无线数据集相似性：衡量监督与无监督机器学习中的距离

Jo˜ao Morais†、Sadjad Alikhani†、Akshay Malhotra‡、Shahab Hamidi-Rad‡、Ahmed Alkhateeb†{joao、alikhani、alkhateeb}@asu.edu， {akshay.malhotra、shahab.hamidi-rad}@interdigital.com

arXiv:2601.01023v1 [cs.LG] 3 Jan 2026

摘要—本文提出一种任务和模型感知框架，用于衡量无线数据集之间的相似性，可支持数据集选择/增强、仿真到真实场景（sim2real）比较、任务特定合成数据生成，并为模型训练/适配到新部署环境提供决策依据。我们根据候选数据集距离度量预测跨数据集迁移能力的效果来评估它们：如果两个数据集之间的距离较小，那么在其中一个数据集上训练的模型应能在另一个数据集上取得良好表现。我们将该框架应用于一项无监督任务——使用自编码器进行信道状态信息（CSI）压缩。利用基于UMAP嵌入并结合Wasserstein距离和欧几里得距离的指标，我们实现了超过0.85的皮尔逊相关系数，反映数据集距离与在一个数据集上训练、在另一个数据集上测试的任务表现之间的关系。我们还将该框架应用于下行链路中的有监督波束预测任务，并采用卷积神经网络。针对该任务，我们通过结合有监督UMAP和针对数据集不平衡的惩罚项，推导出一种标签感知距离。在这两项任务中，所得距离均优于传统基线，并始终与模型可迁移性呈现更强的相关性，支持对无线数据集进行与任务相关的比较。

I. 引言

无线通信中的机器学习（ML）应用正在快速增长 [1]–[5]。为了探索频谱效率的新前沿，工程师和研究人员开始借助计算机科学领域的最新进展，尤其是那些利用神经网络的技术。神经网络等学习机制受到广泛关注，这一点从无线通信会议上人工智能/机器学习论文的年度投稿量增长数倍可见一斑。然而，目前的研究工作过度强调学习方法，而对用于学习的数据关注不足。结果是，无线领域的大多数机器学习研究仍停留在学术环境中，尚未部署到真实世界场景。为弥补这一差距，首先需要解决一些有关模型设计、训练和部署过渡的重要问题：

- 如何为模型训练选择合适的数据集?
- 如何预测模型在实际部署中的性能?
- 如何衡量并确保模型在不同数据集之间的泛化能力?

为回答这些问题，本研究旨在概述数据集相似性及其在无线通信中的应用。数据集相似性/距离度量 [6]是指测量两个数据集之间的接近程度（相似）或差异程度（不同）。为此，我们旨在开发一种方法，在无需进行模型训练的情况下，估计机器学习模型从一个数据集迁移到另一个数据集时的泛化性能。其他本研究的潜在用途包括：i）检测现实世界运行过程中数据的分布偏移；ii）通过选择与目标环境具有更高相似度（从而具有更高可迁移性性能）的数据集，改进迁移学习；以及 iii）利用适当的合成数据增强真实数据。事实上，获取现实世界的无线数据颇具挑战性，因此，数据增强对于实际的机器学习开发与部署至关重要。

机器学习模型的训练数据可以通过几种不同方式获取。第一种是通过真实世界实验。POWDER [7]和 AERPAW [8]等测试平台为研究人员提供实验平台，使其能够接入并在实际部署环境中进行原型验证。第二种方式是使用真实世界数据集，例如来自 DeepSense6G [9] 或纽约大学 [10]的数据。这两种方法能提供宝贵的见解，但通常缺乏全面开发机器学习模型所需的规模和多样性。因此，仿真数据被广泛使用，其主要分为两种方法：随机方法和确定性方法。

随机模型（如 3GPP 定义的模型，包括 CDL、TDL、UMa 和 UMi）已得到广泛采用，并可通过 MATLAB 5G 工具箱等工具使用，以概率方式模拟信道条件。在确定性方法方面，有两种选择。第一种是射线追踪方法，Wireless InSite [11]以及较新的 SionnaRT[12], 等工具均支持此类方法，可进行高保真度、特定场景的仿真，这对 6G 研究尤为重要。第二种是 Deep-MIMO [13], 这类数据集，它们将射线追踪与参数化信道生成相结合，提供了一种混合方法，为确定性、特定场景的射线追踪带来额外的灵活性。然而，尽管已有这些工具，无线通信领域仍缺少健全的数据操作能力——即理解和比较数据集的能力——这阻碍了大规模生成模型的发展。

纯真实世界数据集或纯合成数据集，可能不足以支持全面的模型开发。为此，用仿真数据补充逼真的数据集至关重要。这可以通过射线追踪实现，但重建环境成本高昂，而且为每种设置都这样做并不现实。同样，将随机模型配置为匹配特定环境也很复杂。一种可能的解决方案是数据增强：与其针对每种场景创建全新的数据集，不如利用现有数据集；这些数据集可能与环境并非完全匹配，但在分布上相似。通过谨慎选择或增强数据集，我们可以减少从头生成新数据的需求。这凸显了评估

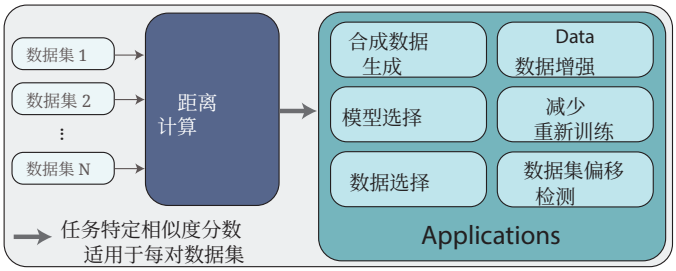

图 1. 数据集距离计算支持的应用示例。

数据集与数据分布之间相似度的重要性，以确保补充数据能够提升模型性能，而不会引入不一致性。

贡献：为此，本文旨在为无线领域研究人员提供在模型训练之前比较数据集的方法，并提供评估是否有必要进行模型再训练的方法，这些数据集可以是纯合成数据，也可以用于真实世界数据增强。图 1 展示了如何将我们的数据集距离计算框架应用于一组数据集，从而支持广泛的应用。该框架还可以仅根据数据集预测模型性能，进一步帮助避免模型再训练。总之，本文主要关注以下贡献：

- 我们开发了一种任务驱动的模型无关框架，用于评估数据集之间的相似度，无需训练额外模型。
- 我们设计了两种距离度量，首先采用类似 UMAP 的拓扑保持降维技术，将数据投影到低维空间中。随后计算低维空间中的距离：第一种距离是应用于通过 K 近邻（KNN）形成的簇上的欧几里得距离；第二种距离基于 Wasserstein距离，应用于各维度的分布之间。我们在有监督和无监督任务上对这两种方法进行了评估，结果显示其性能较以往方法有明显提升。
- 具体而言，针对监督任务，我们提出了一种新颖的有监督距离计算方法，能够有效利用标签信息，并引入惩罚项来提升数据点比较的准确性，从而提高度量准确性和鲁棒性。
- 我们证明了该框架可用于衡量通过度量计算得到的数据集距离与模型性能之间的相关性，为构建任务特定度量以指导数据集选择、模型再训练和基准测试提供了可能。

所提出的框架实现已开源，同时也公开了本研究中使用的所有评估脚本。此外，我们提供了详尽的文档和说明，以复现本文中的所有研究成果。[1]

组织结构：本文组织如下。第二节回顾了该领域的最新进展，探讨了机器学习及相关领域中的数据集相似性，并讨论了不同类型的距离及其应具备的性质。第四节定义了问题，详细说明了数据集、任务的构成，以及将距离度量与模型性能相关联的目标。在第七节中，我们将数据集距离应用于无监督场景，具体而言是CSI压缩，以评估这些距离与测试性能之间的相关性。第IX节将该应用扩展到有监督场景，包括视距状态和波束预测。最后，第十节总结了这项研究可能带来的影响及未来工作。

[1]文档、研究资料及复现说明可在以下网页中找到：*https://wi-lab.net/research/dataset-similarity*

II. 数据集相似性指标的研究现状

数据集相似性指标在领域自适应领域尤为重要 [14], [15]。领域自适应是机器学习的一个子领域，专注于将知识从一个领域（或数据集）迁移到另一个领域，尤其是在两个领域的分布不同的情况下 [16]–[20]。当目标领域的数据稀缺或难以获取 [21], , 但相似源领域的数据丰富时，这种方法得到广泛应用。例如，在无线通信中，人们可能会使用特定城市环境中采集的数据训练模型，但需要将同一模型应用于农村或郊区环境，而那里信号模式和干扰条件有所不同 [22]–[24]。另一种可能的情况是，无法采集真实世界数据，因此模型需要从使用仿真数据转而适应实际情况。领域自适应使机器学习模型能够通过学习这些领域之间的相似之处，同时适应它们的差异，从而更好地泛化 [25]–[27]。这一自适应过程的核心是使用数据集相似性指标，量化两个数据集之间的接近或差异程度。

距离类别：在许多数据驱动的应用中，选择合适的距离度量对于理解数据的结构及其内在关系至关重要。根据数据的性质和具体任务，距离可以分为若干类别。每个类别反映了对数据点组织方式的不同假设：数据点是位于线性子空间 [28],、遵循某种分布 [29],，还是存在于曲流形[30]–[32]上。下面将介绍四种主要的距离度量类别：几何距离、统计距离、子空间距离和基于流形的距离，随后简要讨论机器学习和数据分析中常用的其他类型的距离度量。

几何距离用于衡量通常为欧几里得空间[33]–[35]这类平坦空间中各点之间的直接空间关系。这类距离关注数据点之间的物理或空间差异，通常假设空间规则且呈线性。最常见的几何距离是欧几里得距离，它计算两点之间的直线距离。其他例子包括曼哈顿距离，它对每个维度上的绝对差求和；以及余弦相似度，它衡量向量之间的角距离。当数据点之间的关系不涉及曲率或概率解释时，几何距离直观易懂，常用于聚类、回归和分类等任务。

统计距离用于量化数据集 [29] 中概率分布或统计属性之间的差异。

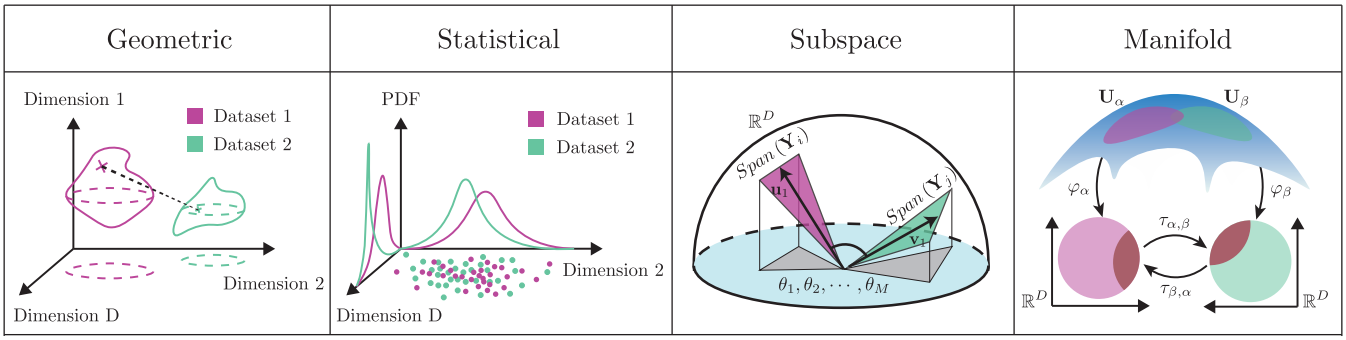

图 2. 可用于数据集之间的距离类别。

这些距离分为散度和积分概率度量（IPMs）。詹森-香农散度等散度*Kullback-Leibler (KL) divergence* [36], [37], 用于衡量一个分布与另一个分布的偏离程度，因此通常具有不对称性。积分概率度量（IPMs）则通过移动概率质量来计算将一个分布转换为另一个分布所需的最小“成本”，例如*Wasserstein*距离 [38] 。其他例子还包括詹森-香农散度，它是 KL 散度的对称形式；以及最大均值差异 *(MMD)* [39], ，用于在再生核希尔伯特空间中比较分布。

