arXiv:2301.04791v2 [stat.ML] 8 May 2023

面向切片 Wasserstein 点云重建的自注意力摊销分布投影优化

Khai Nguyen *1 **Dang Nguyen** *2 **Nhat Ho** [1]

摘要

最大切片瓦瑟斯坦距离（Max-SW）一直被视为解决切片 Wasserstein 距离（SW）投影判别能力不足问题的方法。在涉及多个相互独立的概率测度对的应用中，摊销投影优化用于根据两个输入测度预测“最大”投影方向，而无需多次使用投影梯度上升。尽管效率较高，Max-SW 及其摊销版本仍无法保证度量性，这是因为投影梯度上升存在次优性，且存在摊销差距。因此，我们提出以采用冯·米塞斯–费舍尔（vMF）投影分布的分布式切片 Wasserstein 距离（v-DSW）取代 Max-SW。由于采用任意非退化 vMF 分布的 v-DSW 都是度量，其摊销版本在进行摊销时也能保证度量性。此外，现有摊销模型不具备置换不变性和对称性。为解决这一问题，我们设计了基于自注意力架构的摊销模型。具体而言，我们采用高效自注意力架构，使计算量随支撑点数量线性增长。通过这两项改进，我们提出了自注意力摊销分布投影优化，并展示了其在点云重建及其下游应用中的出色表现。

1. 引言

Wasserstein 距离（Villani，2008年；Peyr´e & Cuturi，2019）已被机器学习领域广泛认可为一种有效工具。例如，Wasserstein 距离被用于探索数据中的簇（Ho 等人，2017）、在不同领域之间迁移知识（Courty 等人，2016；Damodaran 等人，2018）、学习生成模型（Arjovsky 等人，2017；Tolstikhin 等人，2018）、从图中提取特征（Vincent-Cuaz 等人，2022）、比较数据集（Alvarez-Melis & Fusi，2020），以及许多其他应用。尽管 Wasserstein 距离十分有效，但其计算成本极高。具体而言，离散情形下 Wasserstein 距离的计算复杂度和空间复杂度分别为 $\mathcal{O}(m^3 \log m)$ 和 $\mathcal{O}(m^2)$，其中 m 是支撑点的数量。对于需要在不同测度对之间多次 计算 Wasserstein 距离的应用，计算问题会更加严重。例如：深度生成建模（Genevay 等人，2018）、深度领域自适应（Bhushan Damodaran 等人，2018）、数据集比较（Alvarez-Melis & Fusi，2020）、主题建模（Huynh 等人，2020）、点云重建（Achlioptas 等人，2018）等。

通过添加熵正则化（Cuturi，2013），可以在$\mathcal{O}(m^2/\varepsilon^2)$时间内获得 Wasserstein 距离的 ε近似。然而，由于需要存储代价矩阵，这种方法无法降低 $\mathcal{O}(m^2)$的内存复杂度。一种更高效的方法基于一维 Wasserstein 距离的闭式解，这种距离称为切片瓦瑟斯坦距离（Bonneel 等人，2015）。切片瓦瑟斯坦（SW）距离定义为：两个原始测度经随机一维推前后所得测度之间的 Wasserstein 距离的期望。得益于闭式解，SW 可在 $\mathcal{O}(m\log_2 m)$时间内求解，同时具有线性内存复杂度 $\mathcal{O}(m)$。此外，在高维统计推断中，SW 也优于 Wasserstein 距离。具体而言，SW 的样本复杂度（统计估计率）为 $\mathcal{O}(n^{-1/2})$，而 Wasserstein 距离为 $\mathcal{O}(n^{-1/d})$，其中 d 为维度数，n 为数据样本数。凭借这些优良性质，SW 已成功应用于多种领域，例如生成建模（Deshpande 等人，2018年；Nguyen & Ho，2022b；Nguyen 等人，2023）、领域自适应（Lee 等人，2019a）、贝叶斯推断（Nadjahi 等人，2020年；Yi & Liu，2021年）、点云表示学习（Nguyen 等，2021c；Naderializadeh 等人，2021年）等。

*贡献均等 [1]美国德克萨斯大学奥斯汀分校统计与数据科学系 [2] VinAI Research。通信作者：Khai Nguyen <khainb@utexas.edu>。

论文集： 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023。版权所有 2023，归作者所有。

SW 的缺点在于，由于对投影方向采用均匀分布，它对所有投影一视同仁。实际应用中，这种选择并不合适，因为有些投影方向无法区分所关注的两个测度（Kolouri 等人，2018年）。为此，Deshpande 等人（2019年）提出最大切片瓦瑟斯坦距离（Max-SW），通过搜索能够使投影瓦瑟斯坦距离达到最大的最佳投影方向来定义该距离。Max-SW 需要使用投影次梯度上升算法来寻找“最大”切片。因此，在需要多次针对不同测度对计算 Max-SW 的应用中，反复执行优化过程代价高昂。例如，本文关注点云重建应用，在这类应用中，需要计算点云及其重建版本上多个经验测度对之间的 Max-SW。

为了解决这一问题，Nguyen 与 Ho（2022a）提出了摊销投影优化。与其他摊销优化（Shu，2017；Amos，2022）（即学习如何学习）一样，该方法会估计一个摊销模型，根据两个输入经验测度预测最佳投影方向。Nguyen 与 Ho（2022a）提出了三种摊销模型：线性模型、广义线性模型和非线性模型。线性模型假设“最大”投影方向是两个测度的支撑点的线性组合。广义线性模型通过对两个测度的支撑点应用连接函数来引入线性，而非线性模型则使用多层感知器，以获得更强的表达能力。

尽管先前的工作在实践中表现良好，但尚未充分探索摊销优化在切片瓦瑟斯坦设置中的潜力。当前的摊销优化框架存在两个问题。首先，摊销优化的次优性会导致由预测投影方向得到的投影距离失去度量性。具体而言，Max-SW 只有在全局最优解处才具有度量性。因此，使用具有次优解的摊销模型，无法确保所有测度对都满足度量性。失去度量性可能会损害下游应用的性能。其次，当前的摊销模型对两个输入测度的支撑点不具备置换不变性，而且也不对称。置换不变性和对称性至关重要，因为对两个输入经验测度的支撑点进行置换或交换这两个输入经验测度，都不会改变“最大”投影方向。通过在摊销模型中引入置换不变性和对称性，有助于学习更好的摊销模型并缩小摊销差距

本文着重解决当前摊销投影优化框架面临的两个问题。为保持度量性，我们提出摊销分布式投影优化框架，用于预测最佳投影方向分布。具体而言，我们针对分布式切片瓦瑟斯坦距离（DSW）（Nguyen 等人，2021a），采用冯·米塞斯-费舍尔分布（vMF）作为切片分布（Jupp 和 Mardia，1979）进行摊销优化，而非采用 Max-SW。得益于 vMF 的平滑性，即使存在非零摊销差距，也能保持度量性。为满足置换不变性和对称性，我们提出采用自注意力机制（Vaswani 等人，2017）来设计摊销模型。此外，我们采用计算复杂度随支撑点数量线性增长的高效自注意力方法，包括高效注意力（Shen 等人，2021）和线性注意力（Wang 等人，2020a）。

Co贡献。 总而言之，我们的贡献有两方面 ：

1. 首先，我们提出了摊销分布投影摊销框架，用于预测冯·米塞斯–费舍尔分布（vMF）在分布切片 Wasserstein 距离（DSW）中的最佳位置参数。由于 vMF 的平滑性，所有测度对之间的度量性均得到保证。此外，我们通过引入排列不变性和对称性这两种归纳偏置，增强了摊销模型。为提升效率，我们采用两种具有线性复杂度的注意力机制来参数化摊销模型，分别是高效注意力（Shen 等人，2021年）和线性注意力（Wang 等人，2020年a）。结合上述两项改进，我们得到自注意力摊销分布投影摊销框架

2. 其次，我们将这一新框架应用于点云重建问题。具体而言，我们希望训练一个自编码器，使其能够通过潜在表示重建（编码和解码）所有点云。核心思路是将点云视为经验测度，并将切片 Wasserstein 距离用作重建损失。在此，摊销优化提供了一种快速生成切片 Wasserstein 距离投影方向的方法，以区分原始点云与重建点云之间的所有配对。在实证方面，我们表明，与摊销投影优化框架和广泛使用的距离相比，自注意力摊销分布投影摊销能够在 ModelNet40 数据集（Wu 等人，2015年）上得到更优的重建点云。此外，在下游任务中，该新框架还能提高 ModelNet40 上的分类准确率，并生成质量更高的 ShapeNet 椅子。

组织结构。 本文其余部分安排如下：第2节介绍点云重建和常用距离的背景知识。第3节针对点云重建问题，定义了新的摊销分布投影优化框架。第4节通过大量实验，在点云重建、迁移学习、

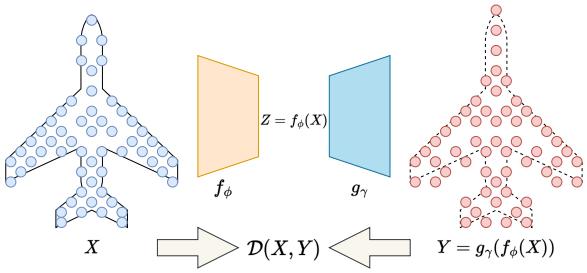

图 *1.* 点云 X （一个平面）的重建。

和点云生成任务上对所提出的方法进行了基准测试。最后，关键结果的证明和补充材料见补充材料。

符号说明。 对于任意 $d \geq 2$，我们记 $\mathcal{U}(\mathbb{S}^{d-1})$ 为单位超球面 $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d \ ||\theta||_2^2 = 1\}$ 上的均匀测度。对于 $p \geq 1$， $\mathcal{P}_p(\mathbb{R}^d)$ 是 $\mathbb{R}^d$ 上所有具有有限 p 阶矩的概率测度的集合。对于任意两个序列 a_n 和 b_n，记号 $a_n = \mathcal{O}(b_n)$ 表示，对于所有 $n \geq 1$，都有 $a_n \leq Cb_n$，其中 C 是某个通用常数。我们记 $\theta\sharp\mu$ 为 μ 通过函数 $f\colon \mathbb{R}^d \to \mathbb{R}$ 推前得到的测度，即 $f(x) = \theta^\top x$。

2. 预备知识

我们首先在第 2.1 节回顾点云重建框架。随后，在第 2.2 节讨论两组点云之间常用的度量。最后，在第 2.3 节介绍适用于点云重建场景的摊销投影优化框架的一个改进定义。

2.1. 点云重建

我们将由 m 个点 $x_1, \ldots, x_m \in \mathbb{R}^d (d \geq 1)$ 组成的点云记为 $X = (x_1, \ldots, x_m) \in \mathbb{R}^{dm}$，它是由点云中的所有点拼接而成的向量。我们将所有可能的点云集合记为 $\mathcal{X} \subset \mathbb{R}^{dm}$。

置换不变度量空间。 给定一个置换一一映射函数 $\sigma\colon [m] \to [m]$，对于所有 $X \in \mathcal{X}$，有 $\sigma(X) \in \mathcal{X}$。此外，我们需要一个度量 $\mathcal{D}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$，使得对于所有 $X \in \mathcal{X}$， $\mathcal{D}(X, \sigma(X)) = 0$，其中 $\sigma(X) = (x_{\sigma(1)}, \ldots, x_{\sigma(m)})$。这里， $\mathcal{D}$ 是一个度量，也就是说，它需要满足非负性、对称性、三角不等式和同一性。二元组 $(\mathcal{X}, \mathcal{D})$ 构成一个点云度量空间。

通过重建学习表示。 由于度量空间复杂，点云的原始表示在实际应用中难以处理。因此，一种著名的方法是将点云映射到另一个空间中的点，例如欧几里得空间，这样更便于应用机器学习算法。具体而言，我们希望估计一个函数 $f_\phi\colon \mathcal{X} \to \mathcal{Z} (\phi \in \Phi)$，其中 $\mathcal{Z}$ 是属于另一个度量空间的集合。然后，我们可以应用机器学习算法处理 $\mathcal{Z}$，而不是 $\mathcal{X}$。估计函数 f_ϕ 最广为人知且有效的方法是使用重建损失。也就是说，给定点云数据集 $p(X)$（ $\mathcal{X}$ 上的分布），我们通过最小化以下目标函数，同时估计 f_ϕ 和函数 $g_\gamma\colon \mathcal{Z} \to \mathcal{X} (\gamma \in \Gamma)$:

$$\min_{\phi \in \Phi, \gamma \in \Gamma} \mathbb{E}_{X \sim p(X)} \mathcal{D}(X, g_\gamma(f_\phi(X))). \tag{1}$$

损失 $\mathbb{E}_{X \sim p(X)} \mathcal{D}(X, g_\gamma(f_\phi(X)))$ 称为重建损失。如果重建损失为 0，则我们几乎必然有 $g_\gamma = f_\phi^{-1}$ p。因此，我们可以通过函数 f_ϕ （称为编码器）和 g_γ （称为解码器），从 $\mathcal{X}$ 转到 $\mathcal{Z}$，再从 $\mathcal{Z}$ 返回 $\mathcal{X}$，而不会丢失信息。图 1 展示了该框架的示意图（Achlioptas 等人，2018年）。学会如何良好地进行重建后，还可以使用自编码器（这一对 (f_ϕ, g_γ)）执行其他点云任务，例如形状插值、形状编辑、形状类比、形状补全、点云分类和点云生成（Achlioptas 等人，2018年）。

2.2. 点云度量空间

现在我们回顾几种著名的度量 $\mathcal{D}$，包括倒角距离（Barrow 等人，1977年）、瓦瑟斯坦距离（Villani，2008年）、切片瓦瑟斯坦（SW）距离（Bonneel 等人，2015年）以及最大切片瓦瑟斯坦（Max-SW）距离（Deshpande 等人，2019年）。

Chamfer 距离。对于任意两个点云 X 和 Y，Chamfer 距离定义如下：$\mathrm{CD}(X, Y) =$

$$\frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} \|x - y\|_2^2 + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} \|x - y\|_2^2, \tag{2}$$

其中， $|X|$ 表示 X中的点数。

Wasserstein 距离。给定两个概率测度 $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ 和 $\nu \in \mathcal{P}_p(\mathbb{R}^d)$，$\mu$与 ν之间的 Wasserstein 距离定义如下:

$$\mathrm{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_p^p d\pi(x, y) \right)^{\frac{1}{p}} \tag{3}$$

