi\circ g_\theta(z)d\zeta(z)-\int_Y\varphi_\xi d\nu\right\} \tag{B.9}$$

。生成器极大化，判别器极小化，各自由一个深度网络交替完成。在优化过程中，解码映射 g_θ 和 Kantorovich 势能函数 $\varphi\xi$ 彼此独立。

定律 B.5：Brenier **方法** Brenier 理论 [] 表明，如果距离函数为 $c(x,y)=1/2|x-y|^2$，那么存在凸函数

$$u:X\to\mathbb{R} \tag{B.10}$$

，被称为是 Brenier 势能，最优传输映射由 Brenier 势能的梯度映射给出，

$$T(x)=\nabla u(x) \tag{B.11}$$

。由保测度条件 $T_*\mu=\nu$，Brenier 势能函数满足所谓的蒙日-安培方程:

$$\det\left(\frac{\partial^2u}{\partial x_i\partial x_j}\right)=\frac{\mu(x)}{\nu\circ\nabla u(x)} \tag{B.12}$$

。关键在于，Brenier 势能和 Kantorovich 势能满足简单的关系:

$$u(x)=\frac{1}{2}|x|^2-\varphi(x) \tag{B.13}$$

判别器计算 Kantorovich 势能，生成器计算 Brenier 势能。在实际优化中，判别器优化后，生成器可以直接推导出来，不必再经过优化过程。

证明 可参考顾险峰与雷娜合著的《最优传输理论与计算》。

参考文献

[1] BITOUK D, KUMAR N, DHILLON S, et al. Face swapping[J/OL]. ACM Transactions on Graphics, 2008: 1–8. http://dx.doi.org/10.1145/1360612.1360638.

[2] CHAN C, GINOSAR S, ZHOU T, et al. Everybody dance now[J/OL]. ITNOW, 2007: 34–34. http://dx.doi.org/10.1093/combul/bwl126.

[3] KORSHUNOVA I, SHI W, DAMBRE J, et al. Fast face-swap using convolutional neural networks[C/OL]//2017 IEEE International Conference on Computer Vision (ICCV). 2017. http://dx.doi.org/10.1109/iccv.2017.397.

[4] WU W, ZHANG Y, LI C, et al. Reenactgan: Learning to reenact faces via boundary transfer [M/OL]. 2018: 622–638. http://dx.doi.org/10.1007/978-3-030-01246-5_37.

[5] LYU S. Deepfake detection: Current challenges and next steps[C/OL]//2020 IEEE International Conference on Multimedia amp; Expo Workshops (ICMEW). 2020. http://dx.doi.org/10.1109/icmew46912.2020.9105991.

[6] AFCHAR D, NOZICK V, YAMAGISHI J, et al. Mesonet: a compact facial video forgery detection network[C/OL]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018. http://dx.doi.org/10.1109/WIFS.2018.8630761. DOI: 10.1109/wifs.2018.8630761.

[7] DANG H, LIU F, STEHOUWER J, et al. On the detection of digital face manipulation[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. http://dx.doi.org/10.1109/cvpr42600.2020.00582.

[8] LI L, BAO J, ZHANG T, et al. Face x-ray for more general face forgery detection[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. http://dx.doi.org/10.1109/cvpr42600.2020.00505.

[9] NGUYEN H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[J]. Cornell University - arXiv,Cornell University - arXiv, 2018.

[10] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: Learning to detect manipulated facial images[C/OL]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019. http://dx.doi.org/10.1109/iccv.2019.00009.

[11] ZI B, CHANG M, CHEN J, et al. Wilddeepfake[C/OL]//Proceedings of the 28th ACM International Conference on Multimedia. 2020. http://dx.doi.org/10.1145/3394171.3413769.

[12] WANG C, DENG W. Representative forgery mining for fake face detection[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. http://dx.doi.org/10.1109/cvpr46437.2021.01468.

[13] GU Q, CHEN S, YAO T, et al. Exploiting fine-grained face forgery clues via progressive enhancement learning[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 735–743. http://dx.doi.org/10.1609/aaai.v36i1.19954.

[14] ZHOU P, HAN X, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017. http://dx.doi.org/10.1109/cvprw.2017.229.

