1 试验结果

在本节中，我们将报告我们的实验结果

1.1 训练过程

AE-OT 模型的训练主要包括两个步骤：训练 AE 和寻找 OT 图。OT 步骤是使用该算法的 GPU 实现来完成的，如第 4 节所述。在 AE 步骤中，在训练过程中，我们采用 Adam 算法 [49] 对中性网络的参数进行优化，学习速率为 0.003 $\beta_1 = 0.5$ $\beta_2 = 0.999$。当 L^2 损失停止下降时，这意味着网络已经找到了一个良好的编码图，我们冻结编码器部分，并继续训练网络进行解码图。编码器冻结前后的训练损失见表 (1)。接下来，为了从给定的分布（这里使用均匀分布）到潜在特征的分布中找到 OT 映射，我们从均匀分布中随机抽取 $100N$ 个随机点来计算能量的梯度。这里，N 是数据集的潜在特征的数量。此外，在实验中，θ_{ij} 被设置为对不同的数据集的不同。具体来说，对于 MNIST 和 Fashion-MNIST 数据集，θ_{ij} 被设置为 0.75，而对于 CIFAR-10 和 CelebA 数据集，它分别被设置为 0.68 和 0.75。

表 1: 编码器冻结前后 AEs 的 L^2 损失

Situation	Dataset			
	MNIST	Fashion-MNIST	CIFAR-10	CelebA
Before	0.0013	0.0026	0.0023	0.0077
After	0.0005	0.0011	0.0018	0.0074

我们的 AE-OT 模型是在一个 Linux 平台上使用 PyTorch 实现的。所有实验均在 GTX1080Ti 上进行。

1.2 传输映射的不连续性试验

在这个实验中，我们想要检验我们的假设：在大多数实际应用中，目标测度的支持是非凸的，奇异集是非空的，概率分布映射沿奇异集是不连续的。

如图 12 所示，我们使用 AE 计算从 CelebA 数据集 $(\sum, \nu)$ 到潜在空间 Z 的编码/解码映射；编码映射 $f_\theta : \sum \to Z$ 在潜在空间上将 ν 向前推到 $(f_\theta)_\# \mu$。在潜在空间中，我们基于第 4 节，$T : Z \to Z$ 中描述的算法计算 OT 映射，其中 T 将单位立方体 ζ 中的均匀分布映射到 $(f_\theta)_\# \nu$。然后，我们从分布 ζ 中随机抽取一个样本 z，并使用解码映射 $g_\xi : Z \to \sum$ 将 $T(z)$ 映射到生成的人类面部图像 $g_\xi \circ T(z)$。图 12(a) 展示了由该 AE-OT 框架生成的真实面部图像。

如果潜在空间中的推进测度 $(f_\theta)_\# \nu$ 的支持是非凸的，则会有一个奇异集 $\sum_k$，其中 $k > 0$。我们想检测出 $\sum_k$ 的存在。我们在潜在空间中随机抽取单位立方体中的线段，然后沿着这条线段进行密集插值，生成面部图像。如图 12(b) 所示，我们找到了一个线段 γ，并在一个有一双棕色眼睛的男孩和一个有一对蓝眼睛的女孩之间生成了一个变形序列。在中间，我们生成了一张有一只蓝眼睛和一只棕色眼睛的脸，这绝对是不现实的，而且在 $\sum$ 之外。这个结果意味着线段 γ 经过一个奇异集 $\sum_k$，其中运输图 T 是不连续的。这也表明了我们的假设是正确的：编码的人脸图像测量在潜在空间上的支持是非凸的。

作为一个副产品，我们发现这个 AE-OT 框架提高了训练速度的 5 倍，并提高了收敛稳定性，因为 OT 步骤是一个凸优化。因此，它为改进现有的 GANs 提供了一种很有前途的方法。