其中， $\Pi(\mu, \nu)$ 是所有边缘分布分别为 μ 和 ν 的耦合集合。由于 Wasserstein 距离最初定义在概率测度空间上，我们需要将点云 $X = (x_1, \ldots, x_m) \in \mathcal{X}$转换为相应的经验概率测度 $P_X = \frac{1}{m} \sum_{i=1}^m \delta_{x_i} \in \mathcal{P}(\mathbb{R}^d)$。因此，对于 $X, Y \in \mathcal{X}$，我们可以使用 $\mathcal{D}(X, Y) = \mathrm{W}_p(P_X, P_Y)$。

切片 Wasserstein 距离。如前所述，Wasserstein 距离的计算开销很大，其时间复杂度为 $\mathcal{O}(m^3 \log m)$，空间复杂度为 $\mathcal{O}(m^2)$。因此，另一种选择是切片瓦瑟斯坦距离（SW）

两个概率测度 $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ 和 $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ 之间的距离为:

$$\mathrm{SW}_p(\mu,\nu) = \left(\mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right)^{\frac{1}{p}}, \quad (4)$$

切片瓦瑟斯坦距离的优点在于，$\mathrm{W}_p(\theta\sharp\mu,\ \theta\sharp\nu)$ 存在闭式解，且

$$\mathrm{W}_p(\theta\sharp\mu,\theta\sharp\nu) = \left(\int_0^1 |F^{-1}_{\theta\sharp\mu}(z) - F^{-1}_{\theta\sharp\nu}(z)|^p dz\right)^{\frac{1}{p}},$$

其中 F^{-1} 表示逆累积分布函数。期望通常通过蒙特卡洛采样近似，即用从 $\mathcal{U}(\mathbb{S}^{d-1})$ 中独立同分布抽取的 θ_1、……、θ_L 的平均值代替。切片瓦瑟斯坦距离的计算复杂度和空间复杂度分别为 $\mathcal{O}(Lm\log_2 m)$ 和 $\mathcal{O}(Lm)$。

最大切片瓦瑟斯坦距离。众所周知，由于采用均匀采样，切片瓦瑟斯坦距离包含许多判别能力较弱的投影。因此，提出最大切片瓦瑟斯坦距离，以使用判别能力最强的投影方向。$\mu \in \mathcal{P}_p(\mathbb{R}^d)$ 与 $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ 之间的最大切片瓦瑟斯坦距离（Max-SW）（Deshpande 等人，2019年）定义如下:

$$\text{Max-SW}_p(\mu,\nu) = \max_{\theta\in\mathbb{S}^{d-1}} W_p(\theta\sharp\mu,\theta\sharp\nu), \quad (5)$$

Max-SW 通常通过投影次梯度上升算法进行计算。当投影次梯度上升算法迭代 $T \geq 1$ 次时，Max-SW 的计算量为 $\mathcal{O}(Tm\log_2 m)$，空间复杂度为 $\mathcal{O}(m)$。SW 和 Max-SW 均已成功应用于点云重建（Nguyen 等，2021c）。

2.3. 摊销投影优化

摊销优化。我们从摊销优化的定义开始。

定义 1。对于上下文空间 $\mathcal{X}$ 中的每个上下文变量 x，$\theta^\star(x)$ 是优化问题 $\theta^\star(x) = \arg\min_{\theta\in\Theta}\mathcal{L}(\theta,x)$ 的解，其中 Θ 是解空间。参数函数 $f_\psi\colon \mathcal{X}\to\Theta$，其中 $\psi\in\Psi$，称为摊销模型，如果

$$f_\psi(x) \approx \theta^\star(x), \quad \forall x\in\mathcal{X}. \quad (6)$$

摊销模型通过摊销优化目标进行训练，该目标定义如下:

$$\min_{\psi\in\Psi}\mathbb{E}_{x\sim p(x)}\mathcal{L}(f_\psi(x),x), \quad (7)$$

其中 $p(x)$ 是 $\mathcal{X}$ 上的概率测度，用于衡量优化问题的“重要性”。

摊销投影优化。现在，我们重新审视带有 $\mathcal{D}(X,\ Y) = \text{Max-SW}_p(P_X,\ P_Y)$ 的点云重建目标:

$$\min_{\phi\in\Phi,\gamma\in\Gamma}\mathbb{E}\left[\max_{\theta\in\mathbb{S}^{d-1}}\mathrm{W}_p(\theta\sharp P_X,\theta\sharp P_{g_\gamma(f_\phi(X))})\right], \quad (8)$$

其中，期望是相对于 $X\sim p(X)$ 而言的。对于每个点云 $X\in\mathcal{X}$，我们都需要通过迭代优化过程计算一个最大切片 Wasserstein 距离。因此，计算量很大。

Nguyen 与 Ho（2022a）的作者提出使用摊销优化（Shu，2017；Amos，2022）来加速该问题。不同于独立求解所有优化问题，他们训练了一个摊销模型，用于预测所有问题的最优解。更具体地说，给定一个参数函数 $a_\psi\colon \mathcal{X}\times\mathcal{X}\to\mathbb{S}^{d-1}$ $(\psi\in\Psi)$，摊销目标为:

$$\min_{\phi\in\Phi,\gamma\in\Gamma}\max_{\psi\in\Psi}\mathbb{E}\mathrm{W}_p(\theta_{\psi,\gamma,\phi}\sharp P_X,\theta_{\psi,\gamma,\phi}\sharp P_{g_\gamma(f_\phi(X))}), \quad (9)$$

其中，期望是相对于 $X\sim p(X)$而言的，并且$\theta_{\psi,\ \gamma,\ \phi} = a_\psi(X,\ g_\gamma(f_\phi(X)))$。上述优化通过交替随机（投影）梯度下降-上升算法求解。因此，在 ϕ 和 γ的每次更新迭代中，计算速度更快。值得注意的是，先前的工作（Nguyen 与 Ho，2022a）考虑的是生成模型应用，该应用不稳定且难以理解。在此，我们将该框架调整至点云重建应用，这更便于探索摊销优化的行为。关于使用 Max-SW 和摊销投影优化训练自编码器的算法，请参阅附录 A.3 中的算法 2-3。

摊销模型。(Nguyen 与 Ho，2022a) 的作者提出了三种摊销模型，其依据是关于线性模型的文献（Christensen，2002）。具体而言，线性摊销模型定义如下:

定义 2。给定 $X,\ Y\in\mathbb{R}^{dm}$，线性摊销模型 定义如下:

$$a_\psi(X,Y) := \frac{w_0 + X'w_1 + Y'w_2}{||w_0 + X'w_1 + Y'w_2||_2},$$

其中，X'和 Y'是大小为 $d\times m$ 的矩阵，由大小为 dm的拼接向量 X 和 Y重塑而成，$\psi = (w_0,\ w_1,\ w_2)$，其中 w_1，$w_2\in\mathbb{R}^m$，且 $w_0\in\mathbb{R}^d$。

类似地，广义线性摊销模型和非线性摊销模型通过在线性模型中引入非线性来定义。我们在附录 A.1 的定义 4-5 中回顾了广义线性摊销模型和非线性摊销模型的定义。

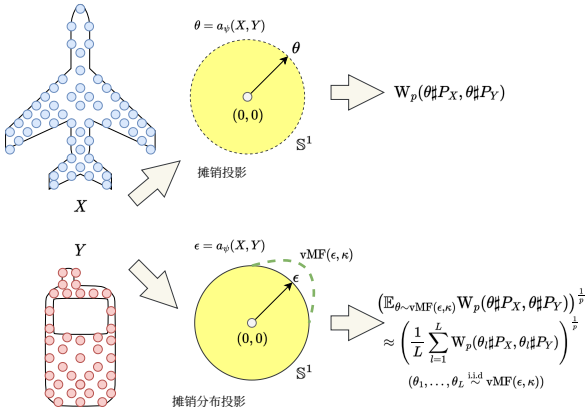

图 *2.* 摊销投影优化与摊销分布投影优化之间的差异。

次优性。 尽管速度更快，摊销优化通常无法恢复优化问题的全局最优解。也就是说，我们记

$$\theta^\star(X) = \mathrm{argmax}_{\theta\in\mathbb{S}^{d-1}}\mathrm{W}_p(\theta\sharp P_X, \theta\sharp P_{g_\gamma(f_\phi(X))})$$

和 $\psi^\star =$

$$\arg\max_{\psi\in\Psi}\mathbb{E}_{X\in p(X)}\left[\mathrm{W}_p(\theta_{\psi,\gamma,\phi}\sharp P_X, \theta_{\psi,\gamma,\phi}\sharp P_{g_\gamma(f_\phi(X))})\right].$$

然后，众所周知，对于度量 c： $\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}\to\mathbb{R}^+$，其摊销差距$\mathbb{E}_{X\sim p(X)}[c(\theta^\star(X)\ a_{\psi^\star}(X,\ g_\gamma(f_\phi(X))))]>0$。优秀的摊销模型应能够最小化摊销差距。然而，在摊销投影优化设置中，我们无法获得 $\theta^\star(X)$，因为投影梯度上升算法只能得到局部最优解。因此，对摊销差距进行细致研究颇具挑战性。

3. 自注意力摊销分布投影优化

在本节中，我们提出自注意力摊销分布投影优化框架。首先，我们在第3.1节介绍摊销分布投影优化，以保持度量性。随后，我们在第3.2节介绍具有对称性和置换不变性的自注意力摊销模型。

3.1. 摊销分布投影优化

当前的摊销投影优化框架旨在预测 Max-SW 中的“最大”投影方向。然而，投影一维 Wasserstein 距离仅在 Max-SW 的全局最优解处，才是概率测度空间上的度量。因此，投影次梯度上升算法（Nietert 等人，2022）得到的局部最优解，以及摊销模型的预测，对于投影瓦瑟斯坦距离而言，只能保证其伪度量性。

命题 *1.* 设投影一维 *Wasserstein* 距离为 $PW_p(\mu,\nu;\theta) = W_p(\theta\sharp\mu,\theta\sharp\nu))$。对于任意 $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)\ (p\geq 1, d\geq 1)$ 以及满足 $\hat{}\theta\neq\hat{}\arg\max_{\theta\in\mathbb{S}^{d-1}} W_p(\theta\sharp\mu,\theta\sharp\nu)$ 的 $\hat{}\theta\in\mathbb{S}^{d-1}$，$PW_p(\mu,\nu;\theta)$ 是 $\mathcal{P}_p(\mathbb{R}^d)$ 上的伪度量，因为它满足对称性、非负性和三角不等式； $\mu=\nu$ 蕴含 $PW_p(\mu,\nu;\theta)=0$，然而，$PW_p(\mu,\nu;\theta)=0$ 并不蕴含 $\mu=\nu$。

命题 1 的证明见附录 B.1。该结果表明，如果重建损失 $\hat{}\mathbb{E}_{X\sim p(X)}[\mathrm{PW}_p(P_X,\ P_{g_\gamma(f_\phi(X))};\theta(X))=0$，则无法推出对于 p-几乎处处的 $X\in\mathcal{X}$， $X=g_\gamma(f_\phi(X))$ 成立。因此，在 Max-SW 重建（公式 8）中，$\max_{\theta\in\mathbb{S}^{d-1}}$ 的局部最大值，以及在摊销 Max-SW 重建（模型设定错误时的公式 9）中，$\max_{\psi\in\Psi}$ 的全局最大值，即使其目标函数取得 0 值，也无法保证完美重建。

摊销分布投影优化。为解决这一问题，我们提出将公式 8 中的 Max-SW 替换为冯·米塞斯-费希尔分布切片瓦瑟斯坦（v-DSW）距离（Nguyen 等人，2021b）：

$$\min_{\phi\in\Phi,\gamma\in\Gamma}\mathbb{E}_{X\sim p(X)}\left[\max_{\epsilon\in\mathbb{S}^{d-1}}\left(\mathbb{E}_{\theta\sim\mathrm{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp P_X,\theta\sharp P_{g_\gamma(f_\phi(X))})\right)^{\frac{1}{p}}\right], \quad (10)$$

其中，$\mathrm{vMF}(\epsilon,\ \kappa)$ 是冯·米塞斯-费舍尔分布，其平均位置参数为 $\epsilon\in\mathbb{S}^{d-1}$，集中参数为 $\kappa>0$; $\text{v-DSW}_p(\mu,\ \nu;\kappa)\ =\max_{\epsilon\in\mathbb{S}^{d-1}}\left(\mathbb{E}_{\theta\sim\mathrm{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu,\ \theta\sharp\nu)\right)^{\frac{1}{p}}$ 是冯·米塞斯-费希尔分布切片瓦瑟斯坦距离。

该优化可通过结合 vMF 重参数化技巧的随机投影梯度上升算法求解。具体而言，通过重参数化的接受-拒绝采样（Davidson 等，2018a），从 $\mathrm{vMF}(\epsilon,\ \kappa)$ 中独立同分布地抽取 $\theta_1,\ \ldots,\ \theta_L\ (L\geq 1)$，并通过蒙特卡罗积分近似$\nabla_\epsilon\mathbb{E}_{\mathrm{vMF}(\epsilon,\kappa)}[\mathrm{W}_p^p(\theta\sharp\mu,\ \theta\sharp\nu)]$。关于 vMF 分布、其抽样算法、重参数化技巧及随机梯度估计量的更多细节，请参阅第 A.2 节。我们在图 2 中直观展示了新的摊销分布投影优化框架与传统摊销投影优化框架之间的区别。相应的摊销目标函数为：

$$\min_{\phi\in\Phi,\gamma\in\Gamma}\max_{\psi\in\Psi}\mathbb{E}_{X\sim p(X)}\left(\mathbb{E}_{\theta\sim\mathrm{vMF}(\epsilon_{\psi,\gamma,\phi},\kappa)}\mathrm{W}_p^p(\theta\sharp P_X,\theta\sharp P_{g_\gamma(f_\phi(X))})\right)^{\frac{1}{p}}, \quad (11)$$

其中 $\epsilon_{\psi,\ \gamma,\ \phi}\ =\ a_\psi(X,\ g_\gamma(f_\phi(X)))$。该优化问题通过一种交替随机（投影）梯度下降-上升算法求解，并采用冯·米塞斯–费舍尔分布重参数化。

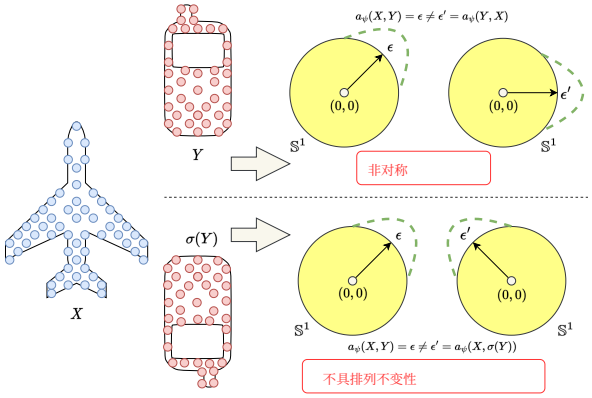

图 3. 二维空间中一个非对称且具置换不变性的摊销模型的可视化。

定理1。 对于任意 $\epsilon \in \mathbb{S}^{d-1}$ 和 $0 \leq \kappa < \infty$，若 $\mathbb{E}_{X\sim p(X)}\left(\mathbb{E}_{\theta\sim vMF(\epsilon,\kappa)}W_p^p(\theta\sharp P_X, \theta\sharp P_{g_\gamma(f_\phi(X))}\right)^{\frac{1}{p}} = 0, X = g_\gamma(f_\phi(X))$ 对 p-几乎处处的 $X \in \mathcal{X}$ 成立。