子空间距离用于衡量子空间之间的关系，而非单个数据点之间的关系 [40], [41]。在许多高维任务中，数据可以表示为子空间，例如在信号处理或计算机视觉中，关键信息在于数据的方向或张成空间，而非各个点本身。格拉斯曼距离[40], 例如，用于衡量格拉斯曼流形上子空间之间的距离；而主角距离则描述子空间之间的夹角。子空间距离尤其适用于可将数据建模为位于低维子空间中的领域，例如无线通信中的 MIMO 系统、生物信息学中的人脸识别和蛋白质折叠。

基于流形的距离考虑到数据往往位于弯曲的非线性空间，而非平坦的欧几里得空间中 [32]。这些距离旨在遵循数据的内在几何结构，同时捕捉局部和全局关系。测地距离例如，衡量沿流形连接两点的最短路径（通常通过邻域图上的最短路径近似）。扩散距离则反映通过此类图上的随机游走所体现的连通性。类似地，t-SNE 等流形学习方法*t-SNE* [31] 和 *UMAP* [30] 利用邻域关系构建保留局部结构的低维嵌入；随后即可在学习得到的表示空间中计算距离。

有些距离用于特定应用，例如代理 *A* 距离（*PAD*） [56]，，它根据模型区分两个数据集时的分类误差来衡量它们之间的可区分性。误差越低，表示距离越小，因此 PAD 适用于领域自适应。另一个重要指标是最大均值差异（*MMD*） [39]，，这是一种基于核的距离，通过分析两个分布在再生

表 I 数据集 X 和 **Y**的距离计算方法

类别	指标	距离计算
几何	成对欧几里得 [42]	$\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{X}_i-\mathbf{Y}_j\|_2$
	质心级欧几里得 [43]	$\|\bar{\mathbf{X}}-\bar{\mathbf{Y}}\|_2$
	簇级欧几里得 [44]	$\frac{1}{c_1 c_2}\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\|\bar{\mathbf{X}}_{C,\,i}-\bar{\mathbf{Y}}_{C,\,j}\|_2$
	余弦距离 [45]	$\frac{\sum_{i=1}^{d}\mathbf{X}_i\mathbf{Y}_i}{\sum_i P(i)\ \text{对数}\left(\frac{P(i)}{Q(i)}\right)}$
统计	Kullback-Leibler 散度 [46]	$\frac{1}{2}\sum_i P(i)\ \text{对数}\left(\frac{P(i)}{M(i)}\right)$
	詹森-香农距离散度 [47]	$+\frac{1}{2}\sum_i Q(i)\ \text{对数}\left(\frac{Q(i)}{M(i)}\right)$
	海灵格距离 [48]	$\frac{1}{\sqrt{2}}\sqrt{\sum_i(\sqrt{P(i)}-\sqrt{Q(i)})^2}$
	Wasserstein距离 [38]	$\inf_{\gamma\in\Gamma(P,Q)}\int\|x-y\|d\gamma(x,\ y)$
	MMD [39]	$\|\mathbb{E}_{X\sim P}[\phi(X)]-\mathbb{E}_{Y\sim Q}[\phi(Y)]\|^2$
	Kolmogorov-Smirnov [49]	监督学习$_x\|F_X(x)-F_Y(x)\|$
	能量距离 [50]	$2\mathbb{E}\|\mathbf{X}-\mathbf{Y}\|-\mathbb{E}\|\mathbf{X}-\mathbf{X}'\|-\mathbb{E}\|\mathbf{Y}-\mathbf{Y}'\|$
	全变差 [51]	$\frac{1}{2}\int\|P(x)-Q(x)\|dx$
子空间	格拉斯曼 [52]	$\|\theta\|_2$
	弦距离 [53]	$\sqrt{\sum_{i=1}^{k}\text{正弦}^2(\theta_i)}$
	阿西莫夫 [54]	θ_k
流形	主成分分析 [55]	嵌入：线性投影（最小重构误差）
	t-SNE [31]	嵌入：保留邻域（最小化 Kullback-Leibler 散度）
	UMAP [30]	嵌入：保留拓扑（邻居图；最小化交叉熵）

核希尔伯特空间中的均值来进行比较 [57]。MMD 常用于生成建模和分布对齐任务。

图 2 总结了各种距离类别背后的概念。然而，理解这些概念或公式，并不能充分说明它们在计算数据集距离时的有效性，尤其是在这些距离预期与模型性能相关的情况下。在接下来的章节中，我们将以数学方式定义问题，并介绍距离的评估方法。我们将探讨距离通常应具备哪些理想属性，以及数据集的哪些属性可能有助于选择合适的距离。

III. 实际计算数据集距离

本节从数学角度定义了探讨数据集相似性/距离所需的基本术语。这些术语

包括数据集、距离，以及几种明确的方式来将距离应用于数据集，即将数据集视为矩阵、应用成对函数、先对点进行聚类再应用此类函数，以及估计各特征的分布或联合数据分布并比较这些分布。此外，本节还展示了对原始数据集应用降维方法后，如何在潜在空间中进行距离计算，以及这样做为何有益。我们先从数据集的一般定义开始。

数据集：数据集由一组 N 维数据点组成，其中每个数据点都表示为一个向量$\mathbf{x} = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^N$。这种向量化形式通用且灵活，可将任意类型的数据重塑为此格式。复值数据可以通过分离实部和虚部进行转换。图像等高维张量可以展平。包含 M 个数据点的数据集 D 可表示为一个集合

$$\mathcal{D} = \{\mathbf{x}_j\}_{j=1}^{M}, \quad \mathbf{x}_j \in \mathbb{R}^N,$$

也可以表示为一个 $M \times N$ 矩阵，其中每一行对应一个数据点，每一列对应一个特征。

数据集距离：数据集距离 d 可定义为一个作用于两个数据集并输出非负数的函数，即

$$d(\mathcal{D}_1, \mathcal{D}_2) : \mathbb{R}^{M_1 \times N} \times \mathbb{R}^{M_2 \times N} \to \mathbb{R}^1.$$

这里， M_1 和 M_2 分别表示数据集 D_1 和 D_2, 中数据点的数量。请注意，两个数据集必须具有相同数量的特征 N。

矩阵距离：根据这一定义，计算数据集之间距离的首要且最直接的方法是应用矩阵距离。一种常见做法是计算两个矩阵之差，再对结果应用矩阵范数，例如弗罗贝尼乌斯范数。然而，这种方法有两个局限：i）要求数据集具有相同数量的数据点；ii）在每个数据集中打乱数据点的顺序可能会导致不同的距离值。

成对距离：第二种方法是计算每对数据点之间的距离，并在所有数据点对上进行累积或取平均值。这称为成对距离（通常是几何距离），可表示为

$$d_p(\mathcal{D}_1, \mathcal{D}_2) = \sum_{j}^{M_1} \sum_{k}^{M_2} f\left(\mathbf{x}_j^{(1)}, \mathbf{x}_k^{(2)}\right) \tag{1}$$

其中，$\mathbf{x}_j^{(i)}$ 表示数据集 D_i 中的第 j 个数据点。函数 $f : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^1$ 接收数据集中的一个二维 N维点，并返回一个非负数值。这种方法面临的挑战是需要对点之间的权重进行非常精确的加权，尤其是在高维情况下。函数 f 还可以应用于数据簇，而不是单个数据点。其缺点是需要采用聚类技术，而聚类技术本身需要对距离度量作出假设才能创建这些簇，并且会带来新的计算需求。尽管如此，聚类可以进行平均，从而在高维情况下和更大的数据集中提高鲁棒性；在将距离度量应用于数据集时，这些都是有用的特性。

分布距离：计算数据集之间距离的最后一种方法，是考虑联合分布或逐特征分布（通常使用统计距离）。在最简单的情况下，逐特征经验分布呈现为直方图的形状，可表示为一个向量，其中每个条目对应于分箱中数据点的频率。如 [58],所述，距离函数

$$d_H(\mathcal{D}_1, \mathcal{D}_2) = d(H(\mathcal{D}_1), H(\mathcal{D}_2)), \tag{2}$$

其中，直方图 H 假设了诸如分箱大小等参数，而这些参数可能因特征而异。分箱大小可以通过启发式方法确定——直方图分箱应尽可能小，以便在比较分布时提供足够的分辨率，但同时又应足够大，以包含足够的样本并对离群值保持稳健。因此，直方图的创建是在后台完成的，大多数实现通常也是如此。这里我们还假设分布得到了准确的表示。不过请注意，对于较小且高方差的数据集，该距离的准确率可能会显著下降。尽管如此，仍有多种距离函数在比较分布方面表现良好，例如能量距离和 Wasserstein距离／地球搬运工距离（EMD）。

潜在空间投影：将数据集转换到低维潜在空间 T(D) 中有两大优势：一是通过突出最重要的特征并消除冗余特征，提高距离计算的准确率；二是使这些计算在计算上更易处理。例如，将一个由 32×32 信道矩阵表示的数据集展平后，会得到一个 1024 维向量。在如此高维的空间中计算距离可能不够准确，因为要进行恰当的分布估计，数据集必须非常庞大。通过应用变换 T，数据会被投影到一个低维空间，在该空间中，距离函数

$$d_T(\mathcal{D}_1, \mathcal{D}_2) = d(T(\mathcal{D}_1), T(\mathcal{D}_2)) \tag{3}$$

在计算上高效，并能反映有意义的数据关系。

降低维度对于基于分布的距离尤其有益，例如Wasserstein距离，这类距离依赖于对多维分布或直方图的估计。在高维空间中，要准确估计这些分布，需要显著更多的数据。降低维度可以简化估计过程，使计算数据集之间的准确距离更加可行。因此，潜在空间投影不仅减轻了计算负担，还能确保距离更好地捕捉数据的内在属性，从而增强可解释性并改善性能预测。

既然已经定义了对数据集进行操作的工具，我们现在可以描述这些距离试图解决的挑战。

IV. 数据集相似性问题定义

在本节中，我们定义数据集相似性问题：计算数据集之间的距离，使其与

这些数据集上的模型性能高度相关。因此，如果两个数据集相似，数据集距离就应该较小，而在一个数据集上训练的机器学习模型应能很好地泛化到另一个数据集。反之，如果距离较大，预期泛化性能会较差。解决这一问题将使我们能够预测在一个数据集上训练的模型在另一个数据集上的表现，而无需进行大量训练和测试。正如前文所述，这对于迁移学习、领域自适应以及模型选择或切换都很有价值。为阐明这一问题，我们将介绍机器学习模型如何在数据集上进行训练和测试，以及如何在使用多个数据集时计算测试分数之间的相关性——这一计算是衡量距离度量质量所必需的。最终目标是创建一种距离度量，准确反映数据集在任务性能方面的关系。