[15] CHEN S, YAO T, CHEN Y, et al. Local relation learning for face forgery detection[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 1081–1088. http://dx.doi.org/10.1609/aaai.v35i2.16193.

[16] GU Z, CHEN Y, YAO T, et al. Delving into the local: Dynamic inconsistency learning for deepfake video detection[Z].

[17] ZHAO H, WEI T, ZHOU W, et al. Multi-attentional deepfake detection[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. http://dx.doi.org/10.1109/cvpr46437.2021.00222.

[18] LI J, XIE H, LI J, et al. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. http://dx.doi.org/10.1109/cvpr46437.2021.00639.

[19] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[M/OL]. 2020: 86–103. http://dx.doi.org/10.1007/978-3-030-58610-2_6.

[20] NADIMPALLI A, RATTANI A. On improving cross-dataset generalization of deepfake detectors*[Z]. 2022.

[21] CHAPELLE O, SCHLKOPF B, ZIEN A. Semi-supervised learning[M/OL]. 2006. http://dx.doi.org/10.7551/mitpress/9780262033589.001.0001.

[22] RUFF L, VANDERMEULEN R, GöRNITZ N, et al. Deep semi-supervised anomaly detection [A]. 2019.

[23] LI X, NI R, YANG P, et al. Artifacts-disentangled adversarial learning for deepfake detection* [Z].

[24] XIAO S, LAN G, YANG J, et al. Mcs-gan: A different understanding for generalization of deep forgery detection*[J/OL]. IEEE Transactions on Multimedia, 2023: 1–13. http://dx.doi.org/10.1109/tmm.2023.3279993.

[25] JU Y, HU S, JIA S, et al. Improving fairness in deepfake detection[Z]. 2023.

[26] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[Z].

[27] LI Y, YANG X, SUN P, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics [C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. http://dx.doi.org/10.1109/cvpr42600.2020.00327.

[28] KINGMA D P, WELLING M. Auto-encoding variational bayes[A]. 2022. arXiv: 1312.6114.

[29] CAO J, MA C, YAO T, et al. End-to-end reconstruction-classification learning for face forgery detection[Z].

[30] GAO G, HUANG H, FU C, et al. Information bottleneck disentanglement for identity swapping [C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. http://dx.doi.org/10.1109/cvpr46437.2021.00341.

[31] SCHMIDHUBER J. Learning factorial codes by predictability minimization[J/OL]. Neural Computation, 1992: 863–879. http://dx.doi.org/10.1162/neco.1992.4.6.863.

[32] HU J, WANG S, LI X. Improving the generalization ability of deepfake detection via disentangled representation learning*[C/OL]//2021 IEEE International Conference on Image Processing (ICIP). 2021. http://dx.doi.org/10.1109/icip42928.2021.9506730.

[33] BAU D, ZHU J, WULFF J, et al. Seeing what a GAN cannot generate[J/OL]. CoRR, 2019, abs/1910.11626. http://arxiv.org/abs/1910.11626.

[34] NAGARAJAN V, KOLTER J. Gradient descent gan optimization is locally stable[J]. Neural Information Processing Systems,Neural Information Processing Systems, 2017.

[35] KHAYATKHOEI M, SINGH M, ELGAMMAL A. Disconnected manifold learning for generative adversarial networks[J]. Neural Information Processing Systems,Neural Information Processing Systems, 2018.

[36] TONEVA M, SORDONI A, DES COMBES R T, et al. An empirical study of example forgetting during deep neural network learning[A]. 2019. arXiv: 1812.05159.

[37] HUANG L, YU W, MA W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[A]. 2023. arXiv: 2311.05232.

[38] KUMAR R, GOYAL A, COURVILLE A C, et al. Maximum entropy generators for energy-based models[J/OL]. CoRR, 2019, abs/1901.08508. http://arxiv.org/abs/1901.08508.

[39] AN D, GUO Y, LEI N, et al. Ae-ot: A new generative model based on extended semi-discrete optimal transport[J]. International Conference on Learning Representations,International Conference on Learning Representations, 2020.