定理1的证明见附录 B.2。证明基于在满足 vMF 分布的光滑性条件时，证明非最优冯·米塞斯-费舍尔分布切片 Wasserstein 距离（v-DSW）具有度量性。值得注意的是，冯·米塞斯-费舍尔分布切片 Wasserstein 距离度量性的证明是新的，因为原始工作（Nguyen 等人，2021b）仅在全局最优条件下证明了其伪度量性。定理1表明，对于 v-DSW 重建（方程 10）中的 $\max_{\epsilon\in\mathbb{S}^{d-1}}$，只需达到局部最优解；对于摊销 v-DSW 重建（方程 11）中的 $\max_{\psi\in\Psi}$，也只需达到局部最优解，即可实现完美重建。

与 SW 和 Max-SW 的比较： 当 $\kappa \to 0$ 时，vMF 分布弱收敛到单位超球面上的均匀分布。因此，我们可以在公式 10 和公式 11 中分别恢复传统的切片瓦瑟斯坦重建。当 $\kappa \to \infty$ 时，vMF 分布弱收敛到位置参数处的狄拉克 δ。因此，我们分别在公式 10 和公式 11 中得到 Max-SW 重建和摊销 Max-SW 重建。然而，当 $0 < \kappa < \infty$ 时，v-DSW 重建和摊销 v-DSW 重建能够找到判别性投影方向的一个区域，同时保持度量性，从而实现完美重建。

3.2. 自注意力摊销模型

现在我们讨论用于摊销优化的摊销模型的参数化。

排列不变性与对称性。 设 X 和 Y 是两个点云，则 P_X 与 P_Y 之间 v-DSW 的最优切片分布 vMF($\epsilon^\star, \kappa$) 可通过运行附录 A.3 中的算法 4 得到。显然，vMF($\epsilon^\star, \kappa$) 对支撑点的置换具有不变性

因为对于置换函数 σ，$P_{\sigma(X)} = P_X$ 和 $P_{\sigma(Y)} = P_Y$。此外，交换 P_X和 P_Y 时，最优切片分布 vMF($\epsilon^\star, \kappa$) 也保持不变，因为 v-DSW 是对称的。然而，当前的摊销模型（见定义 2、附录 A.1 中的定义 4-5）不具置换不变性且非对称，即 $a_\psi(X, Y) \neq a_\psi(X, \sigma(Y))$ 和 $a_\psi(X, Y) \neq a_\psi(Y, X)$。因此，当前的摊销模型可能存在严重的设定错误。我们在图 3 中展示了一个不具对称性和排列不变性的摊销模型的可视化结果。为解决这一问题，我们提出了基于自注意力机制、具有对称性和排列不变性的摊销模型。

自注意力机制。 众所周知，当数据是文本等序列（Devlin 等人，2019年；Liu 等人，2019年；Brown 等人，2020年）或语音（Li 等人，2019年；Wang 等人，2020年b）时，注意力在学习长程依赖方面十分有效。随后，这一机制成功推广到了其他数据类型，包括图像（Carion 等人，2020年；Dosovitskiy 等人，2020年）、视频（Sun 等人，2019年）、图（Dwivedi 和 Bresson，2021）、点云（Zhao 等人，2021年；Guo 等人，2021年）等。现在我们重新审视注意力机制（Vaswani 等人，2017）。给定Q，$K \in \mathbb{R}^{m\times d_k}$，$V \in \mathbb{R}^{m\times d_v}$，缩放点积注意力算子定义如下：

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}_{\mathrm{row}}\left[\frac{QK^T}{\sqrt{d_k}}\right]V \qquad (12)$$

其中，$\mathrm{softmax}_{\mathrm{row}}$ 表示按行计算的 softmax 函数。在自注意力机制中，查询矩阵 Q、键矩阵 K 和值矩阵 V 通常通过将输入序列 X 投影到不同的子空间中计算得到。因此，自注意力机制定义如下。给定 $X \in \mathbb{R}^{m\times d}$，自注意力算子为：

$$\mathcal{A}_\zeta(X) = \mathrm{Att}(XW_q, XW_k, XW_v) \qquad (13)$$

其中 $W_q, W_k \in \mathbb{R}^{d\times d_k}, W_v \in \mathbb{R}^{d\times d_v}$ 和 $\zeta = (W_q, W_k, W_v)$。自注意力算子的二次内存和计算开销臭名昭著。具体而言，给定长度为 m 的输入序列，其时间和空间复杂度均为 $\mathcal{O}(m^2)$。由于我们关注切片 Wasserstein 设置，其计算复杂度至多应为 $\mathcal{O}(m\log m)$，因此传统自注意力并不适用。已有多项工作（Li 等人，2020年；Katharopoulos 等人，2020年；Wang 等人，2020年a；Shen 等人，2021年）致力于将总体复杂度从 $\mathcal{O}(m^2)$ 降至 $\mathcal{O}(m)$。本文采用两种具有线性复杂度的注意力变体：高效注意力（Shen 等人，2021年）和线性注意力（Wang 等人，2020年a）。给定 $X \in \mathbb{R}^{m\times d}$，高效自注意力 定义为：

$$\mathcal{EA}_\zeta(X) = \mathrm{softmax}_{\mathrm{row}}(XW_q)\left[\mathrm{softmax}_{\mathrm{col}}(XW_k)^T(XW_v)\right] \qquad (14)$$

其中 $W_q, W_k \in \mathbb{R}^{d\times d_k}, W_v \in \mathbb{R}^{d\times d_v}$，$\zeta = (W_q, W_k, W_v)$，$\text{softmax}_{\text{col}}$ 表示按列应用 softmax 函数。线性自注意力 为：

$$\mathcal{LA}_\zeta(X) = \text{Att}(XW_q, W_{k1}XW_{k2}, W_{v1}XW_{v2}) \quad (15)$$

其中 W_q，$W_{k2} \in \mathbb{R}^{d\times d_k}$，$W_{v2} \in \mathbb{R}^{d\times d_v}$、$W_{k1}$，$W_{v1} \in \mathbb{R}^{k\times n}$ 和 $\zeta = (W_q, W_{k1}, W_{k2}, W_{v1}, W_{v2})$。投影维度 k 的选择满足 $m \gg k$，从而显著降低内存和空间消耗。

自注意力摊销模型：基于自注意力机制，我们提出了自注意力摊销模型，该模型具有置换不变性和对称性。形式化地，自注意力摊销模型定义如下：

定义 3。给定 $X, Y \in \mathbb{R}^{dm}$，自注意力摊销模型定义如下：

$$a_\psi(X,Y) = \frac{\mathcal{A}_\zeta(X'^\top)^\top \mathbf{1}_m + \mathcal{A}_\zeta(Y'^\top)^\top \mathbf{1}_m}{||\mathcal{A}_\zeta(X'^\top)^\top \mathbf{1}_m + \mathcal{A}_\zeta(Y'^\top)^\top \mathbf{1}_m||_2}, \quad (16)$$

其中 X' 和 Y' 是大小为 $d\times m$ 的矩阵，它们由拼接后的大小为 dm 的向量 X 和 Y 重塑而成；$\mathbf{1}_m$ 是所有元素均为 1 的 m 维向量，且 $\psi = (\zeta)$。

通过将传统自注意力替换为线性自注意力和高效自注意力，我们得到线性自注意力摊销模型 和高效自注意力摊销模型。

命题 2. 自注意力摊销模型具有对称性，且置换不变。

命题 2 的证明见附录 B.3。对称性直接由自注意力摊销模型的定义得出。通过证明结合平均池化的自注意力算子具有置换不变性，即可证明其排列不变性。

与 Set Transformer 的比较。(Lee 等，2019b)的作者也提出了一种方法，以保证集合的置换不变性。我们的工作与他们的工作主要有两点不同。首先，Set Transformer 引入了一种新的注意力机制和一种新的 Transformer 架构，而我们仅提出了一种方法，将任意注意力机制应用于摊销模型，以保留其排列不变性。其次，Set Transformer 通过使用可学习的多头注意力作为聚合方案来维持排列不变性。相比之下，我们仍采用平均池化这一传统的置换不变聚合方案，来汇总自注意力操作所学习到的特征。尽管如此，我们的工作与 Set Transformer 是正交的；换句话说，可以将 Set Transformer 中的技术应用于我们基于注意力的摊销模型。我们将这项研究留待未来工作。

表 *1.*ModelNet40 数据集上的重建与迁移学习性能。倒角距离和切片瓦瑟斯坦距离均乘以 100。

方法	倒角距离(10^{-2},↓)	切片瓦瑟斯坦距离(10^{-2},↓)	EMD(↓)	准确率(↑)	时间 (↓)
CD	1.25 ± 0.03	681.20 ± 16.73	653.52 ± 10.43	86.28 ± 0.34	95
EMD	0.40 ± 0.00	94.54 ± 2.90	168.60 ± 1.57	88.45 ± 0.20	208
SW	0.68 ± 0.01	89.61 ± 3.88	191.12 ± 2.88	87.90 ± 0.27	106
Max-SW	0.68 ± 0.01	88.22 ± 1.45	190.23 ± 0.1	87.97 ± 0.14	116
ASW	0.69 ± 0.01	89.42 ± 5.07	192.03 ± 3.09	87.78 ± 0.20	103
v-DSW	**0.67 ± 0.00**	85.03 ± 3.31	187.75 ± 2.00	87.83 ± 0.40	633
$\mathcal{L}$-Max-SW	1.06 ± 0.03	121.85 ± 5.77	236.87 ± 3.42	87.70 ± 0.23	**94**
$\mathcal{G}$-Max-SW	12.11 ± 0.29	851.07 ± 2.11	829.28 ± 5.53	87.49 ± 0.36	97
$\mathcal{N}$-Max-SW	7.38 ± 3.29	618.74 ± 153.87	648.32 ± 117.03	87.43 ± 0.15	96
$\mathcal{L}$v-DSW	0.68 ± 0.00	85.32 ± 0.54	188.32 ± 0.23	87.70 ± 0.34	114
$\mathcal{G}$v-DSW	0.68 ± 0.01	82.77 ± 0.48	187.04 ± 1.11	87.75 ± 0.19	117
$\mathcal{N}$v-DSW	**0.67 ± 0.00**	83.47 ± 0.49	186.66 ± 0.81	87.84 ± 0.07	115
$\mathcal{A}$v-DSW	0.67 ± 0.01	83.08 ± 1.22	186.27 ± 0.56	88.05 ± 0.17	230
$\mathcal{E}\mathcal{A}$v-DSW	0.68 ± 0.01	82.05 ± 0.40	186.46 ± 0.25	88.07 ± 0.21	125
$\mathcal{L}\mathcal{A}$v-DSW	0.68 ± 0.00	**81.03 ± 0.18**	**185.26 ± 0.31**	**88.28 ± 0.13**	123

表 *2.* 近似冯·米塞斯-费舍尔分布切片瓦瑟斯坦距离（v-DSW）时摊销模型的比较。T 表示投影次梯度上升迭代的次数。

方法	T	距离 (↑)	时间 (↓)
$\mathcal{L}$v-DSW	1	52.73	0.06
$\mathcal{G}$v-DSW	1	50.73	0.07
$\mathcal{N}$v-DSW	1	51.89	0.07
$\mathcal{A}$v-DSW	1	53.07	1.00
$\mathcal{E}\mathcal{A}$v-DSW	1	53.17	0.17
$\mathcal{L}\mathcal{A}$v-DSW	1	**53.83**	0.14
v-DSW	1	51.87	0.1
v-DSW	5	51.90	0.33
v-DSW	10	52.65	0.5
v-DSW	50	53.16	2.00
v-DSW	100	54.39	4.00

4. 实验

为验证我们所提方法的有效性，我们在点云重建任务及其两个下游任务——迁移学习和点云生成——上对方法进行了评估。我们希望回答三个重要问题：

1. 摊销 Max-SW 在处理点云时是否会出现次优性问题? 将 Max-SW 替换为 v-DSW 是否能缓解这一问题?

2. 所提出的摊销分布投影优化框架是否比传统摊销投影优化框架以及常用距离（例如 Chamfer 距离、地球移动距离（Wasserstein 距离）、切片瓦瑟斯坦距离、Max-SW、自适应 SW（ASW）（Nguyen 等，2021c）和 v-DSW）表现更好?

3. 自注意力摊销模型是否优于 Nguyen 与 Ho（2022a）中先前设定错误的摊销模型?