特定任务模型训练：模型定义（架构）、损失函数的选择、训练参数、数据划分及其他配置，都取决于当前的具体任务和数据集。为简便起见，我们仅表示与任务和数据集的依赖关系。首先，我们将针对任务 T 构建并在数据集 D 上训练的机器学习模型 M 定义为

$$\mathcal{M}_{\mathcal{D}}^{\mathcal{T}} = \mathcal{M}^{\mathcal{T}\,\text{train}}(\mathcal{D})\,. \tag{4}$$

还需注意，训练操作通常使用数据集 D 中的特征子集，这些特征称为输入特征。对于无监督任务，所有特征都可能是输入特征。在有监督任务中，用作模型输入和输出的特征并不相同。

特定任务模型推理：训练成功后，也就是说，训练损失充分降低后，向模型提供与训练数据集相似的输入，预期会得到与相应输出训练数据相似的输出。下一步是在可能不同的数据集上对该模型进行推理。这一步与衡量模型可迁移性有关。假设模型在源数据集 D_S 上训练，并在目标数据集 D_T上测试，则模型输出可表示为

$$\mathcal{M}_{\mathcal{D}_S\to\mathcal{D}_T}^{\mathcal{T}} = \mathcal{M}_{\mathcal{D}_S}^{\mathcal{T}\,\text{test}}(\mathcal{D}_T)\,. \tag{5}$$

这些输出随后可用于损失函数，以确定模型在该任务上的性能。

可迁移性性能指标：损失 L 量化了模型在目标数据集上的表现。在此情境下， L 可用于确定哪个源数据集 D_S 更适合用于训练 $\mathrm{M}_{\mathcal{T}}$，以便在目标数据集 D_T上进行推理。具体做法是选择能在目标数据集上取得最高性能的源数据集。或者，等价地，选择与理想性能相比性能下降最小的数据集；理想性能是指模型在目标数据集上进行训练时所达到的性能。用数学形式表示，我们可以将两个数据集之间的模型可迁移性性能 P 定义为

$$\mathcal{P}(\mathcal{D}_i,\mathcal{D}_j) = L\left(\mathcal{M}_{\mathcal{D}_j\to\mathcal{D}_j}^{\mathcal{T}}\right) - L\left(\mathcal{M}_{\mathcal{D}_i\to\mathcal{D}_j}^{\mathcal{T}}\right) \tag{6}$$

距离矩阵和性能矩阵：为了关联距离与性能，我们需要多个示例。每种情况都需要多个数据集。考虑K 个数据集后，我们可以将距离和可迁移性性能汇总成矩阵：

$$\mathbf{D} = \left[d_{i,j}\right] \;:\; d_{i,j} = d(\mathcal{D}_i,\mathcal{D}_j) \tag{7}$$

$$\mathbf{P} = \left[p_{i,j}\right] \;:\; p_{i,j} = \mathcal{P}(\mathcal{D}_i,\mathcal{D}_j) \tag{8}$$

其中 $i,\ j\ =\ 1,\ \ldots\ldots,\ K$。这些矩阵可以绘制为混淆矩阵，如第七节图 3 所示。此处给出的公式允许取每个矩阵中的元素进行相关性计算。

距离与性能之间的相关性：令ρ**(d，p)**表示距离向量 **d** 与性能向量**p**之间的相关系数，它们分别表示数据集对之间的距离和模型性能。我们将**d** 和 **p**定义为

$$\mathbf{d} = \mathrm{vec}(\mathbf{D}) = \left[d_{1,1},\ldots,d_{1,K},d_{2,1},\ldots,d_{2,K},\ldots,d_{K,K}\right] \tag{9}$$

$$\mathbf{p} = \mathrm{vec}(\mathbf{P}) = \left[p_{1,1},\ldots,p_{1,K},p_{2,1},\ldots,p_{2,K},\ldots,p_{K,K}\right], \tag{10}$$

使我们的目标能够被写成 s

$$d_{\mathcal{T}}^{\star} = \underset{d_{\mathcal{T}}}{\mathrm{argmax}}\ \rho(\mathbf{d},\mathbf{p}) \tag{11}$$

其中，距离向量**d** 取决于距离函数$d_{\mathcal{T}}$和数据集$\mathcal{D}$ = {$\mathrm{D}_1,\ldots,\mathrm{D}_K$}。性能向量 **p** 是 $\mathcal{D}$ 和模型 M 的函数。问题在于寻找一个最优距离 $d_{\mathcal{T}}^{\star}$，使该函数生成的距离与性能之间的相关性最大。

从公式（11）出发，下一节介绍用于计算数据集距离并将其与模型性能进行比较的分析框架。随后各节将采用这一框架，并将其应用于特定数据集和任务，以评估所计算的距离与性能之间的相关性是否较高。我们的目标是推导出一种通用方法，用于构建有效、面向特定任务且低复杂度的数据集距离。

V. 评估数据集距离与模型性能相关性的框架

为应对如何选择能够准确预测不同数据集上模型性能的距离函数这一挑战，我们提出了一个综合框架，将计算得到的数据集距离与模型性能指标相关联。该框架通过分析各种距离度量能否反映在数据集之间迁移模型时的性能下降，来评估其适用性。通过有效衡量能够反映模型差异的数据集相似性，我们的框架可用于开发检测数据集偏移的距离、指导生成式数据增强，并为特定任务对最有用的数据集进行排序。

框架概述：在该框架中，每个数据集都要经过两个主要过程：距离计算和性能评估。距离计算步骤计算数据集对之间的距离度量 d，从而生成距离矩阵 **D**。性能评估包括在一个数据集上训练

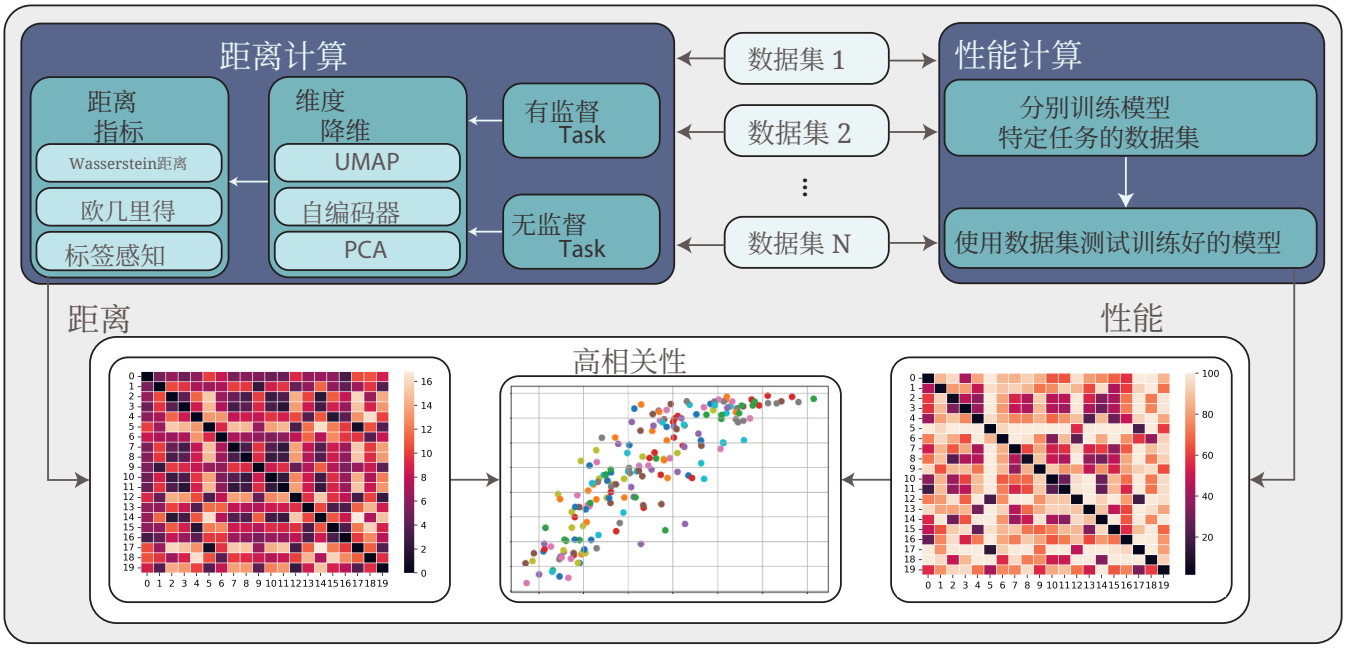

图 3. 用于评估给定任务和模型的数据集距离度量的框架：如果根据该度量，两个数据集彼此接近，那么在其中一个数据集上训练的模型在另一个数据集上测试时应能取得较高性能。使用给定度量计算的距离与模型在给定任务上的性能之间的相关性越高，该度量就越适合用于衡量该任务中的数据集相似性。

机器学习模型，并在其他数据集上对其进行测试，从而得到性能矩阵 **P**。通过将**D** 和**P**向量化并进行相关性分析，我们可以识别其中的模式，并评估距离对模型性能下降的预测效果。该框架的示意图如图 3 所示。

性能计算：性能评估首先是在每个数据集 D_i 上训练模型 M_i，并在来自同一数据集的测试集 P_{ii} 上评估其性能。这为模型在其训练域内确立了基线性能。随后，在其他数据集 D_j（针对 $j \neq i$）上测试训练好的模型 M_i，并记录性能P_{ij}。如果两个数据集 D_i和 D_j相似，我们预期 P_{ij} 会接近 P_{ii}；反之，性能大幅下降则表明数据集之间存在显著差异。目标是评估数据集之间的较大距离d_{ij} 是否对应较大的性能下降$\Delta P_{ij} = P_{ii} - P_{ij}$。

距离计算：距离计算步骤会计算每对数据集D_i和 D_j之间的标量距离 d_{ij}。可采用多种距离度量，包括基于经验概率密度函数（PDF）的统计距离，或在变换空间中计算的距离（例如降维之后的距离）。计算詹森-香农散度或最大均值差异（MMD）等统计距离时，一个关键环节是准确估计数据集的经验 PDF。

联合估计数据集的概率密度函数：主要挑战在于联合估计两个数据集的经验概率密度函数，以确保数据集之间具有可比性。这需要为两个数据集构建箱边界完全相同的直方图。设 $D_i = \{\mathbf{x}_n^{(i)}\}_{n=1}^{M_i}$ 和 $D_j = \{\mathbf{x}_n^{(j)}\}_{n=1}^{M_j}$ 为数据样本，分别来自数据集 D_i和 D_j，其中 M_i和M_j 分别为每个数据集中的样本数量。

我们定义一组通用的箱边界 $\{b_k\}_{k=0}^{K}$，使其覆盖 D_i和 D_j的合并范围：

$$b_0 = \min\left(\min(D_i), \min(D_j)\right), \tag{12}$$

$$b_K = \max\left(\max(D_i), \max(D_j)\right), \tag{13}$$

其中 K 为箱的数量，可使用诸如 $K = \sqrt{M}$的启发式方法设定，其中 $M = \max(M_i, M_j)$。随后，我们针对每个数据集 D_i计算归一化直方图，从而得到经验 PDF：

$$h_i(k) = \sum_{n=1}^{M_i} \delta\left(b_{k-1} \le \mathbf{x}_n^{(i)} < b_k\right), \quad k = 1, 2, \ldots, K, \tag{14}$$