1.3 模式崩溃比较

由于合成数据集由显式分布和已知模态组成，因此可以准确地测量模态坍塌。我们选择了两个在之前的工作中研究或提出的合成数据集：一个二维网格数据集。

为了选择模态崩溃的测量度量，我们采用了三个以前使用的度量 [50,51]。模式的数量计算由生成模型产生的样本所捕获的模式的数量。在这个度量中，如果在该模式的三个标准差内没有产生样本，则被认为是丢失的。高质量样本的百分比衡量在最近模式的三个标准差内产生的样本的比例。第三个度量标准，在参考文献中使用。[51]，是反向回-莱布勒 (KL) 散度。在这个度量中，每个生成的样本被分配到其最近的模式，并且我们计算分配在每个模式上的样本的直方图。然后，该直方图形成一个离散分布，然后计算其与真实数据形成的直方图的 KL 散度。直观地说，这衡量了生成的样本在真实分布的所有模式之间的平衡程度

在参考文献中。[51]，作者用上述三个指标评估了 GAN[26]、ALI[52]、MD[30] 和 PacGAN[51]。每个实验都在相同的生成器架构下进行训练，总共有大约 400k 的训练参数。这些网络在 100k 个样本上被训练了 400 个时代。对于 AE-ot 实验，由于源空间和目标空间都是二维的，因此不需要训练 AE。我们直接计算了一个半离散的 OT，它映射在单位平方上的均匀分布和经验的真实数据分布之间。理论上，OT 恢复所有模式所需的最小真实样本量是每个模式一个样本。然而，这可能会导致在插值过程中产生低质量的样品。因此，对于 OT 计算，我们取 512 个真实样本，并基于此映射生成新的样本。我们注意到，在这种情况下，在 OT 计算中只有 512 个参数需要优化，并且由于凸正定黑森的存在，优化过程是稳定的。我们的结果在表 (2) 中所示，以前的方法的基准测试是从参考文献中复制出来的。[51]. 为了说明这一点，我们在图 13 中绘制了与 GAN 和 PAcGAN 一起绘制的合成数据集上的结果。

表 2: 二维网格数据集的模式折叠比较

	Modes	Samples	Reverse KL
GAN	17.3±0.8	94.8 ±0.7%	0.70 ±0.07
ALI	24.1 ±0.4	95.7 ±0.6%	0.14 ±0.03
MD	23.8 ±0.5	79.9 ±3.2%	0.17 ±0.003
PacGAN2	23.8 ±0.7	91.3 ±0.8%	0.13 ±0.04
PacGAN3	24.6 ±0.4	94.2 ±0.4%	0.06 ±0.02
PacGAN4	24.8 ±0.2	93.6 ±0.6%	0.04 ±0.01
AE-OT	25.0 ±0.0	99.8 ±0.2%	0.007 ±0.002

1.4 与现有技术水平的比较

我们设计了实验来比较我们提出的 AE-OT 模型与最先进的生成模型，包括由 Lucic 等人评估的对抗性模型。[33]，以及 Hoshen 和 Malik 研究的非对抗性模型。[36].

为了进行公平的比较，我们使用了相同的测试数据集和网络体系结构。这些数据集包括 MNIST[53]、MNIST-Dashion[54]、CIFAR-10[55] 和 CelebA[56]，与在 Refs 中测试的数据类似。[31,36]. 该网络架构与 Lucic 等人在参考文献中使用的类似。[33]. 特别是，在我们的 AE-OT 模型中，解码器的网络结构与 Ref 中 GANs 的生成器相同。[33]，编码器与解码器对称。

我们使用 FID 评分 [31] 和 PRD 曲线作为评价标准，将我们的模型与最先进的生成模型进行了比较。FID 评分衡量生成结果的视觉保真度，并对图像损坏具有鲁棒性。然而，FID 评分对模式的添加和删除 [33] 很敏感。因此，我们也使用了 PRD 曲线，它可以量化真实数据集 [32] 上的模式下降程度。

1.4.1 与 FID 评分的比较

FID 得分的计算方法如下：① 通过运行初始网络 [30] 提取生成的图像和真实图像在视觉上有意义的特征；② 用高斯分布拟合真实和生成的特征分布；③ 使用以下公式计算两个高斯分布之间的距离：

$$FID = \|u_r - u_g\|_2^2 + T_r \left[\sum_r + \sum_g -2 \left(\sum_r \sum_g \right)^{\frac{1}{2}} \right] \tag{1}$$