图4. ShapeNet Core-55 数据集中的点云重建定性结果。从上到下依次为输入点云、切片瓦瑟斯坦距离、Max-SW (T = 50)、v-DSW (T = 50) 和 $\mathcal{LA}$v-DSW。

表3.ShapeNet 椅子类别点云生成的性能比较。JSD、MMD-CD 和 MMD-EMD 的值均乘以 100。

方法	JSD ($\times 10^{-2}$, ↓)	MMD ($\times 10^{-2}$, ↓)		覆盖率 (% ↑)		1-NNA (% ↓)	
		CD	EMD	CD	EMD	CD	EMD
CD	17.88 ± 1.14	1.12 ± 0.02	17.19 ± 0.36	23.73 ± 1.69	10.83 ± 0.89	98.45 ± 0.10	100.00 ± 0.00
EMD	5.15 ± 1.52	0.61 ± 0.09	10.37 ± 0.61	41.65 ± 2.19	42.54 ± 2.42	87.76 ± 1.46	87.30 ± 1.22
SW	1.56 ± 0.06	0.72 ± 0.02	10.80 ± 0.11	38.55 ± 0.43	45.35 ± 0.48	89.91 ± 1.17	88.28 ± 0.70
Max-SW	1.63 ± 0.32	0.74 ± 0.01	10.84 ± 0.08	40.47 ± 1.04	47.81 ± 0.78	91.46 ± 0.72	89.93 ± 0.86
ASW	1.75 ± 0.38	0.78 ± 0.05	11.27 ± 0.38	38.16 ± 2.15	45.45 ± 1.40	91.21 ± 0.40	89.36 ± 0.40
v-DSW	1.79 ± 0.17	0.72 ± 0.02	10.73 ± 0.20	37.76 ± 0.71	45.49 ± 1.37	90.23 ± 0.13	88.33 ± 0.95
$\mathcal{L}$v-DSW	1.67 ± 0.07	0.77 ± 0.04	11.10 ± 0.33	37.91 ± 1.84	45.64 ± 2.30	90.42 ± 0.53	88.82 ± 0.38
$\mathcal{G}$v-DSW	1.56 ± 0.22	0.75 ± 0.02	10.99 ± 0.11	37.81 ± 1.70	45.69 ± 0.46	90.32 ± 0.38	88.26 ± 0.28
$\mathcal{N}$v-DSW	**1.44 ± 0.06**	0.75 ± 0.02	10.95 ± 0.09	38.40 ± 1.34	46.28 ± 2.06	90.15 ± 0.80	88.65 ± 0.82
$\mathcal{EA}$v-DSW	1.73 ± 0.21	**0.71 ± 0.04**	**10.70 ± 0.26**	40.03 ± 1.28	**48.01 ± 1.07**	89.98 ± 0.57	88.55 ± 0.38
$\mathcal{LA}$v-DSW	1.54 ± 0.09	0.72 ± 0.03	10.74 ± 0.35	**40.62 ± 1.39**	45.84 ± 1.23	**89.44 ± 0.28**	**87.79 ± 0.37**

实验设置：我们的设置[1]，见附录 C.1，与 ASW 论文中的设置相同。我们将摊销 v-DSW 方法与以下损失函数进行比较：Chamfer 差异（倒角距离），地球移动距离（EMD）、切片瓦瑟斯坦距离（SW）、Max-SW、自适应 SW（ASW），v-DSW 和摊销 Max-SW 变体。对于摊销模型，我们考虑了 6 种不同的模型。前缀 $\mathcal{L}, \mathcal{G}$ 和 $\mathcal{N}$ 分别表示 (Nguyen 与 Ho，2022a) 中的线性、广义线性和非线性摊销模型。 $\mathcal{A}$、$\mathcal{EA}$ 和 $\mathcal{LA}$ 分别表示自注意力、高效自注意力和线性自注意力。基线距离和摊销模型的实现细节分别见附录 C.2 和 C.3。每项实验均

[1]论文代码已发布于 https://github.com/hsgser/Self-Amortized-DSW。

使用三个不同的随机种子运行。我们报告每个实体的平均性能以及标准差。所有实验均在 NVIDIA V100 GPU 上运行。

与倒角距离和 EMD 的比较：本文主要关注将新的摊销框架与传统摊销框架及切片瓦瑟斯坦变体进行比较。本文仅为完整起见提供倒角距离和 EMD 的结果。此外，我们发现 ASW论文中对 EMD 与切片瓦瑟斯坦变体的比较并不公平。具体而言，EMD损失按点云中的点数进行了归一化，而切片瓦瑟斯坦变体没有进行归一化。为解决上述问题，我们修改了 EMD损失的实现，将其乘以点数（本例中为 2048）。作为一种“理想”目标，EMD 的表现优于所有切片瓦瑟斯坦变体。然而，与切片瓦瑟斯坦变体相比，EMD 的计算开销巨大

点云重建：遵循 ASW（Nguyen 等，2021c），我们使用三种差异度量——Chamfer 距离（倒角距离）、切片 Wasserstein 距离（切片瓦瑟斯坦距离）和 EMD——来衡量不同自编码器在 ModelNet40 数据集（Wu 等人，2015年）上的重建性能。定量结果汇总于表 1。对于每种方法，我们仅报告所有超参数选项中表现最佳（以 EMD 为依据）的模型。完整的定量结果（包括标准差）见表 4。我们的方法在这三种差异度量上均取得了最佳表现。相比之下，采用摊销 Max-SW损失的自编码器在此情形下表现不佳，这是由于我们在第 2.3 节讨论的次优性和度量性缺失问题。此外，摊销 v-DSW损失在 3 次运行中的标准差小于 v-DSW。此外，与使用投影次梯度上升算法（例如 Max-SW 和 v-DSW）进行传统计算相比，采用摊销优化可以缩短训练时间。例如，使用 $\mathcal{LA}$v-DSW训练自编码器的一次迭代仅需 123 秒，而使用 v-DSW 则需 633 秒。对于摊销模型而言，基于注意力的摊销模型能够使重建结果与输入之间的 EMD 更低。图4展示了定性结果，表明我们的方法能够成功重建三维点云。完整的定性结果见图6。

摊销间隙：为验证自注意力摊销模型相较于先前设定错误的摊销模型的优势，我们比较了它们近似 v-DSW 的效果。我们从 ShapeNet Core-55 数据集中采样 1000 对点云，构建了一个数据集。由于求解摊销优化时存在内存限制，我们将数据集分为 10 个批次，每批 100 对。我们计算了所有点云对之间的 v-DSW 及其摊销版本，并在表 2 中报告了它们的平均损失值。与先前设定错误的摊销模型相比，基于注意力的摊销模型产生的损失更高，也更接近 v-DSW 的常规计算结果（$T = 100$）。要达到高效/线性自注意力摊销模型的同等水平，需要运行 50 次以上的次梯度迭代，速度慢了 10 倍以上。

迁移学习：我们进一步将上述自编码器学习到的潜向量输入分类器。遵循 ASW论文中的设置，我们以 256 的批量大小训练分类器 500 个周期。优化器与重建实验中使用的相同。表1展示了分类结果。同样，我们发现，使用自注意力摊销 v-DSW 时，准确率有所提升。

点云生成：我们还在三维点云生成任务上评估了我们的方法。遵循（Achlioptas 等人，2018年），我们将 ShapeNet 的椅子类别按 85/5/10 的比例划分为训练集、验证集和测试集。我们在训练集上训练每个自编码器 100 个周期，并在验证集上进行评估。随后，使用该自编码器学习到的潜在编码训练生成器，使其生成潜在编码，这与（Achlioptas 等人，2018年）相同。评估时，我们使用（Yang 等，2019a）中的同一组度量。测试集的定量结果见表3。我们的方法在所有度量上均取得了最佳表现。此外，除 JSD 外，基于注意力的摊销模型在所有度量上的表现都优于以往的摊销模型。完整的定量结果见表9。

5. 结论

我们提出了一种自注意力摊销分布投影优化框架，该框架使用自注意力摊销模型，为每一对概率测度预测投影方向上的最佳判别分布。高效自注意力机制有助于将几何归纳偏置——排列不变性和对称性——注入摊销模型，同时保持较低的计算量。此外，摊销分布投影优化框架保证了所有概率测度对的度量性，尽管摊销差距仍然存在。在实验方面，我们将新提出的框架与传统的摊销投影优化框架以及其他广泛使用的距离度量进行比较，并在点云重建应用及其两个下游任务（迁移学习和点云生成）中验证了该框架的优越性能。

致谢

Nhat Ho 感谢 NSF IFML 2019844 和 NSF 人工智能基础机器学习研究所提供的支持。

参考文献

Achlioptas, P.、Diamanti, O.、Mitliagkas, I. 和 Guibas, L. 学习三维点云的表示和生成模型。发表于国际机器学习会议，第 40–49 页。PMLR，2018年。

Alvarez-Melis, D. 和 Fusi, N. 基于最优传输的几何数据集距离。神经信息处理系统进展，33:21428–21439，2020年。

Amos, B. 关于用于在连续域上学习优化的摊销优化的教程。*arXiv* 预印本 *arXiv:2202.00665*，2022。

Arjovsky, M.、Chintala, S. 和 Bottou, L. 瓦瑟斯坦生成对抗网络。在国际机器学习会议上，第 214–223 页，2017。

Barrow, H. G., Tenenbaum, J. M., Bolles, R. C. 和 Wolf, H. C. 参数化对应与 Chamfer 匹配：两种新的图像匹配技术。技术报告，SRI INTERNATIONAL MENLO PARK CA 人工智能中心，1977年。

Bhushan Damodaran, B., Kellenberger, B., Flamary, R., Tuia, D. 和 Courty, N. Deepjdot：用于无监督领域自适应的深度联合分布最优传输。载于欧洲计算机视觉会议（*ECCV*）论文集，第 447–463 页，2018年。

Bonneel, N., Rabin, J., Peyr´e, G. 和 Pfister, H. 测度的切片和拉东瓦瑟斯坦重心。数学成像与视觉杂志，1(51):22–45，2015年。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shya m, P., Sastry, G., Askell, A. 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Carion, N.、Massa, F.、Synnaeve, G.、Usunier, N.、Kirillov, A. 和 Zagoruyko, S. 使用变换器进行端到端目标检测。载于欧洲计算机视觉会议，第 213–229 页。Springer，2020年。

Chang, A. X.、Funkhouser, T.、Guibas, L.、Hanrahan, P.、Huang, Q.、Li, Z.、Savarese, S.、Savva, M.、Song, S.、Su, H.、Xiao, J.、Yi, L. 和 Yu, F. ShapeNet：信息丰富的三维模型库。技术报告 arXiv:1512.03012 [cs.GR],斯坦福大学 — 普林斯顿大学 — 芝加哥丰田技术学院，2015 年。

Christensen, R. 复杂问题的平面答案，第 35 卷。Springer，2002。

Courty , N., Flamary , R., Tuia, D., 和 Rakotomamonjy , A 。用于领域自适应的最优传输。*IEEE* 模式分析与机器智能汇刊，39(9)：1853–1865，2016年。Cuturi, M. Sinkhorn距离：以光速计算最优传输。载于神经信息处理系统进展，第 2292–2300 页，2013年。Damodaran, B. B.、Kellenberger, B.、Flamary, R.、Tuia, D. 和 Courty, N. Deepjdot：用于无监督域适应的深度联合分布最优传输。载于欧洲计算机视觉会议（*ECCV*）论文集，第 447–463 页，2018年。

Davidson,T. R.,Falorsi,L.,De Cao,N.,Kipf, T.,and Tom- czak, J. M. 超球面变分自编码器。载于第*34*届人工智能不确定性会议 *2018*，*UAI 2018*，第856–865页。人工智能不确定性协会（AUAI），2018a。

Davidson, T. R., Falorsi, L., De Cao, N., and Tomczak, T. K. J. M. 超球面变分自编码器。载于人工智能不确定性会议（*UAI*），2018b。

Deshpande, I., Zhang, Z., and Schwing, A. G. 使用切片 Wasserstein 距离进行生成建模。载于*IEEE*计算机视觉与模式识别会议论文集，第3483–3491页，2018年。

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., 和 Schwing, A. G. 最大切片Wasserstein距离及其在生成对抗网络中的应用。载于*IEEE*计算机视觉与模式识别会议论文集，第10648–10656页，2019年。

Devlin, J., Chang, M.-W., Lee, K., 和 Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。载于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。doi：10.18653/v1/N19-1423。URL https://aclanthology.org/N19-1423。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等。《一幅图像胜过 16×16 个词：用于大规模图像识别的变换器》。*arXiv* 预印本 *arXiv:2010.11929*，2020年。

Dwivedi, V. P. 和 Bresson, X. 将 Transformer 网络推广到图。*AAAI* 图深度学习方法与应用研讨会，2021年。

Genevay, A.、Peyr´e, G. 和 Cuturi, M. 使用Sinkhorn散度学习生成模型。载于人工智能与统计国际会议，第 1608–1617 页。PMLR，2018年。

Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R., a nd Hu, S.-M. Pct：点云Transformer。计算视觉媒体，7(2):187–199，2021年4月。ISSN 2096- 0662。doi: 10.1007/s41095-021-0229-5。URL http://dx.doi.org/10.1007/s41095-021-0229-5。

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. 通过瓦瑟斯坦均值进行多层聚类。载于国际机器学习会议，第1501–1509页，2017年。

Huynh, V.、Zhao, H. 和 Phung, D. OTLDA：一种用于主题建模的几何感知最优传输方法。神经信息处理系统进展，33：18573–18582，2020年。

Jupp, P. E. 和 Mardia, K. V. 矩阵 von Mises-Fisher 分布和宾汉分布的最大似然估计量。《统计年鉴》，7(3)：599–606，1979年。

Katharopoulos, A.、Vyas, A.、Pappas, N. 和 Fleuret, F. Transformer 即 RNN：采用线性注意力的快速自回归变换器。发表于国际机器学习会议，第 5156–5165 页。PMLR，2020年。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. 切片瓦瑟斯坦自编码器。发表于国际表示学习会议，2018年。

Lee, C.-Y., Batra, T., Baig, M. H., and Ulbricht, D. 用于无监督域自适应的切片瓦瑟斯坦差异。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10285–10295页，2019a。

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set Transformer：一种用于基于注意力的置换不变神经网络的框架。发表于国际机器学习会议，第3744–3753页。PMLR，2019b。

Lee, Y., Kim, S., Choi, J., and Park, F. 一种用于点云数据的统计流形框架。发表于国际机器学习会议，第12378–12402页。PMLR，2022。

Li, R., Su, J., Duan, C. 和Zheng, S. 线性注意力机制：一种用于语义分割的高效注意力。*arXiv* 预印本 *arXiv:2007.14902*，2020年。

Li, S., Raj, D., Lu, X., Shen, P., Kawahara, T. 和 Kawai, H. 通过压缩结构和语音属性增强改进基于Transformer的语音识别系统。载于语音国际会议论文集 *2019*年，第4400–4404 页，2019年。doi: 10.21437/Interspeech.2019-2112。URLhttp://dx.doi.org/ 10.21437/Interspeech.2019-2112。

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa：一种鲁棒优化的 BERT 预训练方法。*arXiv*预印本 *arXiv:1907.11692*，2019年。

Naderializadeh, N., Comer, J., Andrews, R., Hoffmann, H., and Kolouri, S. 基于切片Wasserstein嵌入的池化。神经信息处理系统进展，34，2021年。

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., a nd S.ims.ekli, U. 基于切片Wasserstein距离的近似贝叶斯计算。载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第5470–5474页。IEEE，2020年。

Nguyen, K. 和 Ho, N. 面向切片Wasserstein生成模型的摊销投影优化。神经信息处理系统进展，2022a。

Nguyen, K. 和 Ho, N. 重新审视图像上的切片Wasserstein：从向量化到卷积。神经信息处理系统进展，2022b。

Nguyen, K.、Ho, N.、Pham, T. 和 Bui, H. 分布式切片Wasserstein及其在生成模型中的应用。发表于国际学习表征会议，2021a。

Nguyen, K.、Nguyen, S.、Ho, N.、Pham, T. 和 Bui, H. 利用球面切片融合 Gromov Wasserstein 改进关系正则化自编码器。发表于国际学习表征会议，2021b。

Nguyen, K.、Ren, T.、Nguyen, H.、Rout, L.、Nguyen, T. M. 和 Ho, N. 层次切片 Wasserstein 距离。发表于第十一届国际学习表征会议，2023。