其中 $\delta(\cdot)$为指示函数。随后，我们对直方图计数进行归一化，以获得经验 PDF。

$$p_i(k) = \frac{h_i(k)}{\sum_{k=1}^{K} h_i(k)} = \frac{h_i(k)}{M_i}, \tag{15}$$

$$p_j(k) = \frac{h_j(k)}{\sum_{k=1}^{K} h_j(k)} = \frac{h_j(k)}{M_j}. \tag{16}$$

通过使用通用箱边界和相同数量的箱，我们确保 p_i 和p_j 可以直接比较。K 和$\{b_k\}$的选择至关重要：箱数过多可能导致出现空箱，使 PDF 估计不准确；箱数过少则可能过度平滑分布，掩盖重要差异。

灵活性与挑战：我们的框架具有灵活性，可根据任务和数据集的特点执行预处理步骤，例如降维或聚类。高维数据集在距离

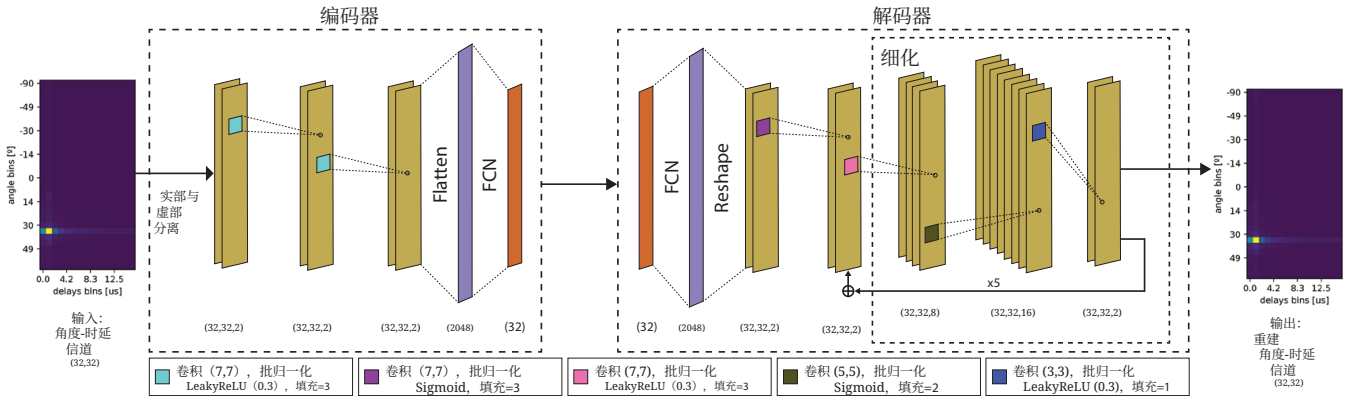

图4. 架构　　用于无监督 CSI 压缩任务的模型。该模型深受 CSINet 启发+ [59　　].

计算方面面临维度灾难带来的挑战。降维技术（如主成分分析、UMAP）可以将数据投影到较低维度的空间中，从而缓解这些挑战；在这些空间中，距离更有意义，也更易于计算。此外，聚类可以通过聚类质心概括数据集，进一步简化问题并降低计算复杂度。不过，需要注意的是，某些聚类算法可能会带来额外的计算开销。

依赖因素与注意事项：该框架的有效性取决于多个因素：

- **数据集特征**：数据分布、样本量和特征空间的变化会影响距离计算和模型性能。
- **任务性质**：有监督和无监督任务可能需要采用不同的距离计算方法（例如，标签感知方法与标签无关方法）。
- **模型选择**：在不同数据集上使用一致的模型可以最大限度地减少模型依赖性带来的差异，从而更清晰地分析数据集相似性。

通过仔细考虑这些依赖因素，我们旨在确保计算出的距离能够准确反映模型性能差异。在后续章节中，我们将把该框架应用于有监督和无监督任务，并使用所述经验概率密度函数估计方法评估各种距离度量。通过全面评估数据集距离化的潜在候选方法，我们旨在找出最有效的距离度量，并将研究结果推广到其他领域。

VI. 通过潜在空间投影实现数据集距离化

在高维数据集中，计算能够准确反映数据集间关系并与模型性能相关的距离是一项艰巨的任务，主要原因在于噪声、无关特征以及数据本身的复杂性。传统距离度量应用于原始高维空间时，往往无法捕捉关键的底层结构，这是因为噪声和无关特征的存在。为解决这一问题，必须将数据集映射到经过变换的空间中，以突出最符合当前任务需求的特征。这类似于 Johnson-Lindenstrauss (JL) [60] 定理等降维技术致力于保留距离，或主成分分析[55]旨在提高可分性的做法。我们需要的变换既要保留**局部邻近性**——确保在任务相关特征方面彼此接近的数据集在新空间中仍然接近——也要保留**全局结构**，使数据集之间更广泛的关系得以维持。这样，我们就能计算出对任务更有意义的距离，从而提高这些距离与不同数据集上模型性能之间的相关性。

这种转换可以通过**基于图的方法**来实现，根据数据集的局部邻域和全局连接对它们之间的关系进行建模。该图捕捉了数据集在与任务相关的特征方面的关联，确保特征相似的数据集在转换后的空间中聚集在一起。同时，这种方法保留了全局结构，确保彼此距离较远、但在与任务相关的更广泛层面上具有相似性的数据集也能得到**恰当的流形式表示**。因此，转换后的空间成为一种**类似流形的表示**，其中数据集之间的距离不仅反映它们在特征空间中的接近程度，也体现其结构属性，具体取决于任务。这样，我们就能计算出更有可能与在一个数据集上训练的模型在另一个数据集上的表现相关联的距离，最终改进迁移学习、领域自适应和模型选择。这种面向任务的数据转换方法确保我们保留最有意义的信息，同时滤除噪声，从而更准确、高效地比较数据集。

本节将介绍一种新颖的方法，利用**统一流形逼近与投影（UMAP）**将原始数据转换到编码后的潜在空间，从而测量数据集之间的距离。按照前文的定义，设数据集为 $D_i = \{\mathbf{x}_j^{(i)}\}_{j=1}^{M_i}$，每个数据集包含 M_i 个数据点，记作 $\mathbf{x}_j^{(i)} \in \mathbb{R}^N$。我们的目标是在能够比原始特征空间更有效地体现数据集内在几何性质和拓扑性质的空间中，计算这些数据集之间的距离。

将 UMAP 用于潜在空间：我们使用 UMAP 将

高维数据投影到低维潜在空间中。此转换由函数f_{UMAP}定义：$\mathbb{R}^N \to \mathbb{R}^d$，其中$d \ll N$。随后，编码数据集表示为$\tilde{D}_i = \{\tilde{\mathbf{x}}_j^{(i)} = f_{\text{UMAP}}(\mathbf{x}_j^{(i)})\}_{j=1}^{M_i}$。UMAP 通过在高维空间中构建加权 k 近邻图，形成模糊拓扑表示，同时捕捉局部和全局数据结构。之所以采用模糊表示（即非二元的集合归属关系），部分原因在于它们能够为迭代优化提供更平滑的代价函数。随后，UMAP 使用模糊集合的加权集合归属关系，对低维嵌入进行优化，以保留该模糊单纯形集。UMAP 算法中实现这些功能的步骤包括：

1)**从高维数据构建模糊单纯形集**：UMAP 构建一个加权图，其中权重表示数据点之间的连接概率，从而有效捕捉局部邻域关系。2) **定义模糊拓扑表示**：随后，通过考虑这些概率，UMAP 创建一个模糊拓扑空间，用以表示更广泛的数据流形结构。3) **低维空间中的优化**：UMAP 寻找一种低维嵌入，使高维和低维空间中的模糊拓扑表示之间的交叉熵最小化。

与其他流形学习技术相比，UMAP 采用这种方法能够保留更多数据的全局和局部结构。原因如下：i) PCA 是线性的，可能无法捕捉非线性结构；ii) t-SNE 关注局部邻域，可能会扭曲全局结构；iii) 自编码器高度依赖网络架构和训练过程，更侧重于数据重构，而非维持内在关系。

UMAP 空间中的欧几里得距离：在潜在空间中，我们探索欧几里得距离的不同形式，以量化 D_1 与 D_2 之间的间隔：

1)**成对欧氏距离**：计算两个数据集中所有点对之间的距离，

$$d_{\text{pairwise}} = \frac{1}{M_1 M_2} \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \left\| \tilde{\mathbf{x}}_j^{(1)} - \tilde{\mathbf{x}}_k^{(2)} \right\|_2 . \tag{17}$$

该指标通过考虑所有可能的点对点距离，提供全面的度量。

2)**基于聚类的欧氏距离**：使用 k 均值 [61],层次聚类[62], 或 DBSCAN [63] 等聚类算法，将每个~编码后的数据集 D_i 划分为 K_i 个簇。每个簇由其质心$\mathbf{c}_l^{(i)}$ 表示，该质心是该簇中各点的平均值：

$$\mathbf{c}_l^{(i)} = \frac{1}{|C_l^{(i)}|} \sum_{\tilde{\mathbf{x}} \in C_l^{(i)}} \tilde{\mathbf{x}}, \tag{18}$$

其中，$C_{(i)l}$是数据集 D_i 中簇 l 内的点集，$|C_l^{(i)}|$ 是该簇中的点数。

基于聚类的欧几里得距离定义如下：

$$d_{\text{cluster}} = \frac{1}{K_1 K_2} \sum_{l=1}^{K_1} \sum_{m=1}^{K_2} \left\| \mathbf{c}_l^{(1)} - \mathbf{c}_m^{(2)} \right\|_2 . \tag{19}$$

这种方法能突出簇级别的结构差异，但其计算复杂度取决于所采用的聚类技术。

3) **平均欧氏距离**：当 $K_1 = K_2 = 1$时，这是基于聚类的距离的一种特殊情况，可简化为数据集均值向量之间的距离，

$$d_{\text{average}} = \left\| \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} \right\|_2, \text{ where } \bar{\mathbf{x}}^{(i)} = \frac{1}{M_i} \sum_{j=1}^{M_i} \tilde{\mathbf{x}}_j^{(i)}. \tag{20}$$

这提供了一种粗略但计算效率最高的数据集分离度度量。

UMAP 空间中的 Wasserstein距离：除了欧几里得度量，我们还将每个数据集视为一个多元分布，计算潜在空间中数据集之间的 Wasserstein距离（地球搬运距离）。为简化计算，我们按维度分别处理这些分布，从而能够计算潜在空间中每个维度 n 的一维 Wasserstein距离。数据集 D_1 和 D_2 的n维经验分布之间的一维 Wasserstein距离定义为：

$$W_n = \int_0^1 \left| F_n^{(1)^{-1}}(t) - F_n^{(2)^{-1}}(t) \right| dt, \tag{21}$$

其中，$F_n^{(i)^{-1}}(t)$ 是数据集 D_i的第 n维的逆累积分布函数（分位数函数）。直观而言，Wasserstein距离衡量的是将一个分布的概率质量搬运至与另一个分布相匹配所需的最小成本，其中成本由搬运的概率质量（泥土）乘以搬运距离来衡量。