[40] TENENBAUM J B, SILVA V D, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J/OL]. Science, 2000: 2319–2323. http://dx.doi.org/10.1126/science.290.5500.2319.

[41] LEI N, SU K, CUI L, et al. A geometric view of optimal transportation and generative model [J/OL]. CoRR, 2017, abs/1710.05488. http://arxiv.org/abs/1710.05488.

[42] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. International Conference on Learning Representations,International Conference on Learning Representations, 2016.

[43] GU X D. 深度学习和几何[EB/OL]. 2018. https://mp.weixin.qq.com/s?__biz=MzA3NTM4MzY1Mg==&mid=2650814604&idx=1&sn=2c8765becfe9df06e33ee6c4ed59792f&c

hksm=8485ce87b3f24791707d181b9aabda2f4a7b3a027f5912592aee329607f19549c4a8374ebdbd&token=52299169&lang=zh_CN#rd.

[44] HAJIJ M, ZAMZMI G, CAI X. Persistent homology and graphs representation learning[J/OL]. CoRR, 2021, abs/2102.12926. https://arxiv.org/abs/2102.12926.

[45] MONTESUMA E F, MBOULA F N, SOULOUMIAC A. Recent advances in optimal transport for machine learning[A]. 2023. arXiv: 2306.16156.

[46] IMFELD M, GRALDI J, GIORDANO M, et al. Transformer fusion with optimal transport[Z].

[47] AN D, GUO Y, ZHANG M, et al. Ae-ot-gan: Training gans from data specific latent distribution[M/OL]. 2020: 548–564. http://dx.doi.org/10.1007/978-3-030-58574-7_33.

[48] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C/OL]// GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

[49] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. http://dx.doi.org/10.1109/cvpr.2017.195.

[50] ZHOU H, LU C, YANG S, et al. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts[J]. Cornell University - arXiv,Cornell University - arXiv, 2021.

[51] KINGMA D, BA J. Adam: A method for stochastic optimization[A]. 2014.

[52] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[J/OL]. International Journal of Computer Vision, 2020: 336–359. http://dx.doi.org/10.1007/s11263-019-01228-7.

[53] SU J. Variational inference: A unified framework of generative models and some revelations [A]. 2018. arXiv: 1807.05936.

[54] TISHBY N, ZASLAVSKY N. Deep learning and the information bottleneck principle[J/OL]. CoRR, 2015, abs/1503.02406. http://arxiv.org/abs/1503.02406.

[55] ALEMI A A, FISCHER I, DILLON J V, et al. Deep variational information bottleneck[A]. 2019. arXiv: 1612.00410.

[56] LI K, MALIK J. Implicit maximum likelihood estimation[J/OL]. CoRR, 2018, abs/1809.09087. http://arxiv.org/abs/1809.09087.

[57] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein gan[A]. 2017. arXiv: 1701.07875.

致谢

打字到这里，我应该述说个人过去四年来的艰难历程，以供后人参考。

作为一名南昌大学软件工程专业的本科学生，虽然本科期间所学专业是软件工程，对深度学习有浓厚的兴趣和热情。在我的专业学习过程中，虽然在学校学习的大多数课程只是草草带过，但我通过课外学习让我对深度学习的原理和应用有了初步的了解和实践。我在第一学年就意识到要学好人工智能理论，一开始就应该重视基础数学，特别是高等数学、线性代数、离散数学与概率统计等与人工智能发展息息相关的课程，因此我在学校中与数学相关的课程取得了不错的成绩，特别是高等数学（二）、线性代数、离散数学的课程分数达到了 90+ 的好成绩。

从第二学年开始，计算机四大课的出现除了数据结构之外我才发现个人对计算机软件工程上的各种工程细节是不大擅长甚至是不感兴趣的，于是个人决定不求各大科目的完课程度，在一个教育资源平平的城市中决定要像后来清华顾险峰教授对莘莘学子那番言语“不要迷信盲从，不要过于追逐潮流，要独立思考，目光长远，追求深刻，特别是学习艰深的前沿科学，求真求美，培养出难以被取代的核心能力”一样，不求对目前热门的计算机令人摸不着头脑的工程问题进行死磕，只求学习真正有长远学科价值且思想底蕴深刻的人工智能理论。

一开始我的出发点就是学习 coursera 上面久负盛名的吴恩达深度学习课程，当然通过我个人的努力也成功拿到了 coursera 发放的证书；同样地，因为当时还想通过考取企业认证的人工智能工程师来多一份就业机会，于是考取了华为初级人工智能工程师认证。但是这些依旧解不开我的困惑，特别是当我通过调用 tensorflow、华为封装的 DNN 模型等现成框架时：为什么是这样做?