Nguyen, T., Pham, Q.-H., Le, T., Pham, T., Ho, N., andHua, B.-S. 用于学习三维点云表示的点集距离。载于 *IEEE/CVF* 国际计算机视觉会议（*ICCV*）论文集，2021c。

Nietert, S., Sadhu, R., Goldfeld, Z. 和 Kato, K. 切片 Wasserstein 距离的统计、鲁棒性与计算保证。神经信息处理系统进展，2022。

Peyr´e, G. 和 Cuturi, M. 计算最优传输：及其在数据科学中的应用。机器学习基础与趋势®，11(5-6):355–607，2019年。

Pham, Q.-H., Uy, M. A., Hua, B.-S., Nguyen, D. T., Roig, G., 和 Yeung, S.-K. LCD：用于二维-三维匹配的学习式跨域描述符。载于 *AAAI*人工智能大会，2020年。

Qi, C. R., Su, H., Mo, K., 和 Guibas, L. J. Pointnet：用于三维分类与分割的点集深度学习。载于 *IEEE*计算机视觉与模式识别会议论文集，第 652–660 页，2017。

Shen, Z., Zhang, M., Zhao, H., Yi, S., 和 Li, H. 高效注意力：具有线性复杂度的注意力。载于*IEEE/CVF* 冬季计算机视觉应用会议论文集，第 3531–3539 页，2021。Shu, R. 摊销优化
h
t
t
p://ruishu.io/2017/11/07/a
mortized-optimization/。个人博客，2017。Sra, S. 机器学习中的方向统计学：简要综述。*arXiv* 预印本 *arXiv:1605.00316*，2016。Sun, C., Myers, A., Vondrick, C., Murphy, K., 和 Schmid, C. VideoBERT：用于视频与语言表示学习的联合模型。载于*IEEE/CVF* 国际计算机视觉会议，第 7464–7473 页，2019。Temme, N. M. 特殊函数：数学物理经典函数导论。John Wiley & Sons，2011。Tolstikhin, I., Bousquet, O., Gelly, S., 和 Schoelkopf, B. 瓦瑟斯坦自编码器。载于国际学习表征会议，2018。Vaswani, A., Shazeer, N., Par
mar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., 和 Polosukhin, I. 注意力就是你所需要的一切。神经信息处理系统进展，30，2017。Villani, C. 最优传输：新旧理论，第 338 卷。Springer科学与商业媒体，2008年。Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., 和 Courty, N. 结合最优传输距离的基于模板的图神经网络。神经信息处理系统进展，2022。Wang, S., Li, B. Z., Khabsa, M., Fang, H., 和 Ma, H. Linformer：具有线性复杂度的自注意力。*arXiv* 预印本 *arXiv:2006.04768*，2020a。

Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., Zhang, F., 等。用于混合语音识别的基于Transformer的声学建模。载于*ICASSP 2020-2020 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6874–6878 页。IEEE，2020b。

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., 和 Xiao, J. 3D ShapeNets：体积形状的深度表征。载于 *IEEE*计算机视觉与模式识别会议论文集，第 1912–1920 页，2015年。

Yang, G., Huang, X., Hao, Z., Liu, M.-Y., Belongie, S., and Hariharan, B. Pointflow：基于连续归一化流的 3D 点云生成。发表于*IEEE/CVF* 国际计算机视觉会议论文集，第 4541–4550 页，2019a。Yang, J., Zhang, Q., Ni, B., Li, L., Liu, J., Zhou, M., and Tian, Q. 利用自注意力和 Gumbel子集采样对点云建模。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 3323–3332 页，2019b。Yi, M. and Liu, S. 切片瓦瑟斯坦变分推断。发表于近似贝叶斯推断进展研讨会第四届会议，2021年。Zhao, H., Jiang, L., Jia, J., Torr, P. H., and Koltun, V. Point Transformer。发表于*IEEE/CVF* 国际计算机视觉会议论文集，第 16259–16268 页，2021年。

《用于切片瓦瑟斯坦点云重建的自注意力摊销分布投影优化》补充材料

在本补充材料中，我们首先在附录 A 中提供一些补充材料，包括附录 A.1 中广义线性摊销模型和非线性摊销模型的定义、附录 A.2 中 von Mises-Fisher 分布切片瓦瑟斯坦的计算细节，以及附录 A.3 中自编码器的训练算法。接下来，我们在附录 B 中汇总正文中略去的证明。随后，我们在附录 C 中讨论实验设置。最后，我们在附录 D 中展示其他实验结果。

A. 补充材料

A.1. 摊销模型

现在我们回顾　广义线性摊销模型和非线性摊销模型（Nguyen 与 Ho，2022a　）.

定义 4。给定 X、$Y \in \mathbb{R}^{dm}$，广义线性摊销模型定义如下：

$$f_\psi(X,Y) := \frac{g_{\psi_1}(X)'w_1 + g_{\psi_1}(Y)'w_2}{||g_{\psi_1}(X)'w_1 + g_{\psi_1}(Y)'w_2||_2^2}, \tag{17}$$

其中，X' 和 Y' 是大小为 $d \times m$ 的矩阵，由大小为 dm、w_1、$w_2 \in \mathbb{R}^m$、$w_0 \in \mathbb{R}^d$、$\psi_1 \in \Psi_1$、g_{ψ_1}的拼接向量 X 和 Y 重塑而成：$\mathbb{R}^{dm} \to \mathbb{R}^{dm}$、$\psi = (w_0, w_1, w_2, \psi_1)$，以及 $g_{\psi_1}(X) = (x'_1, ..., x'_m)$ 和$g_{\psi_1}(Y) = (y'_1, ..., y'_m)$。具体而言，令$g_{\psi_1}(X) = (W_2\sigma(W_1x_1) + b_0, ..., W_2\sigma(W_1x_m) + b_0)$，其中 $\sigma(\cdot)$ 是 Sigmoid 函数，$W_1 \in \mathbb{R}^{d\times d}$、$W_2 \in \mathbb{R}^{d\times d}$和 $b_0 \in \mathbb{R}^d$。

定义 5。给定 $X, Y \in \mathbb{R}^{dm}$，非线性摊销模型定义如下：

$$f_\psi(X,Y) := \frac{h_{\psi_2}(X'w_1 + Y'w_2)}{||h_{\psi_2}(X'w_1 + Y'w_2)||_2^2}, \tag{18}$$

其中，X' 和 Y' 是大小为 $d \times m$ 的矩阵，由大小为 dm的拼接向量 X 和 Y 重塑而成，w_1，$w_2 \in \mathbb{R}^m$, $\psi_2 \in \Psi_2$, h_{ψ_2}：$\mathbb{R}^d \to \mathbb{R}^d$, $\psi = (w_1, w_2, \psi_2)$，以及$h_{\psi_2}(x) = W_4\sigma(W_3x)) + b_0$ 其中 $\sigma(\cdot)$是 Sigmoid 函数。

A.2. von Mises-Fisher 分布切片 Wasserstein 距离

我们首先从冯·米塞斯-费舍尔分布（vMF）的定义开始。冯·米塞斯-费舍尔分布（*vMF*）（Jupp & Mardia，1979年）是定义在单位超球面 $\mathbb{S}^{d-1}$ 上的概率分布，其密度函数为：

$$f(x|\epsilon,\kappa) := C_d(\kappa)\exp(\kappa\epsilon^\top x), \tag{19}$$

其中，$\epsilon \in \mathbb{S}^{d-1}$ 是位置向量，$\kappa \geq 0$ 是浓度参数，$C_d(\kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}$ 是归一化常数。这里，I_v 是阶数为 v 时的第一类修正贝塞尔函数（Temme，2011年）。

vMF分布是一种连续分布，其概率质量集中在均值 ϵ 周围，并且当x 远离 ϵ 时，其概率密度会降低。当 $\kappa \to 0$ 时，vMF在分布意义下收敛于 $\mathcal{U}(\mathbb{S}^{d-1})$；当 $\kappa \to \infty$ 时，vMF在分布意义下收敛于以 ϵ 为中心的狄拉克分布（Sra，2016）。

重参数化拒绝采样：vMF分布的采样过程基于拒绝采样程序。我们在算法1中回顾采样过程（Davidson等人，2018a；Nguyen 等人，2021b）。该算法对提议分布进行重参数化。现在，我们针对一般函数 $f(\theta)$ 推导 $\nabla_\epsilon \mathbb{E}_{\text{vMF}(\theta|\epsilon,\kappa)}[f(\theta)]$的梯度估计量，以便求出优化问题 $\max_{\epsilon\in\mathbb{S}^{d-1}} \mathbb{E}_{\text{vMF}(\theta|\epsilon,\kappa)}[f(\theta)]$中的最大值 ϵ^*。

在 $d > 0$ 维度中，令（ϵ，κ）为vMF分布的参数。我们记 $b = \frac{-2\kappa+\sqrt{4\kappa^2+(d-1)^2}}{d-1}$ 为两个条件分布：$g(\omega \mid \kappa) = \frac{2b^{1/2(d-1)}}{(\pi^{d/2})\Gamma(d/2)\mathcal{C}_d(\kappa)\exp(\omega\kappa)}(1-\omega^2)^{\frac{1}{2}(d-3)}$Beta$(\frac{1}{2}(d-1), \frac{1}{2}(d-1))$，$r(\omega|\kappa) = \frac{1-\omega^{2\,1/2(d-3)}}{\omega|^{d-1}}$，分布 $s(\psi) := $Beta$(\frac{1}{2}(d-1), \frac{1}{2}(d-1))$，函数 $h(\psi, \kappa) = \frac{1-(\frac{1+b}{1+b}\psi)}{[\,1+b\,-\,1-b}$，分布 $\pi_1(\psi|\kappa) = s(\psi)\,g(h(\psi,\kappa)|\kappa)\,r(h(\psi,\kappa)|\kappa)$，$\pi_2(v) := \mathcal{U}(\mathbb{S}^{d-2})$，以及函数 $T(\omega, \overset{1-}{v}, \overset{1-b}{\epsilon}) = \left(I - 2\frac{e_1-\epsilon}{||e_1-\epsilon||_2}\frac{e_1-\epsilon}{||e_1-\epsilon||_2}^\top\right)(\omega, \sqrt{1-\omega^2}v^\top)^\top := = \theta$.

算法1从vMF分布中采样

输入：位置 ϵ、浓度 κ、维度 d、单位向量 $e_1 = (1, 0, .., 0)$ 抽取 $v \sim \mathcal{U}(\mathbb{S}^{d-2})$

$b \leftarrow$ $-2\kappa+\sqrt{4\kappa^2+(\quad d-1\ ^2)}$, $a \leftarrow (d-1)+2\kappa+\sqrt{4\kappa^2+(\quad d-1\ ^2)}$, $m \leftarrow$ $\frac{4ab}{(1+b)}$ $(d-1)\log(d-1)$ 重复 抽取 $\psi \sim$ Beta $(\frac{1}{2}(d-1)$、$\frac{1}{2}(d-1))\omega \leftarrow h(\psi, \kappa) =$ $1-$ $(\quad 1+b\ \psi\ )$ $(\ )\psi^t \leftarrow$ $1-\frac{2ab}{1-b}$ $(\ )\psi$抽取 $u \sim \mathcal{U}([0, 1])$, 直到 $(d-1)\log(t) - t + m \geq$ $\log(u)h_1 \leftarrow (\omega, \sqrt{1-\omega^2}v^\top)^\top\epsilon' \leftarrow e_1 - \epsilon^u = \frac{\epsilon'}{||\epsilon'||_2}$ $U = I - 2uu^\top$输出： Uh_1

我们可以通过以下引理2（Davidson等，2018b）得到梯度估计量：

$$\nabla_\epsilon \mathbb{E}_{\text{vMF}(\theta|\epsilon,\kappa)}\big[f(\theta)\big] = \nabla_\epsilon \mathbb{E}_{(\psi,v)\sim\pi_1(\psi|\kappa)\pi_2(v)}\Big[f\big(T(h(\psi,\kappa),v,\epsilon)\big)\Big] = \mathbb{E}_{(\psi,v)\sim\pi_1(\psi|\kappa)\pi_2(v)}\Big[\nabla_\epsilon f\big(T(h(\psi,\kappa),v,\epsilon)\big)\Big].$$

在 v-DSW 的情况下，我们有 $f(\theta) = \text{W}_p^p(\theta\sharp\mu,\ \theta\sharp\nu)$。因此，我们有：

$$\nabla_\epsilon \mathbb{E}_{\text{vMF}(\theta|\epsilon,\kappa)}\big[\text{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\big] = \mathbb{E}_{(\psi,v)\sim\pi_1(\psi|\kappa)\pi_2(v)}\Big[\nabla_\epsilon \text{W}_p^p(f\big(T(h(\psi,\kappa),v,\epsilon)\sharp\mu, f\big(T(h(\psi,\kappa),v,\epsilon)\sharp\nu\big)\big)\Big].$$

然后，我们可以使用蒙特卡洛估计方案得到梯度估计量：

$$\nabla_\epsilon \mathbb{E}_{\text{vMF}(\theta|\epsilon,\kappa)}\big[\text{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\big] \approx \frac{1}{L}\sum_{i=1}^{L}\Big[\nabla_\epsilon \text{W}_p^p(f\big(T(h(\psi_i,\kappa),v_i,\epsilon)\sharp\mu, f\big(T(h(\psi_i,\kappa),v_i,\epsilon)\sharp\nu\big)\big)\Big],$$

其中 $\{\psi_i\}_{i=1}^L \sim \pi_1(\psi|\kappa)$ 是独立同分布的，$\{v_i\}_{i=1}^L \sim \pi_2(v)$ 是独立同分布的，且 L 是投影的数量。从 $\pi_1(\psi|\kappa)$ 中采样等价于 vMF 采样过程中的接受-拒绝方案；采样 $\pi_2(v)$ 则直接来自 $\mathcal{U}(\mathbb{S}^{d-2})$。值得注意的是，$\nabla_\kappa \mathbb{E}_{\text{vMF}(\theta|\epsilon,\kappa)}\big[f(\theta)\big]$ 的梯度估计量可以通过对数导数技巧推导得到；不过，由于我们在 v-DSW 中不优化 κ，因此这里不需要用到它。

A.3 训练算法

使用 Max-SW 训练点云自编码器：我们在算法 2 中给出了使用 Max-SW 训练自编码器的算法。该算法包含一个嵌套循环：一个用于训练自编码器，另一个用于寻找 Max-SW 的最大投影方向。

使用摊销投影优化训练点云自编码器：我们在算法 3 中给出了使用摊销投影优化训练点云自编码器的算法。采用摊销优化后，寻找最大投影方向的内循环被移除。

使用 v-DSW 训练点云自编码器：我们在算法 4 中给出了使用 v-DSW 训练自编码器的算法。该算法包含一个嵌套循环：一个用于训练自编码器，另一个用于寻找 v-DSW 的最佳投影方向分布。