具体来说，该积分计算的是数据集在维度 n上的两个累积分布函数（CDF）之间的面积。这表示将单位概率质量从一个分布搬运到另一个分布所需的平均距离。数据集之间的总体 Wasserstein距离则是所有维度上的平均值：

$$W = \frac{1}{d} \sum_{n=1}^{d} W_n. \tag{22}$$

该度量能够有效捕捉数据集在所有维度上的分布差异，为衡量它们的相似度提供稳健的指标。当满足以下条件时，它优于欧几里得距离：i）数据集包含足够多的点，能够准确估计经验分布；ii）维度数量较少，因为过多的维度可能会稀释在更相关的特征上计算出的分布距离值。接下来，我们将探讨如何进一步改进这种距离方法，使其能够适用于高维潜在空间。

结合特征重要性进行增强：当已知潜在空间中的某些特征比其他特征更重要时，可以扩展欧几里得指标和 Wasserstein距离指标。可以使用以下方法评估特征重要性：

例如置换特征重要性（PFI） [64], 偏依赖（PD） [65] 图、局部可解释模型无关解释（LIME） [66], 或 SHAP（SHapley 加性解释） [67] 值。根据特征重要性对距离进行加权，可以得到一种更细致的度量，反映每个维度的相对重要性。

重要注意事项：我们利用 UMAP 创建潜在空间，以忠实地表示数据集的底层结构。不过，这种 UMAP 的应用需要注意一些细节。首先，将 UMAP 应用于高维输入空间时，计算成本可能很高。因此，可以先使用 PCA 或其他计算成本较低的（可能是线性的）降维技术来缩减输入空间，然后再在这个中间潜在空间上使用 UMAP。此外，UMAP 需要通过距离函数来比较数据点。对于无线信道，我们发现最合适的距离函数是相关距离。不过，如果复杂度和可扩展性比性能更重要，也可以使用效率更高的欧几里得距离。此外请注意，将 UMAP 应用于一组数据集时，应在单个 UMAP 模糊集中联合处理所有数据集。否则，无法保证独立编码彼此一致，从而可能导致错误的距离。

UMAP 参数：为潜在空间中的最小距离和邻居数量等参数设置适当的数值至关重要。在 K 近邻（KNN）语境下，邻居数量决定了 UMAP 中每次计算会受到多少个最近邻居的影响。邻居数量越多，对全局结构的优先级就越高。邻居数量较少则能更好地保留局部结构，但会牺牲对全局的理解。平衡非常关键；对于所研究的数据集（每个数据集大约包含 5000 个样本），我们发现 32 个邻居的效果很好。点之间的最小距离应保持较小，以便更广泛的数据关系得以体现。换言之，如果数据点之间的关系相似，较小的距离允许它们在表示中彼此更接近。我们选择潜在空间中最小距离为 0.1。

总体而言，与传统方法相比，这种方法具有显著优势，能够对数据集相似性进行更细致且更易于解释的评估。接下来，我们将在无监督 CSI 压缩任务中应用所提出的距离，并将其与传统方法进行比较。

VII. 无监督情形：CSI 压缩

本节将探讨在一个无监督机器学习任务——CSI 压缩——中如何将距离应用于数据集。我们选择先从无监督学习任务入手，因为用于模型输入和作为标签的数据是相同的，即 $\mathrm{N}^{I_{\mathcal{T}}}=\mathrm{N}^{O}_{\mathcal{T}}=\mathrm{N}$。在此任务中，我们探究距离与模型性能之间的相关性。总之，第四节中定义的变量取值如下：

- T：CSI 压缩
- P：编码信道与真实信道之间的归一化均方误差
- M：卷积自编码器，架构如图4所示
- $\mathcal{D}$：来自 ASU Campus DeepMIMO数据集的 20 个区域

在本节剩余部分中，我们将按照上述顺序更详细地介绍各变量。在说明第五节框架的适用条件后，我们将展示所得结果并进行讨论。

CSI压缩任务：信道压缩的过程涉及将高维无线信道（可能包含数百或数千个复数条目）转换为低维表示，通常不超过几十个实数条目。信道反馈通常以原始形式占用大量空间，通过这种压缩技术可以变得更加紧凑，从而更高效地传输信道信息。无线信道可以表示为一个矩阵 $\mathbf{H}$ ，其维度为 N_{BS} 个基站天线乘以 N_{sub}个子载波。通常，用于压缩的信道会先经过预处理：沿矩阵的每个维度进行傅里叶变换，将其转换到角度-时延域，再将所得矩阵展平。在此任务中，我们考虑 $N_{BS}=32$ 和 $N_{sub}=32$；将信道转换到角度-时延域后，去掉最后 16 个时延抽头，因为它们几乎不包含任何信息。编码维度设为 $N_{enc}=32$，从而实现 32 倍压缩率。对于性能指标 P，在常见的重建任务中，通常使用归一化均方误差（NMSE）：

$$\mathrm{NMSE}_{\mathrm{dB}}(\mathbf{H},\hat{\mathbf{H}})=10\log_{10}\frac{||\mathbf{H}-\hat{\mathbf{H}}||_F^2}{||\mathbf{H}||_F^2}. \tag{23}$$

自编码器模型：用于执行信道重建任务的模型是一个自编码器，其架构如图4所示。该图展示了一个 32×16 复数信道矩阵，通过分离实部和虚部将其扩展为 $32\times16\times2$ 。随后，该矩阵经过若干卷积层和一个全连接层，进入以橙色标示的32维潜在空间。接着，解码器接收该潜在空间，并通过一个全连接层将其重建到高维空间。经过重塑后，数据再通过卷积层和细化网络，该过程重复三次，最终输出重建的信道，并使用最小平方误差（MSE）损失与输入进行比较。值得注意的是，该模型通过两种方式进行训练：首先，为了进行性能评估，每个区域分别训练一个自编码器。在同一区域进行测试时，模型的归一化均方误差（NMSE）损失低于 –20dB，但在未训练过的区域中损失较高。其次，为了进行距离计算，使用所有数据集的数据训练一个特殊的单一模型，该模型包含五个细化网络，而非三个，在32维空间中实现了强大的压缩效果，各区域的平均归一化均方误差（NMSE）为 –20 dB。该模型用于降维，将输入空间转换为编码空间以进行距离计算。之后，我们将原始空间和潜在空间中的距离结果进行比较，结果表明，自编码器既能实现高效编码，也具有出色的数据集距离化性能。

数据集： 我们使用基于 DeepMIMO 生成的亚利桑那州立大学（ASU）校园光线追踪数据集。该数据集包含约 130000 个数据点，其中

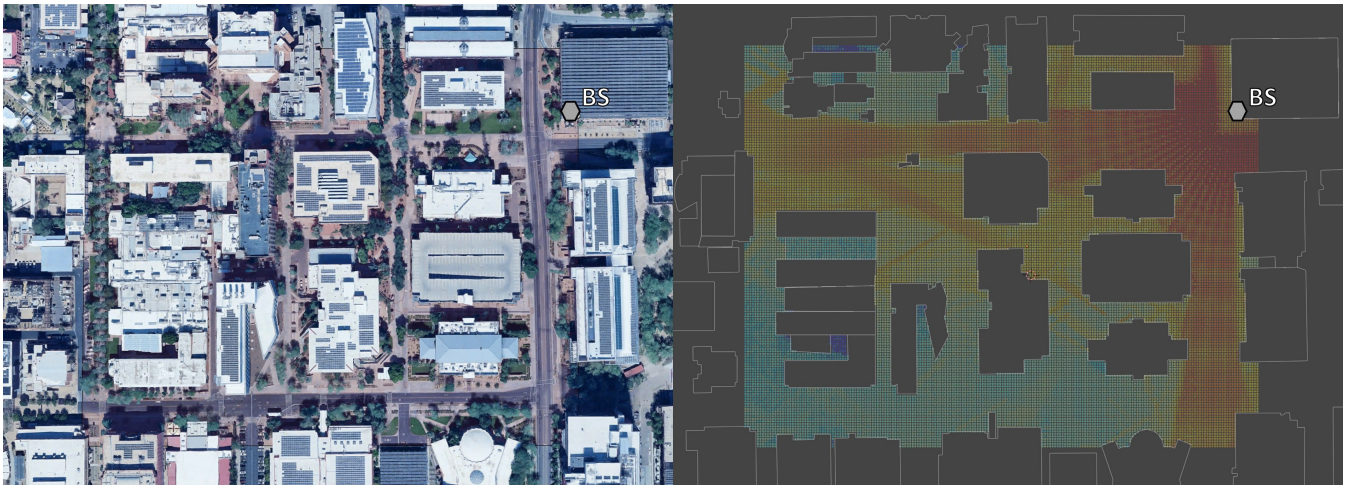

图 5. ASU 校园 DeepMIMO 数据集的真实（左）和渲染（右）俯视图。两幅图中均展示了基站。需要注意的是，建筑物和其他场景资源都是 3D 的，其高度对屋顶绕射有显著影响。合成对应物中的网格表示基站采用标准离散傅里叶变换码本时的接收功率。

约 90000 个代表具有有效信道的用户（位于建筑物外），覆盖面积为 410 米 × 320 米，用户分辨率为每米一名。射线追踪仿真考虑了多条传播路径，包括视距、反射、漫散射和绕射，因此仿真相当全面且更贴近现实。图 5 展示了该数据集的保真度，其中对比了来自谷歌地球的真实地理数据与数字渲染图。

输入空间中的距离：表 VII 展示了通过计算信道状态信息压缩任务的性能与采用多种距离方法得到的数据集距离之间的相关性而获得的结果。我们展示了三类距离（几何、统计和子空间距离）以及 PAD。由于前文已将聚类和降维方法定义为计算距离的辅助工具，本表省略了依赖降维方法的流形方法。此外，我们还纳入了聚类距离和质心欧几里得距离，前者分别考虑 3 个簇，后者仅考虑 1 个簇的质心。从计算时间可以看出这些选择的影响：计算均值远快于对所有数据进行聚类。总体而言，Wasserstein距离和能量距离（均属于统计积分概率度量）的表现更优，与模型性能的相关性分别为 0.52 和 0.55。积分概率度量优于 f-散度。几何距离受到维度灾难的影响，相关性较低。子空间距离也有类似情况：尽管需要消耗计算资源，其表现仍然不佳。此外，子空间距离表现不佳也可能归因于主角的计算方式，这使得难以察觉非常相似的子空间之间的差异。原始空间中表现最好的距离是 PAD。PAD 能够学习如何对数据进行分类，因此在维度较高且数据易于分离时也能表现良好。

降维：降维等空间变换方法有助于将原始数据集中的信息提炼到一个便于计算距离的空间中。我们在图 6 中评估了其中几种方法：将由地理位置彼此接近的信道构成的 20 个数据集

表 II 在原始空间中计算的距离及模型性能的计算时间和相关性

类别	距离名称	相关性	计算时间（秒）
几何	成对欧氏距离	0.36	11
	聚类欧氏距离	0.37	166
	质心欧氏距离	0.34	3
	余弦距离	-0.07	5255
统计	詹森-香农距离	0.14	80
	海灵格距离	0.15	81
	Wasserstein距离	**0.52**	562
	柯尔莫哥洛夫-斯米尔诺夫	0.47	381
	全变差	0.15	78
	MMD（线性）	-0.08	79
	最大均值差异（径向基函数）	-0.06	135
	能量距离	**0.55**	735
子空间	格拉斯曼	-0.11	15223
	弦距离	-0.10	14810
	阿西莫夫	-0.03	15594
其他	PAD	**0.64**	952