其中，μ_r 和 μ_g 分别表示真实分布和生成分布的平均值；$\sum_r$ 和 $\sum_g$ 表示这些分布的方差

比较结果汇总见表 (3) 和表 (4)。各种 GANs 的统计数据来自 Lucic 等人的 [33]，而非对抗性生成模型的统计数据来自 Hoshen 和 Malik[36]。总的来说，我们提出的模型比其他最先进的生成模型获得了更好的 FID 分数。

表 3: 与 FID-I 进行的定量比较

Dataset	Adversarial				
	MM GAN	NS GAN	LS GAN	WGAN	BEGAN
MNIST	9.8	6.8	7.8	6.7	13.1
Fashion-MNIST	29.6	26.5	30.7	21.5	22.9
CIF AR-10	72.7	58.5	87.1	55.2	71.4
CelebA	65.6	55.0	53.8	41.3	**38.9**

最佳结果以粗体显示。

表 4: 与 FID-II 之间的定量比较

Dataset	Non-adversarial			Reference	
	VAE	GLO	CLANN	AE	AE-OT
MNIST	23.8	49.6	8.6	5.5	**6.4**
Fashion-MNIST	58.7	57.7	13.0	4.7	**10.2**
CIF AR-10	155.7	65.4	46.5	28.2	**38.1**
CelebA	85.7	52.4	46.3	67.5	68.4(**28.6**)

最佳结果以粗体显示。

理论上，我们的 AE-OT 模型的 FID 分数应该与预先训练过的 AEs 模型的分数很接近；我们的实验也验证了这一点。

我们的 AE 的固定网络架构采用了 Lucic 等人的 [33]；它的容量不足以编码 CIFAR-10 或 CelebaA，所以我们不得不对这些数据集进行降采样。我们从 CIFAR-10 中随机选择 25k 图像，从 CelebA 中选择 10k 图像来训练我们的模型。即便如此，我们的模型在 CIFAR-10 中获得了最好的 FID 分数。由于 InfoGAN 模型的有限容量，CelebA 的 AE 性能的 FID 为 67.5 并不理想，进一步导致生成的数据集的 FID 为 68.4。通过在 AE 体系结构中增加两个卷积层，CelebA 的 L^2 损失小于 0.03，FID 得分优于其他所有的模型（28.6，如表 (4) 的括号所示）。

1.4.2 与 PRD 曲线的比较

FID 评分是衡量生成分布与真实数据分布差异的有效方法，但它主要关注精度，不能准确地捕获生成模型可以覆盖的真实数据的哪个部分。参考文献中提出的方法。[32] 将分布之间的差异分解为两个组成部分：精度和召回率。

给定一个参考分布 P 和一个学习到的分布 Q，精度直观地衡量从 Q 中获得的样本的质量，而召回率则衡量 Q 所覆盖的 P 的比例。

我们使用了 Sajjadi 等人在参考文献中提出的 (F_8，$F_{1\,8}$) 的概念。[32] 来量化精度和查全率的相对重要性。图 14 总结了比较结果。每个点代表一个具有一组超参数的特定模型。一个点越靠近右上角，模型的性能就越好。蓝色和绿色的点表示参考文献中评估的 GANs 和 VAEs。[32]，黄点代表 Ref 中的 GLANN 模型。[36]，红点是我们的 AE-OT 模型.

很明显，我们提出的模型在 MNIST 和 Dashion-MNIST 上优于其他模型。对于 CIFAR-10 数据集，我们的模型的精度略低于 GANs 和 GLANN，但查全率最高。对于 CelebA 数据集，由于 AE 的容量有限，我们的模型的性能并不令人印象深刻。然而，在 AE 中增加了两个卷积层后，我们的模型获得了最好的分数。

1.4.3 视觉比较

图 15 显示了我们提出的方法生成的图像与 Lucic 等人研究的 GANs 生成的图像之间的视觉比较。[33] 和 Hoshen 和 Malik 研究的非对抗性模型。[36]. 第一列显示了原始图像，第二列显示了 AE 生成的结果，第三列显示最好的生成结果的差距在 Lucicetal.[33]，第四列显示所生成的结果霍申和马利克 [36] 的模型，和第五列显示我们的方法的结果。很明显，我们的方法可以生成高质量的图像，并覆盖了所有的模式。