使用摊销分布投影优化训练点云自编码器：我们在算法 5 中给出了采用摊销分布投影优化训练点云自编码器的算法。借助摊销分布优化，无需再通过内循环寻找最佳投影方向分布。

Alg法 2训练 g p点云自编码器与最大切片瓦瑟斯坦距离

输入： 点云分布 $p(X)$、学习率 η、切片学习率 η_s、模型最大迭代次数 $\mathcal{T}$、切片最大迭代次数 T、小批量大小 k。

初始化：初始化编码器 f_ϕ 和解码器 g_γ当 ϕ, γ 未收敛或未达到 $\mathcal{T}$时执行 从 $p(X)\nabla_\phi = 0$中独立同分布地采样一个小批量 $X_1, \ldots, X_k$ $\nabla_\gamma = 0$对于 $i = 1$ 到 k 执行 初始化 θ当 θ 未收敛或未达到 T时执行$\theta = \theta + \eta_s \cdot \nabla_\theta \mathrm{W}_p(\theta\sharp P_{X_i}$, $\theta\sharp P_{g_\gamma(f_\phi(X_i))})$ # 可以使用其他更新规则$\theta = \frac{\theta}{||\theta||_2}$ # 投影回单位超球面 $\mathbb{S}^{d-1}$结束循环$\nabla_\phi = \nabla_\phi + \frac{1}{k}\nabla_\phi \mathrm{W}_p(\theta\sharp P_{X_i}, \theta\sharp P_{g_\gamma(f_\phi(X_i))})\nabla_\gamma = \nabla_\gamma + \frac{1}{k}\nabla_\gamma \mathrm{W}_p(\theta\sharp P_{X_i}$, $\theta\sharp P_{g_\gamma(f_\phi(X_i))})$ 结束循环$\phi = \phi - \eta \cdot \nabla_\phi$ # 可以使用其他更新规则 $\gamma = \gamma - \eta \cdot \nabla_\gamma$ # 可以使用其他更新规则 结束循环 返回： ϕ, γ

Algorithm 3 Traini$^{\text{ng po}}$ int-cloud aut$_{\text{oenco}}$ der with amortized projecti$^{\text{on op}}$ timizati

on 输入：点云分布 $p(X)$、学习率 η、切片学习率 η_s、模型最大迭代次数 $\mathcal{T}$、小批量大小 k。初始化：初始化编码器 f_ϕ、解码器 g_γ 和摊销模型 $a_\psi$while ϕ, γ, ψ 未收敛或达到 $\mathcal{T}$do 从 $p(X)\nabla_\phi = 0$中独立同分布采样一个小批量 $X_1, \ldots, X_k$ $\nabla_\gamma = 0, \nabla_\psi = 0$for $i = 1$到 k do$\theta_{\psi,\gamma,\phi} = a_\psi(X_i, g_\gamma(f_\phi(X_i)))$ $\nabla_\psi = \nabla_\psi + \frac{1}{k}\nabla_\psi \mathrm{W}_p(\theta_{\psi,\gamma,\phi}\sharp P_{X_i}, \theta_{\psi,\gamma,\phi}\sharp P_{g_\gamma(f_\phi(X_i))})\nabla_\phi = \nabla_\phi + \frac{1}{k}\nabla_\phi \mathrm{W}_p(\theta_{\psi,\gamma,\phi}\sharp P_{X_i}, \theta_{\psi,\gamma,\phi}\sharp P_{g_\gamma(f_\phi(X_i))})\nabla_\gamma = \nabla_\gamma + \frac{1}{k}\nabla_\gamma \mathrm{W}_p(\theta_{\psi,\gamma,\phi}\sharp P_{X_i}, \theta_{\psi,\gamma,\phi}\sharp P_{g_\gamma(f_\phi(X_i))})$ end for$\psi = \psi + \eta_s \cdot \nabla_\psi$ # 也可以使用其他更新规则$\phi = \phi - \eta \cdot \nabla_\phi$ # 也可以使用其他更新规则 $\gamma = \gamma - \eta \cdot \nabla_\gamma$ # 也可以使用其他更新规则 end while 返回： ϕ, γ

B. 证明

B.1. 命题1的证明

我们首先回顾两个概率测度 μ 和 ν之间投影一维 Wasserstein 距离的定义：$\hat{}\,\hat{}\,\hat{}\,\hat{}\,\mathrm{PW}_p(\mu, \nu; \theta) = \mathrm{W}_p(\theta\sharp\mu, \theta\sharp\nu)$ for $\theta \neq \mathrm{argmax}_{\theta\in\mathbb{S}^{d-1}}\mathrm{W}_p(\theta\sharp\mu, \theta\sharp\nu)$。

非负性和对称性： 由于 Wasserstein 距离具有非负性和对称性，投影瓦瑟斯坦距离的非负性和对称性可直接由其定义得到。

算法 4使用冯·米塞斯-费舍尔分布式切片 Wasserstein 距离训练点云自编码器

输入： 点云分布 $p(X)$、学习率 η、切片学习率 η_s、模型最大迭代次数 $\mathcal{T}$、切片最大迭代次数 T、小批量大小 k、投影数量 L 以及浓度超参数 κ。

初始化： 初始化编码器 f_ϕ 和解码器 g_γ

当 ϕ, γ 未收敛或达到 $\mathcal{T}$**时**

从 $p(X)$ 中采样一个小批量 $\{X_1, \ldots, X_k\}$，样本独立同分布

$\nabla_\phi = 0, \nabla_\gamma = 0$

对于 $i = 1$至 k **执行**

初始化 ϵ

当 ϵ 未收敛或未达到 T**时**

通过算法1中的重参数化接受-拒绝采样，从冯·米塞斯–费舍尔分布(ϵ, κ)中独立同分布地采样 θ_1^ϵ、……、θ_L^ϵ

$\epsilon = \epsilon + \eta_s \cdot \frac{1}{L} \sum_{l=1}^{L} \nabla_\epsilon \mathrm{W}_p(\theta_l^\epsilon \sharp P_{X_i}, \theta_l^\epsilon \sharp P_{g_\gamma(f_\phi(X_i))})$ # 也可以使用其他更新规则

$\epsilon = \frac{\epsilon}{||\epsilon||_2}$ #投影回单位超球面 $\mathbb{S}^{d-1}$

结束 while 循环

通过算法1中的重参数化接受-拒绝采样，从 vMF(ϵ, κ) 中独立同分布地抽取 $\theta_1^\epsilon, \ldots, \theta_L^\epsilon$。$\mathrm{W}_p(\theta_l^\epsilon \sharp P_{X_i}, \theta_l^\epsilon \sharp P_{g_\gamma(f_\phi(X_i))})\nabla_\gamma = \nabla_\gamma + \frac{1}{k} \frac{1}{L} \sum_{i=l}^{L} \nabla_\gamma \mathrm{W}_p(\theta_l^\epsilon \sharp P_{X_i}, \theta_l^\epsilon \sharp P_{g_\gamma(f_\phi(X_i))})$

结束 for 循环

$\phi = \phi - \eta \cdot \nabla_\phi$ # 可以使用其他更新规则

$\gamma = \gamma - \eta \cdot \nabla_\gamma$ # 可以使用其他更新规则

结束 while

返回： ϕ， γ

算法 5使用摊销投影优化训练点云自编码器

输入：点云分布 $p(X)$，学习率 η，切片学习率 η_s，模型最大迭代次数 $\mathcal{T}$，小批量大小 k。初始化：初始化编码器 f_ϕ、解码器 g_γ 和摊销模型 a_ψ当 ϕ, γ, ψ 未收敛或未达到 $\mathcal{T}$时执行 从 $p(X)\nabla_\phi = 0$ 中抽取一个小批量 $\{X_1, \ldots, X_k\}$，各样本独立同分布 $\nabla_\gamma = 0, \nabla_\psi = 0$对 $i = 1$ 到 k 执行$\epsilon_{\psi,\gamma,\phi} = a_\psi(X_i, g_\gamma(f_\phi(X_i)))$ 通过算法 $\nabla_\psi = \nabla_\psi + \frac{1}{k} \frac{1}{L} \sum_{i=l}^{L} \nabla_\psi$ 中的重参数化接受-拒绝抽样，从冯·米塞斯–费舍尔分布 vMF($\epsilon_{\psi,\gamma,\phi}, \kappa$) 中抽取 θ^ψ，$_{\gamma\,\phi}, , \ldots, \theta^\psi$，$_{\gamma\,\phi}$，各样本独立同分布 $\mathrm{W}_p(\theta^\psi$，$_{\gamma\,\phi} \sharp P_{X_i}, , \theta^\psi$，$_{\gamma\,\phi} \sharp P_{g_\gamma}, (f_\phi(X_i)))$ $\mathrm{W}_p(\theta^\psi$，$_{\gamma\,\phi} \sharp P_{X_i}, , \theta^\psi$，$_{\gamma\,\phi} \sharp P_{g_\gamma}, (f_\phi(X_i)))$ $\mathrm{W}_p(\theta^\psi$，$_{\gamma\,\phi} \sharp P_{X_i}, , \theta^\psi$，$_{\gamma\,\phi} \sharp P_{g_\gamma}, (f_\phi(X_i)))$ 结束 for$\psi = \psi + \eta_s \cdot \nabla_\psi$ # 可以使用其他更新规则 $\phi = \phi - \eta \cdot \nabla_\phi$ # 可以使用其他更新规则$\gamma = \gamma - \eta \cdot \nabla_\gamma$ # 可以使用其他更新规则 结束 while 返回： ϕ, γ

三角不等式： 对于任意三个概率测度 $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\mathbb{R}^d)$，有：

$$\begin{aligned} \mathrm{PW}_p(\mu_1, \mu_3; \hat{\theta}) &= \mathrm{W}_p(\hat{\theta} \sharp \mu_1, \hat{\theta} \sharp \mu_3) \\ &\leq \mathrm{W}_p(\hat{\theta} \sharp \mu_1, \hat{\theta} \sharp \mu_2) + \mathrm{W}_p(\hat{\theta} \sharp \mu_2, \hat{\theta} \sharp \mu_3) \\ &= \mathrm{PW}_p(\mu_1, \mu_2; \hat{\theta}) + \mathrm{PW}_p(\mu_2, \mu_3; \hat{\theta}), \end{aligned}$$

其中，第一个不等式由 Wasserstein 距离的三角不等式得出。

^ ^**恒等性：** 若 $\mu=\nu$，则根据 Wasserstein 距离的恒等性，$\mathrm{PW}_p(\mu,\ \nu;\theta)=0$。然而，若 $\mathrm{PW}_p(\mu,\ \nu;\ \theta)$ =0，则存在 $\theta'\in\mathbb{S}^{d-1}$，使得 $0=\mathrm{PW}_p(\mu,\ \nu;\ \hat{}\theta)<\mathrm{PW}_p(\mu,\ \nu;\theta')$。令 $\mathcal{F}[\gamma](w)=\int_{\mathbb{R}^{d'}}e^{-i\langle w,x\rangle}d\gamma(x)$ 为 $\gamma\in\mathcal{P}(\mathbb{R}^{d'})$ 的傅里叶变换，对于任意 $t\in\mathbb{R}$，有

$$\begin{aligned}\mathcal{F}[\mu](t\theta')&=\int_{\mathbb{R}^d}e^{-it\langle\theta',x\rangle}d\mu(x)=\int_{\mathbb{R}}e^{-itz}d\theta'\sharp\mu(z)=\mathcal{F}[\theta'\sharp\mu](t)\\&\neq\mathcal{F}[\theta'\sharp\nu](t)=\int_{\mathbb{R}}e^{-itz}d\theta'\sharp\nu(z)=\int_{\mathbb{R}^d}e^{-it\langle\theta',x\rangle}d\nu(x)=\mathcal{F}[\nu](t\theta').\end{aligned}$$

因此，我们有 $\mu\neq\nu$。证明完毕。

B.2. 定理1的证明

我们首先证明非最优的 冯·米塞斯-费希尔分布切片 Wasserstein 距离（v-DSW）的度量性。对于任意两个概率测度 $\mu,\ \nu\in\mathcal{P}_p(\mathbb{R}^d)$，非最优的 冯·米塞斯-费希尔分布切片 Wasserstein 距离（v-DSW）定义如下:

$$\text{v-DSW}_p(\mu,\nu;\epsilon,\kappa)=\left(\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right)^{\frac{1}{p}},$$

其中 $\epsilon\in\mathbb{S}^{d-1}$ 和 $0<\kappa<\infty$。

引理1. 对于任意 $\epsilon\in\mathbb{S}^{d-1}$ 和 $\kappa<\infty$，*v-DSW*$_p(\cdot,\cdot;\epsilon,\kappa)$ 是概率测度空间上的有效度量 .