转换到二维空间，以便于可视化。我们的目标是直观地了解这些方法，并将二维表示与它们的特性联系起来。图中显示，PCA 编码的信息在各数据集之间看起来非常相似，无法区分它们。另一方面，t-SNE 能够分离这些数据集，但无法保留全局结构。这一点从数据集彼此混杂、相似的信道并未彼此靠近可以看出。而 UMAP 则保留了原始数据中的局部和全局结构，同时舍弃了几乎所有其他信息。

潜在空间中的距离：在通过各种降维技术获得的编码空间中，我们观察到距离反映数据集相似性的趋势始终一致。在这些空间中，子空间距离始终较小，这表明数据集在编码后彼此高度对齐。然而，大多数统计距离——例如詹森-香农距离、海灵格距离和全变差距离——表现并不

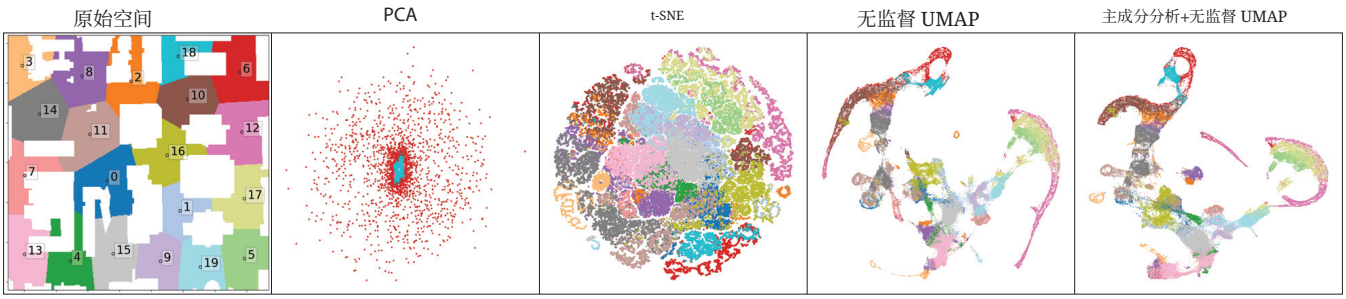

图6. 将不同降维方法应用于我们的数据集后得到的潜在空间可视化。从左到右依次为：通过基于邻近度的聚类选取各数据集后的原始空间、主成分分析、t-SNE、UMAP，以及主成分分析（将数据集降至 ~ 100 分量）与 UMAP 的组合。

理想。造成这种表现不佳的一个可能原因是，这些距离的取值范围仅限于 0 到 1，因此在处理不相交的分布时，敏感度可能受到限制。相比之下，Wasserstein距离没有这一限制，其数值可以达到大得多的范围（例如数千），因此在区分分布差异显著的数据集时更为稳健。

比较编码空间：我们的结果表明，在主成分分析编码空间中，PAD 在各项评估指标中表现最佳。这可能是因为主成分分析作为一种线性降维方法，能够保留数据的全局结构，使 PAD 可以有效区分不同数据集。尽管 t-SNE 仅将数据降至两个分量，但其平均表现往往优于主成分分析，这表明 t-SNE 对局部数据结构的侧重增强了其有意义地组织数据的能力。UMAP 在所有情况下都稳定优于 t-SNE，其中欧几里得距离和 Wasserstein距离在该空间中的表现尤为出色。基于欧几里得距离的指标在 UMAP 中表现强劲，说明编码空间非常适合进行距离计算，因为 UMAP 流形学习所确立的内在几何结构发挥了作用。同样，Wasserstein距离表现出色，这是因为它能够捕捉分布差异，且不受有限取值范围的限制。在这里，经过所有区域训练的自编码器可作为基准，衡量通过以重构为重点的编码所能达到的最高性能。虽然它展现了性能上限，但在此情境下，与其他方法相比，其实际用途可能有限，因为计算距离需要训练多个独立模型。此外，正如我们在第IX节中所示，它在监督任务上的表现远不及其他方法。

无监督任务的要点：关键要点是，选择合适的降维方法可能会产生重大影响，有时甚至比选择之后应用的最优距离度量更为重要。尽管如此，Wasserstein距离和欧几里得距离分别凭借其鲁棒性和简洁性表现出色。在评估无监督任务中的数据集距离后，我们展示了两项监督任务的结果。这一点尤其重要，因为无线通信中的大多数机器学习任务都是有监督的，但许多距离度量并未将标签视为数据集中的独立元素，而是像处理输入一样处理它们

表III 模型性能与在通过保留不同降维方案（即主成分分析、t-SNE、UMAP 和自编码器）的 2 个或 32 个分量获得的降维空间中计算的距离之间的相关性。

类别	距离	降维			
		主成分分析32	t-SNE2	UMAP 2	自编码器 32
几何	成对欧氏距离	0.37	0.58	**0.83**	**0.87**
	聚类欧氏距离	0.41	0.59	**0.84**	**0.91**
	质心欧氏距离	0.35	0.59	**0.86**	**0.93**
	余弦距离	0.30	0.41	0.42	**0.94**
统计	KL散度	0.32	0.62	0.52	**0.85**
	詹森-香农距离	-0.08	0.15	0.12	0.07
	海灵格距离	-0.08	0.17	0.13	0.07
	Wasserstein距离	0.47	0.68	**0.85**	**0.92**
	柯尔莫哥洛夫-斯米尔诺夫	0.57	0.32	0.46	0.22
	全变差	-0.06	0.13	0.10	0.04
	MMD（线性）	-0.17	0.10	0.04	0.04
	最大均值差异（径向基函数）	-0.13	0.06	0.04	-0.02
	能量距离	0.56	0.42	0.60	0.25
子空间	格拉斯曼	-0.14	0.02	-0.22	-0.05
	弦距离	-0.14	0.02	-0.22	-0.05
	阿西莫夫	-0.12	0.03	-0.23	-0.06
其他	PAD	**0.75**	0.68	0.71	0.66

。为此，我们提出了一种纳入标签的新距离度量，其性能可能优于忽略标签信息的方法。

VIII. 标签感知数据集距离

本节介绍一种名为**标签感知数据集距离**的新型度量方法。该方法纳入标签信息，以改进对带标签数据集之间距离的计算。传统上，距离计算仅基于输入特征，或将输入按标签分组，这会导致输入与输出之间关系的丢失。然而，通过有意利用标签信息，我们可以更全面地衡量数据集相似性，尤其是在不同数据集的标签分布存在差异时。

适用性：标签感知方法既可应用于编码空间（例如由 UMAP 生成的空间），也可应用于原始特征空间。该方法的关键组成部分是计算**惩罚项**，以处理不平衡标签分布。当一个数据集包含另一个数据集中没有的标签时，这两个数据集就被视为不平衡。对于标签并集 $L = L_1 \cup L_2$ 中的每个标签 l，其中 L_i 是数据集 D_i 中的标签集合，我们定义一个惩罚项 P_l 和一个标签特定距离 d_l。

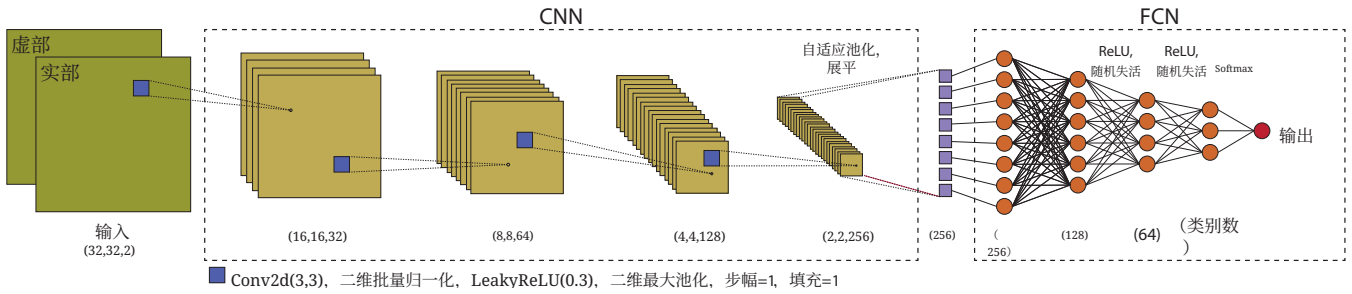

图 7. 用于监督任务的卷积神经网络模型

惩罚项 P_l**：** 标签 l 的惩罚项设为所有待考察数据集中任意两个带有标签 l 的点之间的最大距离。设共有 S 个数据集，并令 D_s^l 表示数据集 D_s 中仅包含标签 l 样本的子集，其中 $s \in \{1, 2, \ldots\ldots, S\}$。惩罚项定义如下：

$$P_l = \max_{\substack{\mathbf{x}\in D_p^l,\ \mathbf{y}\in D_q^l \\ p,q\in\{1,2,\ldots,S\}}} d(\mathbf{x},\mathbf{y}),$$

其中， $d(\mathbf{x}, \mathbf{y})$ 是所选的距离度量（例如，编码空间中的欧几里得距离）。此惩罚项确保所考察的数据集组内保持一致，并考虑标签特定的差异，因为某些标签可能天然具有较大的距离。

标签特定距离 d_l**：** 对于两个数据集 D_1 和 D_2中都存在的标签，我们仅使用带有标签 l的样本来计算距离：

$$d_l = d\left(D_1^l, D_2^l\right),$$

其中， $d\left(D_1^l,\ D_2^l\right)$ 表示应用于标签 l对应子集的距离计算方法（例如，成对欧氏距离）。这实际上将问题转化为在每个标签类别内进行无监督距离计算。对于只出现在一个数据集中的标签，我们赋予 $d_l = P_l/2$ 作为惩罚，以惩处这种不平衡

$$d_l = \begin{cases} d\left(D_1^l, D_2^l\right), & \text{if } l \in L_1 \cap L_2, \\ P_l/2, & \text{if } l \in L_1 \Delta L_2, \end{cases}$$

其中， Δ 表示集合之间的对称差。

聚合标签特定距离： 计算每个标签的距离后，我们将它们聚合起来，得到总体标签感知数据集距离 $d_{\text{label-aware}}$：

$$d_{\text{label-aware}} = \frac{1}{|L|}\sum_{l\in L} d_l,$$

其中， $|L|$ 是L_1 和 L_2的并集中不同标签的总数。两个数据集中均不存在的标签不纳入计算，以确保只有相关标签会影响距离。

通过纳入针对某个数据集中缺失标签的惩罚项，标签感知距离能够有效惩罚不平衡标签分布，反映出在此类数据集之间迁移模型时可能面临的挑战。A

一个简单的例子是，仅使用视距数据训练的模型难以预测非视距标签。这种方法能够更细致地比较数据集，同时考虑标签的存在与分布，并可与现有的距离计算方法结合，应用于编码后的特征空间或原始特征空间。下一节将展示这种方法与其他应用于监督任务的方法的比较。