这个时候我的问题分为了两个方面：一个是人工智能与人类社会关系中各种范畴的关系；一个是人工智能的传统理科背景。两个问题中，我果断向哲学研究进发，首先接触到的是各种精神分析理论、现象学流派，我在其中也很幸运地受到了来自同济大学和复旦大学哲学系的帮助，了解到前沿的人工智能理论也受到哲学不小的影响，比如现在新近的具身性智能就是来自于法国哲学家梅洛-庞蒂的具身现象学，人工智能的流形分布假设也是来自一

些早期精神分析师的系统研究，甚至在加州大学伯克利分校的研究德国哲学家海德格尔的著名学者德雷福斯就较早一步推动具身智能的观点，令我在未来的学术研究中受益匪浅。

当然对人工智能的传统理科背景我也没有放松。在我的学术背景方面，我参与了两个与深度学习相关的研究项目。深度学习的快速发展和广泛使用很大程度上得益于一系列简单好用且强大的编程框架，例如 Pytorch 和 Tensorflow 等等，但大多数从业者只是这些框架的“调包侠”，对于这些框架内部的细节实现却了解甚少。出于希望从事深度学习底层框架的开发工作，个人在线上参与 CMU 10-414/714 课程，我实现了一个基于现代深度学习框架 pytorch 模仿的 needle 框架，通过该项目学习了对神经网络的后向传播机制、pytorch 在 ndarray 上的设计哲学、以及一定的基于 cuda 的并行式编程等技术栈。第二个项目就是这项基于图像语义解耦技术的人脸图像检测，我负责设计和实现了一个的人脸二分类模型，并受到顾险峰教授的最优传输理论启发用于对真样本和伪造样本的聚类，目前正在做实验验证其是否提高生成样本的质量和多样性。

在四年前下定决心申请中科院之前，我也是万般犹豫，因为我意识到那时自身英语水平的不足，所以为了克服自身的短板，我花了四年时间将四六级和专四专八单词全部记忆来加深对英语的理解和运用，也在四六级笔试分别取得了 562 分和 560 分的成绩。即便我自身在运用英语上仍存在较大差距，但是我相信一定可以迎头赶上，打开进入国际学术交流的大门。

顺便一提，在这大学本科期间我已经接触过很多同学，不论社科文方面的哲学，还是就是在贵校学习的研究生同辈，发现他们选择专业并非出自天然的热爱，而是为当时的世俗价值观所裹挟，真正的目的是专业后面的利益和名誉，因而学术层面无法前行太远。尤其伴随着周围人的不理解，乃至家庭层面的误解，我经常在孤身一人思考：如果一个资质平庸的年轻人，仅凭一腔热血而投身其中，是否会落得一事无成，抱憾终生？急于发表文章而不下苦功加强学术修养的浮躁盛况，难道是所有人将来的一面镜子？论文在当今变成批量流水线式的生产，人类依旧无法理解其运作机理，也无从判断其优劣，甚至已经出现“取消人类思想的介入”让“机器自主发现规律”的论调，是深度学习乃至整个人工智能的未来？我作为一个在传统学术价值观念下深受其影响的人，在这些问题中感到不自量力，可能就算攻读研究生毕业后的就业目标也只是成为一名普通的深度学习的研究员或工程师，为人工智

能的发展和应用尽些微薄之力。

在这里感谢学校、特别是导师和辅导员原谅我过去的一些鲁莽之举，过去的误解随着时间冰释花绽，离开成为了最后的序曲。

感谢过去在线上线下帮助过我的各大高校好友，是他们让我知道社会百态，他们背后的幸苦付出令我无语言表。

“浮云一别后，流水十年间。”

蒋涛

2024 年 3 月 28 日