证明。现在我们证明 v-DSW 满足非负性、对称性、三角不等式和恒等性。

非负性和对称性： v-DSW 的非负性和对称性直接源于 Wasserstein 距离的非负性和对称性。

三**角不等式：** 对于任意三个概率测度 μ_1、 μ_2和 $\mu_3\in\mathcal{P}_p(\mathbb{R}^d)$，我们有

$$\begin{aligned}\text{v-DSW}_p(\mu_1,\mu_3;\epsilon,\kappa)&=\left(\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu_1,\theta\sharp\mu_3)\right)^{\frac{1}{p}}\\&\leq\left(\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\left[\mathrm{W}_p(\theta\sharp\mu_1,\theta\sharp\mu_2)+\mathrm{W}_p(\theta\sharp\mu_2,\theta\sharp\mu_3)\right]^p\right)^{\frac{1}{p}}\\&\leq\left(\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu_1,\theta\sharp\mu_2)\right)^{\frac{1}{p}}+\left(\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu_2,\theta\sharp\mu_3)\right)^{\frac{1}{p}}\\&=\text{v-DSW}_p(\mu_1,\mu_2;\epsilon,\kappa)+\text{v-DSW}_p(\mu_2,\mu_3;\epsilon,\kappa)\end{aligned}$$

同一性： 根据定义，如果 $\mu=\nu$，我们得到 $\text{v-DSW}_p(\mu,\ \nu;\epsilon,\ \kappa)=0$。现在，我们需要证明，如果 $\text{v-DSW}_p(\mu,\ \nu;\epsilon,\ \kappa)=0$，则 $\mu=\nu$。

如果 $\text{v-DSW}_p(\mu,\nu;\ \epsilon,\kappa)=0$，我们有 $\left(\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right)^{\frac{1}{p}}=0$，这意味着 $\mathbb{E}_{\theta\sim\text{vMF}(\epsilon,\kappa)}\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)=0$。因此，对于 $\text{vMF}(\epsilon,\kappa)$，几乎必然有 $\mathrm{W}_p(\theta\sharp\mu,\theta\sharp\nu)=0\ \theta\in\mathbb{S}^{d-1}$。利用 Wasserstein 距离的同一性，我们得到：对于 $\text{vMF}(\epsilon,\kappa)$，几乎必然有 $\theta\sharp\mu=\theta\sharp\nu\ \theta\in\mathbb{S}^{d-1}$。由于 $\text{vMF}(\epsilon,\kappa)$ 在 $0<\kappa<\infty$ 时的支撑集覆盖所有 $\mathbb{S}^{d-1}$，对于任意 $t\in\mathbb{R}$ 和 $\theta\in\mathbb{S}^{d-1}$，我们有:

$$\begin{aligned}\mathcal{F}[\mu](t\theta)&=\int_{\mathbb{R}^d}e^{-it\langle\theta,x\rangle}d\mu(x)=\int_{\mathbb{R}}e^{-itz}d\theta\sharp\mu(z)=\mathcal{F}[\theta\sharp\mu](t)\\&=\mathcal{F}[\theta\sharp\nu](t)=\int_{\mathbb{R}}e^{-itz}d\theta\sharp\nu(z)=\int_{\mathbb{R}^d}e^{-it\langle\theta,x\rangle}d\nu(x)=\mathcal{F}[\nu](t\theta),\end{aligned}$$

其中，$\mathcal{F}[\gamma](w)=\int_{\mathbb{R}^{d'}}e^{-i\langle w,x\rangle}d\gamma(x)$ 表示 $\gamma\in\mathcal{P}(\mathbb{R}^{d'})$ 的傅里叶变换。随后，根据傅里叶变换的单射性，我们得到 $\mu=\nu$。证明完毕。 □

为简化记法，我们将 $\text{v-DSW}(X,Y;\ \epsilon,\kappa)=\text{v-DSW}(P_X,P_Y;\ \epsilon,\kappa)$ 记作 $X,Y\in\mathcal{X}$ 是两个点云，$P_X=\frac{1}{m}\sum_{i=1}^m\delta_{x_i}$，且 $P_Y=\frac{1}{m}\sum_{i=1}^m\delta_{y_i}$。我们将 v-DSW 从概率测度空间上的度量推广到点云空间 $\mathcal{X}$。

推论1. 对于任意 $\epsilon \in \mathbb{S}^{d-1}$ 和 $\kappa < \infty$，*v-DSW*$_p(\cdot, \ ; \ \epsilon, \kappa)$ 都是点云空间 $\mathcal{X}$上的有效度量。

证明。 由于 $P_X, P_Y \in \mathcal{P}_p(\mathbb{R}^d)$，非负性、对称性、三角不等式和恒等性均可直接由引理1得出。现在只需证明 v-DSW 对置换保持不变。根据经验概率测度的定义，这一性质显而易见。对于任意置换函数 σ $P_X = \frac{1}{m}\sum_{i=1}^m \delta_{x_i} =$，有 $\frac{1}{m}\sum_{i=1}^m \delta_{x_{\sigma\ i}} = P_{\sigma(X)}$ 由此完成证明。 □

现在继续证明定理1。若 $\mathbb{E}_{X\sim p(X)}\left(\mathbb{E}_{\theta\sim \text{vMF}(\epsilon,\kappa)} \text{W}_p^p(\theta\sharp P_X, \theta\sharp P_{g_\gamma(f_\phi(X))})\right)^{\frac{1}{p}} = 0$，则得到$\left(\mathbb{E}_{\theta\sim \text{vMF}(\epsilon,\kappa)} \text{W}_p^p(\theta\sharp P_X, \theta\sharp P_{g_\gamma(f_\phi(X))})\right)^{\frac{1}{p}}$ =v-DSW$(X, g_\gamma(f_\phi(X)); \epsilon, \kappa) = 0$对几乎处处的 p 成立 $X \in \mathcal{X}$。根据推论1，可得 $X = g_\gamma(f_\phi(X))$ 对几乎处处的 p 成立 $X \in \mathcal{X}$。证明完成。

B.3. 命题 2 的证明

我们首先回顾定义 3 中自注意力摊销模型的定义：

$$a_\psi(X, Y) = \frac{\mathcal{A}_\zeta(X'^\top)^\top \mathbf{1}_m + \mathcal{A}_\zeta(Y'^\top)^\top \mathbf{1}_m}{||\mathcal{A}_\zeta(X'^\top)^\top \mathbf{1}_m + \mathcal{A}_\zeta(Y'^\top)^\top \mathbf{1}_m||_2},$$

对称性： 由于自注意力摊销模型对 X 和 Y 使用相同的注意力权重 ζ，交换 X 和 Y 会得到相同的结果 $a_\psi(X, Y) = a_\psi(Y, X)$。

排列不变性： 基于 Yang 等人（2019b，附录A）的结果，我们证明自注意力摊销模型具有置换不变性。具体而言，我们有：

$$\begin{aligned}
\mathcal{A}_\zeta(X'^\top)^\top \mathbf{1}_m &= \text{Att}(X'^\top W_q, X'^\top W_k, X'^\top W_v)^\top \mathbf{1}_m \\
&= \left(\text{softmax}_{\text{row}}\left[\frac{X'^\top W_q W_k^\top X'}{\sqrt{d_k}}\right] X'^\top W_v\right)^\top \mathbf{1}_m \\
&= \left(\text{softmax}_{\text{row}}\left[\frac{\sigma(X)'^\top W_q W_k^\top \sigma(X)'}{\sqrt{d_k}}\right] \sigma(X)'^\top W_v\right)^\top \mathbf{1}_m \\
&= \mathcal{A}_\zeta(\sigma(X)'^\top)^\top \mathbf{1}_m.
\end{aligned}$$

同样，该证明对线性自注意力和高效自注意力均成立。

C. 实验设置

本节首先介绍点云重建、迁移学习和点云生成的训练过程及架构详情。然后，我们介绍实验中不同距离的实现细节和超参数设置。

C.1. 点云重建与下游应用的详细信息

点云重建： 我们采用 ASW（Nguyen 等，2021c）中的相同设置来训练自编码器。我们使用了 Pham 等人（2020）提出的 Point-Net（Qi 等人，2017）变体，其嵌入维度为 256。自编码器和分类器的架构如图 5 所示。我们的自编码器在 ShapeNet Core-55 数据集（Chang 等人，2015年）上训练，批次大小为 128，点云大小为 2048。我们使用 SGD 优化器训练 300 个周期，初始学习率为 1e-3，动量为 0.9，权重衰减为 5e-4。

接下来，我们将详细介绍如何开展点云重建的两项下游应用。

迁移学习： 在自编码器的潜在空间上训练分类器。具体而言，我们通过预训练编码器提取输入三维点云的 256 维潜向量（低于 (Lee et al., 2022) 中的设置）。然后，将该向量输入一个多层感知机，其隐藏层大小分别为 512 和 256。最后一层输出一个 40 维向量，表示对 ModelNet40 数据集中 40 个类别的预测。

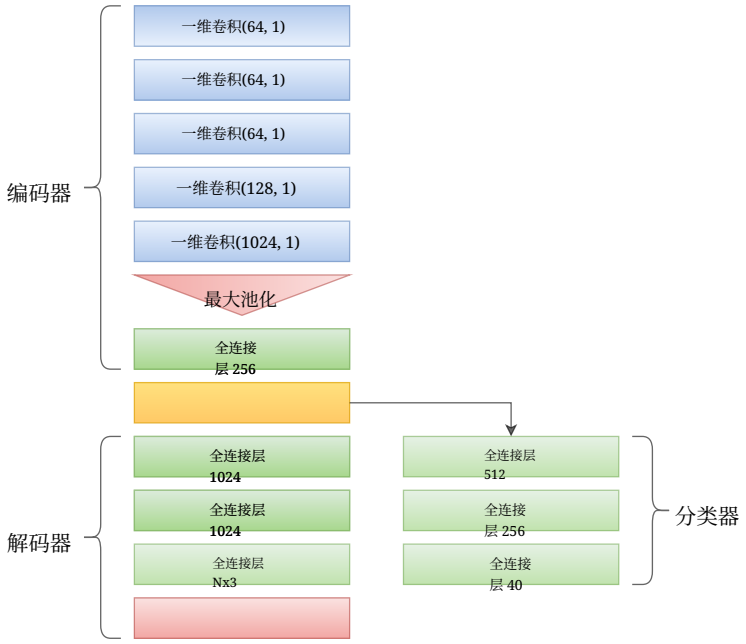

图 5. 的架构 我们实验中的 Point-Net 变体。对于迁移学习，我们使用一个简单的分类器，包含 3 个全连接 - 连接层。所有层均 默认后接 ReLU 激活和批归一化，但最后几层除外。

点云生成： 我们的生成模型按如下方式在自编码器的潜在空间中进行训练。首先，我们通过预训练编码器提取输入三维点云的 256 维潜向量。然后，从正态分布 $\mathcal{N}(0,\ \mathbb{I}_{64})$ 中采样一个 64 维向量，其中 $\mathbb{I}_{64}$ 为 $64x64$ 单位矩阵，并将其输入生成器；生成器同样输出一个 256 维向量。最后，生成器通过最小化生成潜在编码与真实潜在编码之间的最优传输距离进行学习。

C.2. 基线距离的详细信息

我们想强调的是，对于 Chamfer 距离、EMD、切片瓦瑟斯坦距离和 Max-SW，我们使用了 (Nguyen 等，2021c) 中报告的同一组超参数。

Chamfer 距离和 EMD： 我们使用 (Yang 等，2019a) 中的 CUDA 实现。

切片瓦瑟斯坦距离：我们使用包含 100 个切片的蒙特卡洛估计。

Max-SW： 我们使用投影次梯度上升算法优化投影。使用 Adam 优化器进行训练，初始学习率为 1e-4。迭代次数 T 从 $\{1, 10, 50\}$ 中选取。

自适应 SW：我们采用 Nguyen 等（2021c）中的算法1，并使用以下相同的参数集： $N_0 = 2$， s =1， $\epsilon = 0.5$，以及 $M = 500$。

v-DSW：我们使用随机投影梯度上升算法来优化公式 19 中的位置向量 ϵ，同时将 v-DSW 及其所有摊销版本的浓度参数 κ 固定为 1。与 Max-SW 类似，它使用 Adam 优化器进行训练，初始学习率为 1e-4。迭代次数 T 根据任务性能从 $\{1, 10, 50\}$ 中选取。直观而言，增加迭代次数会带来更好的近似，使其更接近最优值，但计算成本也会相应提高。与 SW 一样，我们还使用包含 100 个切片的蒙特卡洛估计。

C.3. 摊销切片 Wasserstein 距离的细节

线性、广义线性和非线性模型：我们采用 (Nguyen 与 Ho，2022a) 中的官方实现。

基于自注意力的模型：我们在实验中改编了相应论文中的官方实现。对于所有变体， d_v 设为 3，与点云的维度相同，而 d_k 则从 $\{16, 32, 64, 128\}$中选取。在公式 15 中，投影维度 k 从 $\{64, 128\}$中选取。

训练摊销模型 **els:** 学习率设为 1e-3，优化器设为 Adam（Kingma & Ba，2014年） h

表 *4.* 不同自编码器在 ModelNet40 数据集上的重建和迁移学习性能。对于 v-DSW 和 Max-SW，T 表示投影次梯度上升迭代的次数。在表 1 中，v-DSW 和 Max-SW 的 T 均为 = 50 次迭代。除 EMD 外，所有重建损失均乘以 100。

方法	倒角距离 (10^{-2}, ↓)	切片瓦瑟斯坦距离 (10^{-2}, ↓)	EMD (↓)	准确率 (↑)	时间 (↓)
CD	1.25 ± 0.03	681.20 ± 16.73	653.52 ± 10.43	86.28 ± 0.34	95
EMD	0.40 ± 0.00	94.54 ± 2.90	168.60 ± 1.57	88.45 ± 0.20	208
SW	0.68 ± 0.01	89.61 ± 3.88	191.12 ± 2.88	87.90 ± 0.27	106
Max-SW (T = 1)	0.69 ± 0.01	87.60 ± 0.95	190.88 ± 0.40	88.05 ± 0.23	97
Max-SW (T = 10)	0.69 ± 0.01	90.72 ± 0.58	192.82 ± 0.73	87.82 ± 0.37	102
Max-SW (T = 50)	0.68 ± 0.01	88.22 ± 1.45	190.23 ± 0.1	87.97 ± 0.14	116
ASW	0.69 ± 0.01	89.42 ± 5.07	192.03 ± 3.09	87.78 ± 0.20	103
v-DSW（T = 1）	**0.67 ± 0.01**	87.29 ± 1.49	188.52 ± 1.47	87.87 ± 0.28	115
v-DSW（T = 10）	0.68 ± 0.00	87.44 ± 1.07	189.97 ± 1.04	87.98 ± 0.23	205
v-DSW（T = 50）	**0.67 ± 0.00**	85.03 ± 3.31	187.75 ± 2.00	87.83 ± 0.40	633
$\mathcal{L}$-Max-SW	1.06 ± 0.03	121.85 ± 5.77	236.87 ± 3.42	87.70 ± 0.23	**94**
$\mathcal{G}$-Max-SW	12.11 ± 0.29	851.07 ± 2.11	829.28 ± 5.53	87.49 ± 0.36	97
$\mathcal{N}$-Max-SW	7.38 ± 3.29	618.74 ± 153.87	648.32 ± 117.03	87.43 ± 0.15	96
$\mathcal{L}$v-DSW（我们的方法）	0.68 ± 0.00	85.32 ± 0.54	188.32 ± 0.23	87.70 ± 0.34	114
$\mathcal{G}$v-DSW（我们的方法）	0.68 ± 0.01	82.77 ± 0.48	187.04 ± 1.11	87.75 ± 0.19	117
$\mathcal{N}$v-DSW（我们的方法）	**0.67 ± 0.00**	83.47 ± 0.49	186.66 ± 0.81	87.84 ± 0.07	115
$\mathcal{A}$v-DSW（我们的方法）	0.67 ± 0.01	83.08 ± 1.22	186.27 ± 0.56	88.05 ± 0.17	230
$\mathcal{EA}$v-DSW（我们的方法）	0.68 ± 0.01	82.05 ± 0.40	186.46 ± 0.25	88.07 ± 0.21	125
$\mathcal{LA}$v-DSW（我们的方法）	0.68 ± 0.00	**81.03 ± 0.18**	**185.26 ± 0.31**	**88.28 ± 0.13**	123