IX. 监督情形：波束预测

本节将数据集距离度量应用于一项有监督机器学习任务，具体而言，即波束预测。用于这些监督任务的数据集与无监督情形中使用的数据集相同，只是针对每项任务添加了相应的标签。通过将数据集距离度量应用于这些监督学习问题，我们旨在评估在无监督任务中表现良好的方法在有监督任务中的表现。此外，我们还分析通过所提出的标签感知方法单独考虑标签信息所带来的益处。总之，第四节中定义的变量取值如下：

- T：任务：最优波束预测
- P：准确率（%）
- M：深度卷积网络，图7中的架构
- $\mathcal{D}$：来自ASU数据集的20个区域及相应的任务标签

有监督任务：波束预测任务（T）要求对信道应用一组波束成形器，并确定在用户端产生最高接收功率的波束成形器。在此场景中，我们采用包含32个单元的天线阵列，以及由32个波束组成的相应码本；每个波束都作为分类中的一个独立标签。对于这两个任务，我们采用相同的性能指标，即top-1准确率，该指标衡量模型正确预测出确切标签的实例所占的百分比。从数学上讲，top-1准确率 $A_{\text{top-1}}$ 定义为：

$$A_{\text{top-1}} = \frac{1}{N}\sum_{i=1}^{N}\delta(y_i, \hat{y}_i)$$

其中， N 是样本总数， y_i 是第 i个样本的真实标签，$\hat{y}_i$ 是第 i个样本的预测标签。请注意，此处 $\delta(y_i,\ \hat{y}_i)$表示指示函数，使得

$$\delta(y_i, \hat{y}_i) = \begin{cases} 1, & \text{if } y_i = \hat{y}_i \\ 0, & \text{if } y_i \neq \hat{y}_i \end{cases}$$

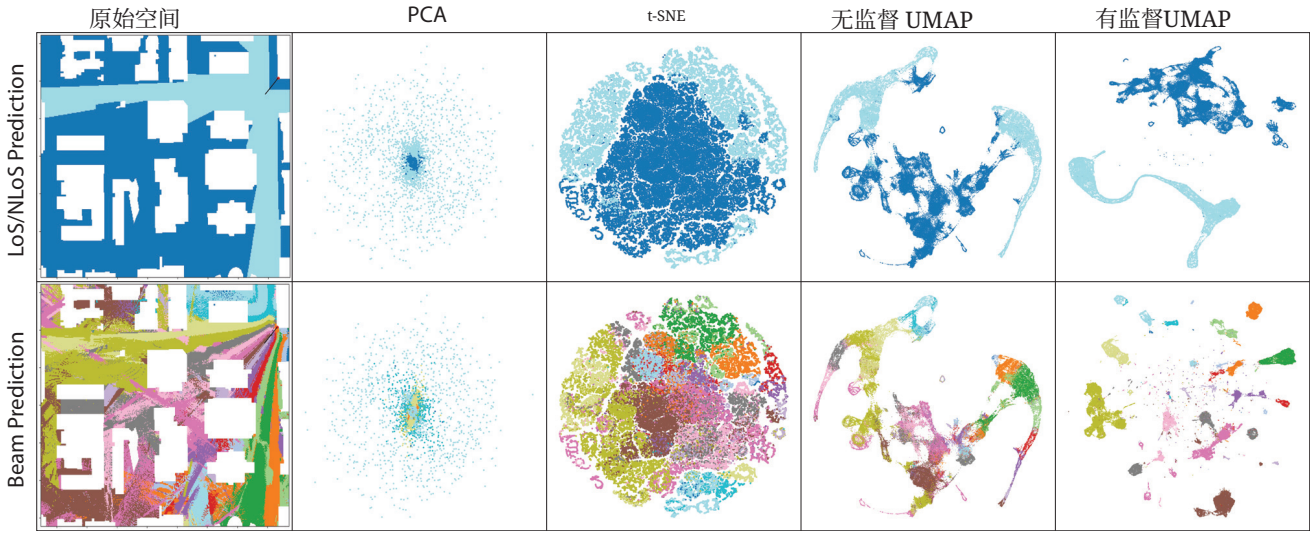

图 8. 潜在空间的比较（PC A、t-SNE、UMAP）与 XY/笛卡尔坐标中的原始空间。颜色表示不同的标签（视距状态或波束索引）。图中显示，UMAP 能够提供更清晰的分离效果，重叠更少，同时更忠实地保留全局结构。

表 IV 原始距离与监督任务中模型性能的相关性。

类别	距离名称	相关性	
		视距标识	波束预测
几何	成对欧氏距离	0.43	0.07
	聚类欧氏距离	0.42	0.12
	质心欧氏距离	0.32	0.06
	余弦距离	-0.05	-0.05
统计	詹森-香农距离	0.08	0.14
	海灵格距离	0.09	0.15
	全变差	0.07	0.14
	Wasserstein距离	**0.55**	0.23
	柯尔莫哥洛夫-斯米尔诺夫	0.50	0.42
	能量距离	**0.63**	0.33
	MMD（线性）	-0.10	-0.01
	最大均值差异（径向基函数）	-0.12	-0.01
子空间	格拉斯曼	-0.09	-0.08
	弦距离	-0.08	-0.07
	阿西莫夫	0.12	-0.06
其他	PAD	0.45	**0.61**
所提出的方法	标签感知欧几里得	0.18	**0.75**
	标签感知瓦瑟斯坦	**0.53**	0.11

图8最左侧的图展示了ASU数据集中每个标签的情况，以及每种降维方法如何改变标签空间。数据集的划分方式也与之前完全相同，因此，给定区域/数据集可能有多个不同的标签，具体取决于数据集所在的位置和特定任务。

卷积模型：为完成这些任务，我们使用了如图7所示的卷积神经网络（CNN）。该网络架构由多层二维卷积、批归一化、Leaky ReLU激活和最大池化组成。这些操作应用于输入数据的实部和虚部。最后几层采用全连接网络处理提取的特征，并输出与每项任务中标签数量相对应的 Softmax 归一化类别概率。

输入空间中的距离：在输入空间中，我们的结果表明，某些距离（例如 Proxy-A距离 和诸如 Wasserstein距离、柯尔莫哥洛夫-斯米尔诺夫距离和能量距离等积分概率度量，能够有效衡量数据集相似性。所提出的标签感知欧几里得距离表现良好，尤其是在标签不平衡程度较低的波束预测任务中。相比之下，标签感知 Wasserstein距离的表现不够理想，这可能归因于原始输入空间的高维度。此外，视距（LoS）识别任务面临着重大挑战，因为许多数据集仅包含一个标签。对于这一特定任务，我们仅展示每个标签至少有一个样本的数据集的结果，因为模型无法泛化到包含未见标签的数据集。标签感知方法通过引入基于标签不平衡程度的惩罚项来缓解这一问题。然而，这些方法仍会受到维度灾难的影响，而且由于惩罚项需要估算，标签感知方法也与具体任务和数据集绑定。

潜在空间中的距离：在考察降维方法的结果时，我们观察到，能够编码数据中全局关系的技术（例如 UMAP 和自编码器）平均而言表现更好。PCA 作为一种线性方法，无法捕捉这些全局结构，因此主要适用于欧几里得距离和 PAD 等线性距离。在偏好采用线性降维方法的情况下，标签感知欧氏距离优于其他指标。我们还发现，标签感知距离往往超过其他方法，或取得相近的性能。有趣的是，监督 UMAP 利用标签增强数据分离效果；在标签较少的情况下，其表现略优于 UMAP，而在标签较多的情况下则较差——这可能是因为过度强调标签信息，导致数据点之间的分离程度超过了它们在输入空间中的自然差异。总体而言，我们观察到，与欧几里得距离、Wasserstein距离以及相关性得分相比，标签感知方法能够提升性能，并且编码空间中的相关性得分高于原始空间。

监督任务的结论：结果表明

表 V 降维空间中的距离与监督任务模型性能的相关性。

Task	距离	降维方法				
		主成分分析 32	t-SNE 2	UMAP 2	自编码器 32	SUP UMAP 2
LoS ID	成对欧氏距离	0.43	0.38	0.46	0.41	**0.73**
	聚类欧氏距离	0.40	0.45	0.51	0.49	**0.63**
	质心欧氏距离	0.35	0.32	0.47	0.47	**0.68**
	KL散度	0.23	0.35	0.17	0.51	0.21
	Wasserstein距离	0.50	0.43	0.45	0.47	**0.71**
	PAD	0.44	0.38	0.37	0.34	0.36
	标签感知欧氏距离	0.10	0.22	**0.78**	0.40	**0.80**
	标签感知 Wasserstein 距离	0.28	0.25	**0.75**	0.45	**0.81**
Beam Prediction	成对欧氏距离	0.07	0.62	0.73	**0.76**	0.63
	聚类欧氏距离	0.12	0.56	0.71	**0.75**	0.56
	质心欧氏距离	0.05	0.70	**0.79**	**0.79**	0.72
	KL散度	0.25	0.60	0.50	0.72	0.74
	Wasserstein距离	0.15	0.72	**0.76**	**0.79**	0.72
	PAD	**0.73**	0.66	0.67	0.63	0.70
	标签感知欧氏距离	**0.74**	**0.78**	**0.79**	**0.81**	**0.81**
	标签感知 Wasserstein 距离。	0.19	**0.77**	**0.80**	**0.79**	**0.81**

标签感知距离是有效的。我们之前在 UMAP 编码空间中使用 Wasserstein 距离或欧几里得距离的方法，在无监督情形下仍然很有前景。不过，有效利用标签信息时，标签感知距离更有潜力带来显著提升。尽管这些方法在我们的实验中表现良好，我们也承认，利用标签信息的算法仍有改进空间。尤其是，惩罚项过于依赖具体问题，目前还无法充分推广。尽管如此，对于标签较多的任务，我们建议使用标签感知欧几里得距离；对于标签较少的情况，则建议使用 Wasserstein距离。如果难以确定，则可选择欧几里得距离，因为它在大多数情况下仍表现良好，而且复杂度显著较低。

X. 结论

在本研究中，我们提出了一个用于无线通信数据集距离化的综合框架，建立了数据集距离与模型准确率之间首批直接相关性之一。通过利用由主成分分析、t-SNE、UMAP 和自编码器等降维技术得到的潜在空间，我们有效捕捉了数据的内在结构，从而能够在无监督和有监督任务中稳健地计算距离。我们进一步提出了一种新颖的标签感知数据集距离度量，将标签信息纳入距离计算过程，从而提供了更稳健的相似度衡量方式。大量评估结果表明，我们所提出的距离度量优于原始空间中的传统方法，以及各种降维技术，尤其是在视距识别和波束预测等任务中。这一理解和量化数据集相似性的基础性举措，有助于改进数据选择、减少不必要的模型再训练，并促进机器学习模型在无线系统中的高效部署。未来工作将在这些发现的基础上，拓展我们的方法以覆盖更广泛的无线任务，将其应用于真实世界数据集，并与先进模型集成，进一步推动该领域的发展。