表*5*。 ShapeNet Core-55 数据集中点云重建的定量结果（以 EMD 衡量）。

方法	PC1	PC2	PC3	PC4	PC5	PC6	Avg
SW	141.07	139.50	118.83	99.11	150.28	128.46	129.54
Max-SW（T = 50）	145.15	131.76	112.13	116.73	139.91	115.79	126.91
ASW	139.17	126.55	115.49	**91.07**	153.87	114.84	123.50
v-DSW（T = 50）	133.06	146.99	105.65	105.66	137.32	**110.50**	123.20
$\mathcal{N}$v-DSW	132.60	127.57	100.81	94.31	131.04	116.34	117.11
$\mathcal{EA}$v-DSW	139.64	**124.28**	100.34	98.33	**123.59**	115.05	116.87
$\mathcal{LA}$v-DSW	**130.21**	127.00	**96.75**	98.09	132.33	114.11	**116.41**

表*6*。 改变 L 时切片瓦瑟斯坦距离和 $\mathcal{LA}$v-DSW 的重建结果。倒角距离和切片Wasserstein距离均乘以 100。

方法	L	倒角距离（10^{-2}, ↓）	切片瓦瑟斯坦距离（10^{-2}, ↓）	EMD（↓）	Time
SW	50	**0.67 ± 0.00**	90.17 ± 2.97	190.97 ± 1.87	100
	100	0.68 ± 0.01	**89.61 ± 3.88**	191.12 ± 2.88	107
	200	**0.67 ± 0.00**	89.54 ± 4.57	191.21 ± 3.87	111
	500	0.67 ± 0.01	88.20 ± 4.22	190.14 ± 2.35	142
$\mathcal{LA}$v-DSW	50	0.68 ± 0.01	85.88 ± 4.03	188.80 ± 2.55	133
	100	0.68 ± 0.00	**81.03 ± 0.18**	**185.26 ± 0.31**	123

$(\beta_1,\ \beta_2) = (0,\ 0。9)$。

D. 其他实验结果

点云重建： 表 4 展示了点云重建实验的完整定量结果。对于 Max-SW 和 v-DSW，我们改变梯度迭代次数 T，使其取值为 $\{1, 10, 50\}$。由于倒角距离并非真正的距离，因此我们根据切片瓦瑟斯坦距离和 EMD 损失选择最佳迭代次数（优先考虑 EMD 损失，其次是切片瓦瑟斯坦距离）。正如

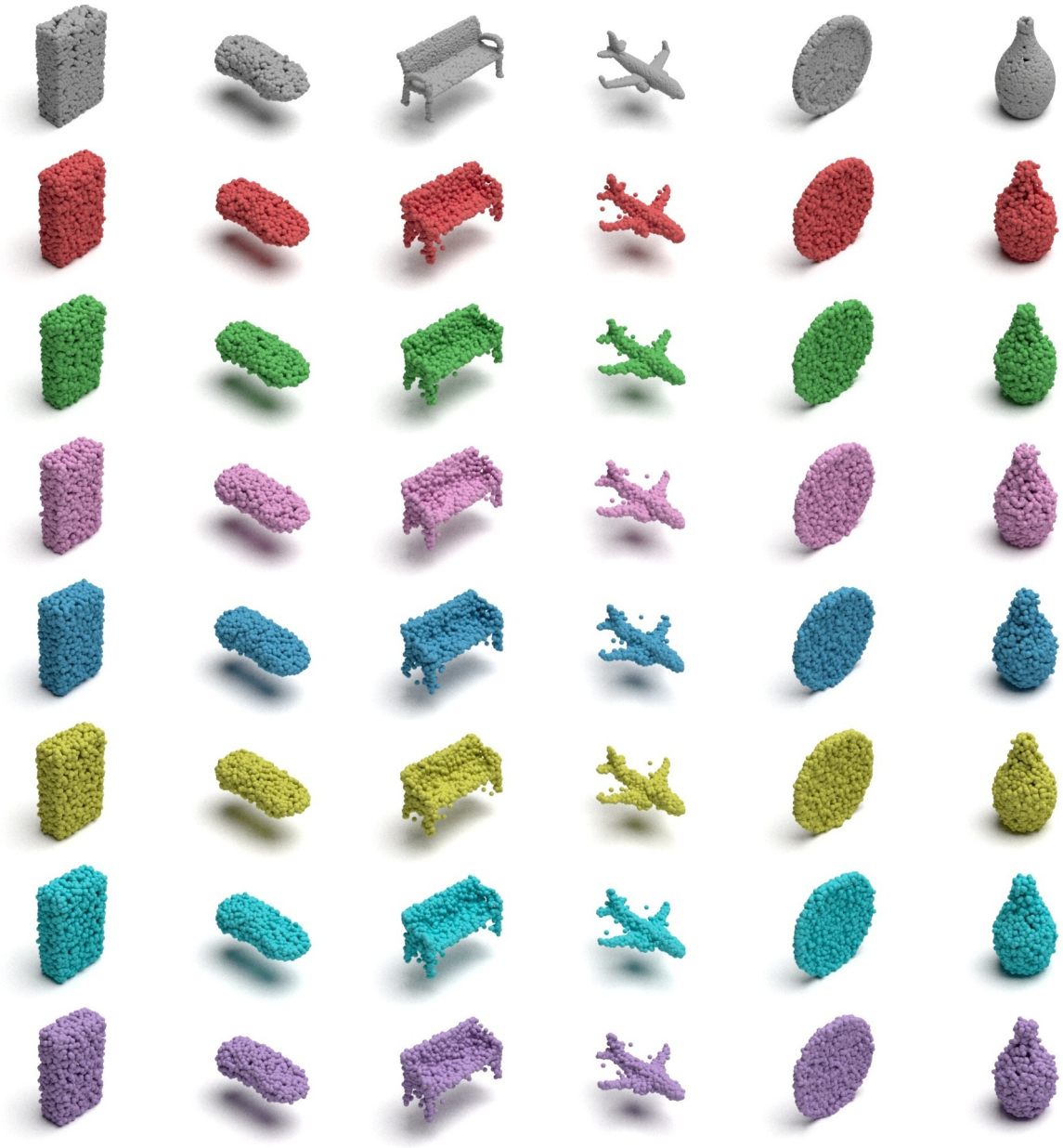

图6. ShapeNet Core-55 数据集上点云重建的定性结果。从上到下依次为：输入、切片瓦瑟斯坦距离、Max-SW (T = 50)、ASW、v-DSW (T = 50)、$\mathcal{N}$v-DSW、$\mathcal{EA}$v-DSW 和 $\mathcal{LA}$v-DSW。

表7. 改变投影次梯度上升迭代次数 (T) 时 v-DSW 的重建结果。倒角距离和切片Wasserstein距离均乘以 100。

方法	倒角距离 (10^{-2}, ↓)	切片瓦瑟斯坦距离 (10^{-2}, ↓)	EMD (↓)
v-DSW (T = 0)	0.67 ± 0.01	88.63 ± 2.30	189.81 ± 1.19
v-DSW (T = 1)	0.67 ± 0.01	87.29 ± 1.49	188.52 ± 1.47
v-DSW (T = 10)	0.68 ± 0.00	87.44 ± 1.07	189.97 ± 1.04
v-DSW (T = 50)	**0.67 ± 0.00**	85.03 ± 3.31	187.75 ± 2.00
$\mathcal{LA}$v-DSW	0.68 ± 0.00	**81.03 ± 0.18**	**185.26 ± 0.31**

表*8.* 改变 κ时 $\mathcal{LA}$v-DSW的重建结果。CD和SWD均乘以100。

κ	倒角距离 ($10^{-2}, \downarrow$)	切片瓦瑟斯坦距离 ($10^{-2}, \downarrow$)	EMD ($\downarrow$)
0.1	**0.67 ± 0.00**	81.88 ± 1.09	185.30 ± 0.94
1	0.68 ± 0.00	**81.03 ± 0.18**	**185.26 ± 0.31**
10	0.85 ± 0.01	96.01 ± 4.24	208.46 ± 4.03

表*9.* ShapeNet 椅子类别点云生成的性能比较。对于 v-DSW 和 Max-SW，T 表示投影次梯度上升迭代的次数。在表3中，v-DSW 和 Max-SW 分别进行 T = 50 次和 10 次迭代。JSD、MMD-CD 和 MMD-EMD 均乘以 100。

方法	JSD (10^{-2}, $\downarrow$)	MMD (10^{-2}, $\downarrow$)		覆盖率 (% $\uparrow$)		1-NNA (% $\downarrow$)	
		CD	EMD	CD	EMD	CD	EMD
CD	17.88 ± 1.14	1.12 ± 0.02	17.19 ± 0.36	23.73 ± 1.69	10.83 ± 0.89	98.45 ± 0.10	100.00 ± 0.00
EMD	5.15 ± 1.52	0.61 ± 0.09	10.37 ± 0.61	41.65 ± 2.19	42.54 ± 2.42	87.76 ± 1.46	87.30 ± 1.22
SW	1.56 ± 0.06	0.72 ± 0.02	10.80 ± 0.11	38.55 ± 0.43	45.35 ± 0.48	89.91 ± 1.17	88.28 ± 0.70
Max-SW（T = 1）	1.74 ± 0.22	0.78 ± 0.05	11.05 ± 0.31	39.39 ± 2.28	46.82 ± 0.79	92.15 ± 0.95	90.20 ± 0.87
Max-SW（T = 10）	1.63 ± 0.32	0.74 ± 0.01	10.84 ± 0.08	40.47 ± 1.04	47.81 ± 0.78	91.46 ± 0.72	89.93 ± 0.86
Max-SW (T = 50)	1.57 ± 0.26	0.80 ± 0.05	11.25 ± 0.34	37.81 ± 1.69	46.23 ± 0.64	92.15 ± 0.72	90.35 ± 0.28
ASW	1.75 ± 0.38	0.78 ± 0.05	11.27 ± 0.38	38.16 ± 2.15	45.45 ± 1.40	91.21 ± 0.40	89.36 ± 0.40
v-DSW（T = 1）	1.84 ± 0.17	0.75 ± 0.03	11.02 ± 0.21	38.26 ± 1.46	45.35 ± 1.70	90.08 ± 0.48	87.81 ± 0.16
v-DSW（T = 10）	1.48 ± 0.17	0.77 ± 0.02	11.09 ± 0.09	37.22 ± 0.96	43.77 ± 0.39	90.40 ± 1.05	88.87 ± 1.04
v-DSW（T = 50）	1.79 ± 0.17	0.72 ± 0.02	10.73 ± 0.20	37.76 ± 0.71	45.49 ± 1.37	90.23 ± 0.13	88.33 ± 0.95
$\mathcal{L}$v-DSW（我们的方法）	1.67 ± 0.07	0.77 ± 0.04	11.10 ± 0.33	37.91 ± 1.84	45.64 ± 2.30	90.42 ± 0.53	88.82 ± 0.38
$\mathcal{G}$v-DSW（我们的方法）	1.56 ± 0.22	0.75 ± 0.02	10.99 ± 0.11	37.81 ± 1.70	45.69 ± 0.46	90.32 ± 0.38	88.26 ± 0.28
$\mathcal{N}$v-DSW（我们的方法）	**1.44 ± 0.06**	0.75 ± 0.02	10.95 ± 0.09	38.40 ± 1.34	46.28 ± 2.06	90.15 ± 0.80	88.65 ± 0.82
$\mathcal{EA}$v-DSW（我们的方法）	1.73 ± 0.21	**0.71 ± 0.04**	**10.70 ± 0.26**	40.03 ± 1.28	**48.01 ± 1.07**	89.98 ± 0.57	88.55 ± 0.38
$\mathcal{LA}$v-DSW（我们的方法）	1.54 ± 0.09	0.72 ± 0.03	10.74 ± 0.35	**40.62 ± 1.39**	45.84 ± 1.23	**89.44 ± 0.28**	**87.79 ± 0.37**

从表中可以看出，增加梯度上升迭代次数（T）能够提升 Max-SW 和 v-DSW 的重建性能，但会带来额外的计算量开销，尤其是对于 v-DSW。然而，在 T 的所有取值下，Max-SW 和 v-DSW 的重建性能（以切片瓦瑟斯坦距离和 EMD 衡量）通常都不如我们的摊销方法。此外，我们的摊销方法在 3 次运行中的标准差更小，因此比使用梯度上升方法的传统优化方法更加稳定。定性结果如图6所示。相应的 EMD 定量结果见表5。可以看出，我们的摊销 v-DSW 方法具有更优的性能。

关于投影数量（L）。在我们的实验中，L 与 ASW论文中的设置一样，固定为 100。这里，我们对投影数量 L 进行了消融研究，并在表6中报告结果。从表中可以看出，增加投影数量能够提升 EMD 方面的性能，但会增加运行时间。我们发现，$\mathcal{LA}$v-DSW 在 $L = 50$ 和 $L = 100$ 的设置下，比 $L = 500$ 的切片瓦瑟斯坦距离计算更快，同时在切片瓦瑟斯坦距离和 EMD 评估指标方面表现更好。与 $L = 200$的切片瓦瑟斯坦距离相比，$\mathcal{LA}$ v-DSW 在 $L = 50$ 下的计算时间大致相同，但切片瓦瑟斯坦距离和 EMD 评估指标更低。

关于位置向量的选择 ϵ。我们在此重申，v-DSW 的最优位置向量 $\epsilon^\star$ 是通过附录A.3中的算法4计算得到的。为展示其有效性，我们将其与随机位置 ϵ进行比较，即 T = 0。表7显示，对 vMF分布的位置参数进行优化有助于改善重建效果。此外，与随机初始化的位置 ϵ 以及使用传统方法优化得到的位置相比，我们的摊销优化仍能获得更好的重建分数。因此，采用摊销优化确实可能带来优势。

关于参数的选择 κ。首先需要指出的是，在我们的实验中，所有 v-DSW 和摊销 v-DSW 方法的 κ 均设为 1。实际应用中，可以通过网格搜索来选择参数 κ。在这里，我们通过改变 $\mathcal{LA}$v-DSW 的 $\kappa \in \{0.1, 1, 10\}$ 进行消融研究，并在表8中报告结果。从表中可以看出，$\kappa = 1$ 取得了最佳 EMD 结果。

点云生成。 我们在表9中汇总了点云生成的完整定量结果。对于 Max-SW 和 v-DSW，我们再次改变 $\{1, 10, 50\}$ 中的梯度迭代次数 T。需要注意的是，由于内存不足， $\mathcal{A}$v-DSW 无法用于此实验；而摊销 Max-SW 的性能太差，因此本实验未报告它们的结果。从表中可以看出，摊销 v-DSW 方法在 7 项度量中的 7 项上都取得了最佳性能。使用多于一次的梯度上升迭代（$T \in \{10, 50\}$）确实能提升 Max-SW 和 v-DSW 的生成性能，但代价是增加计算量。