XI. 未来工作

未来工作应解决两项局限：（i）许多距离计算都需要重新训练嵌入模型，这妨碍了在线使用；（ii）UMAP 式映射通常需要访问完整数据集，且难以低成本扩展。一个自然的方向是通过以小批量目标函数训练的**可扩展、可复用的嵌入空间**，借助**参数化神经映射**实现流式/持续更新。在监督设置中，标签感知编码器可以塑造潜在空间，使其更好地反映可迁移性。大型无线模型（LWM） [68]提供了一条减少逐任务重新训练需求的途径，即提供通用嵌入：只需将新样本嵌入一次，并在共享空间中计算距离，即可比较数据集；必要时只需进行轻量校准。此外，本文针对有限任务集评估的方法还可扩展并改进，以适用于更广泛的无线任务，包括混合输入任务，例如将信噪比（SNR）与信道数据结合，或纳入角度和到达时间。将我们的技术应用于真实世界数据也至关重要，特别是实现基于射线追踪的场地特定数据集生成，使其能够贴近实际部署场景。另一个有前景的方向是为生成模型 [69], 创建距离感知损失函数，以提升生成数据的质量和相关性。此外，我们的方法还可扩展至涉及离散或类别型输入变量的任务。归根结底，应重点实现端到端应用，例如数据生成、数据集排序、数据集压缩，以及异常值或数据集偏移检测。随着这些进展，我们期望能够更有规划地使用数据集，从而推动更优的数据驱动方法。

参考文献

[1] C.-K. Wen、W.-T. Shih 和 S. Jin，“大规模 MIMO CSI 反馈的深度学习”，*IEEEWireless Communications Letters*，第 7 卷，第 5 期，第 748–751 页，2018 年。[2] A. Taha、M. Alrabeiah 和 A. Alkhateeb，“利用压缩感知和深度学习实现大型智能表面”，*IEEE Access*，第 9 卷，第 44 304–44 321 页，2021 年。[3] A. Alkhateeb、S. Alex、P. Varkey、Y. Li、Q. Qu 和 D. Tujkovic，“面向高移动性毫米波系统的深度学习协同波束成形”，*IEEE Access*，第 6 卷，第 37 328–37 348 页，2018 年。[4] M. Alrabeiah 和 A. Alkhateeb，“利用 6 GHz 以下信道进行毫米波波束与阻塞预测的深度学习”，*IEEE Transactions on Communications*，第 68 卷，第 9 期，第 5504–5518 页，2020 年。[5] F. B. Mismar、B. L. Evans 和 A. Alkhateeb，“面向 5G 网络的深度强化学习：联合波束成形、功率控制与干扰协调”，*IEEE Transactions on Communications*，第 68 卷，第 3 期，第 1581–1592 页，2020 年。[6] S. Parthasarathy 和 M. Ogihara，“利用数据集相似性进行分布式挖掘”，载于 并行与分布式处理，J. Rolim 编，柏林、海德堡：Springer Berlin Heidelberg，2000 年，第 399–406 页。[7] J. Breen、A. Buffmire、J. Duerig、K. Dutt、E. Eide、M. Hibler、D. Johnson、S. K. Kasera、E. Lewis、D. Maas、A. Orange、N. Patwari、D. Reading、R. Ricci、D. Schurig、L. B. Stoller、J. Van der Merwe、K. Webb 和 G. Wong，“POWDER：开放式无线数据驱动实验研究平台”，载于 第 *14* 届国际无线网络测试平台、实验评估与特征化研讨会论文集，系列：WiNTECH ’20。纽约，纽约州，美国：计算机协会，2020 年，第 17–24 页。 [在线]。可访问：https://doi.org/ 10 1145 3411276 3412204 。/。I. Guvenc、M. [8] Sichitiu、R. Dutta、O. Ozdemir 和 R. Gyurek，“AERPAW：无线与无人机研究国家级设施”，*IEEE Communications Technology News*，2023年，
h
t
t
p
s://
w
w
w.comsoc.org/publications/ctn/aerpaw-national-facility-wireless-and-drone-research。

[9] A. Alkhateeb、G. Charan、T. Osman、A. Hredzak、J. Morais、U. Demirhan 和 N. Srinivas，“Deepsense 6G：大规模真实世界多模态感知与通信数据集”，IEEE Communications Magazine，2023年。[10] T. S. Rappaport、S. Sun、R. Mayzus、H. Zhao、Y. Azar、K. Wang、G. N. Wong、J. K. Schulz、M. Samimi 和 F. Gutierrez，“面向 5G蜂窝通信的毫米波移动通信：它将奏效！”，IEEE Access，第 1 卷，第 335–349 页，2013 年。[11] Remcom，“Wireless InSite 射线追踪软件”，https://www.remcom.com/wireless-insite-em-propagation-software，访问日期：2024-10-06。[12] J. Hoydis、F. A. Aoudia、S. Cammerer、M. Nimier-David、N. Binder、G. Marcus 和 A. Keller，“Sionna RT：用于无线电传播建模的可微射线追踪”，2023 年。 [在线]。可访问：https://arxiv.org/abs/2303.11103。A. Alkhateeb，“[13] DeepMIMO：面向毫米波和大规模 MIMO 应用的通用深度学习数据集”，载于信息论与应用研讨会（ITA）论文集，美国加利福尼亚州圣迭戈，2019 年 2 月，第 1–8 页。[14] I. Siegert、R. B¨ock 和 A. Wendemuth，“利用基于主成分分析的数据集相似度度量改进跨语料库情感识别”，Computer Speech & Language，第 51 卷，第 1–23 页，2018 年。[15] A. Achara 和 R. K. Pandey，“揭示潜在模式：研究数据集相似性、性能与泛化能力”，Neurocomputing，第 573 卷，第 127205 页，2024 年。[16] S. Ben-David、J. Blitzer、K. Crammer 和 F. Pereira，“领域自适应的表征分析”，载于《神经信息处理系统进展》，B. Sch¨olkopf、J. Platt 和 T. Hoffman 编，第 19 卷，MIT Press，2006 年。[17] A. Farahani、S. Voghoei、K. Rasheed 和 H. R. Arabnia，“领域自适应简述”，载于《数据科学与信息工程进展》，R. Stahlbock、G. M. Weiss、M. Abou-Nasr、C.-Y. Yang、H. R. Arabnia 和 L. Deligiannidis 编，Cham：Springer International Publishing，2021 年，第 877–894 页。[18] M. Wang 和 W. Deng，“深度视觉领域自适应：综述”，Neurocomputing，第 312 卷，第 135–153 页，2018 年。[19] E. Tzeng、J. Hoffman、K. Saenko 和 T. Darrell，“对抗判别领域自适应”，载于 IEEE 计算机视觉与模式识别会议论文集（CVPR），2017 年 7 月。[20] A. Chadha 和 Y. Andreopoulos，“对抗判别领域自适应的改进技术”，IEEE Transactions on Image Processing，第 29 卷，第 2622–2637 页，2020 年。[21] W. M. Kouw 和 M. Loog，“无目标标签领域自适应综述”，IEEE Transactions on Pattern Analysis and Machine Intelligence，第 43 卷，第 3 期，第 766–785 页，2021 年。[22] J. Shi、M. Sha 和 X. Peng，“通过基于深度学习的领域自适应将无线网状网络配置从仿真迁移到现实”，载于第 18 届 USENIX网络系统设计与实现研讨会（NSDI 21），USENIX Association，2021 年 4 月，第 887–901 页。[23] Y. Yuan、G. Zheng、K.-K. Wong、B. Ottersten 和 Z.-Q. Luo，“基于迁移学习和元学习的快速下行链路波束成形自适应”，《IEEE无线通信汇刊》，第 20 卷，第 3 期，第 1742–1755 页，2021 年。[24] H. Zou、J. Yang、Y. Zhou、L. Xie 和 C. J. Spanos，“通过无监督对抗领域自适应实现稳健的 WiFi 设备无关手势识别”，载于 2018 年第 27 届国际计算机通信与网络会议（ICCCN），2018 年，第 1–8 页。[25] P. O. Pinheiro，“利用相似性学习进行无监督领域自适应”，2018 年。[26] Y. Fu、Y. Wei、G. Wang、Y. Zhou、H. Shi 和 T. S. Huang，“自相似分组：一种简单的无监督跨领域行人重识别自适应方法”，载于 IEEE/CVF 国际计算机视觉会议论文集（ICCV），2019 年 10 月。[27] M. Peng、Z. Li 和 X. Juan，“基于相似性的领域自适应网络”，Neurocomputing，第 493 卷，第 462–473 页，2022 年。[28] L. Wang、X. Wang 和 J. Feng，“子空间距离分析及其在基于自适应贝叶斯算法的人脸识别中的应用”，《模式识别》，第 39 卷，第 3 期，第 456–464 页，2006 年。[29] T. van Erven 和 P. Harremos，“R´enyi 散度与 Kullback-Leibler 散度”，IEEE Transactions on Information Theory，第 60 卷，第 7 期，第 3797–3820 页，2014 年。[30] J. M. L. McInnes、J. Healy，“UMAP：用于降维的统一流形近似与投影”，arXiv 预印本 arXiv:1802.03426，2018 年。[31] L. Van der Maaten 和 G. Hinton，“使用 t-SNE 可视化数据”，Journal of Machine Learning Research，2008 年。

[32] P. Doll´ar、V. Rabaud 和 S. Belongie，《非等距流形学习：分析与算法》，ICML ’07：第24届国际机器学习会议论文集，第241–248页，2007年。[33] L. Liberti、C. Lavor、N. Maculan 和 A. Mucherino，《欧几里得距离几何及其应用》，2012年。[34] P.-E. Danielsson，《欧几里得距离映射》，《计算机图形学与图像处理》，第14卷，第3期，第227–248页，1980年。[35] R. De Maesschalck、D. Jouan-Rimbaud 和 D. Massart，《马哈拉诺比斯距离》，化学计量学与智能实验室系统，第50卷，第1期，第1–18页，2000年。[36] S. Kullback 和 R. A. Leibler，《论信息与充分性》，《数学统计年刊》，第22卷，第1期，第79–86页，1951年。[37] T. Narayana，《信息论与统计学》，作者 Solomon Kullback。Wiley，纽约，1959年。xvii + 395 页。12.50。加拿大数学公报，第4卷，第1期，第85–86页，1961年。[38] V. M. Panaretos 和 Y. Zemel，《Wasserstein距离的统计学方面》，《统计学及其应用年度评论》，第6卷，第1期，第405–431页，2019年3月。[39] G. K. Dziugaite、D. M. Roy 和 Z. Ghahramani，《通过最大均值差异优化训练生成式神经网络》，2015年。[40] J. Hamm 和 D. D. Lee，《格拉斯曼判别分析：基于子空间的学习的统一视角》，ICML ’08：第25届国际机器学习会议论文集，第376–383页，2008年。[41] K. Ye 和 L.-H. Lim，《舒伯特簇与不同维数子空间之间的距离》，2016年。[42] D. Hilbert，《欧几里得几何基础》，Springer，1899年。[43] A. Sarmiento、I. Fond´on、I. Dur´an-D´ıaz 和 S. Cruces，《基于质心的聚类与 $\alpha\beta$-散度》，《熵》，第21卷，第2期，2019年。[44] M. Ahmed、R. Seraj 和 S. M. S. Islam，《K均值算法：综合综述与性能评估》，《电子学》，第9卷，第8期，2020年。[45] P. Xia、L. Zhang 和 F. Li，《利用余弦相似度集成学习相似度》，《信息科学》，第307卷，第39–52页，2015年。[46] S. Kullback 和 R. Leibler，《论信息与充分性》，《数学统计年刊》，1951年。[47] B. Fuglede 和 F. Topsoe，《詹森-香农散度与希尔伯特空间嵌入》，载于国际信息论研讨会，2004年。ISIT 2004。论文集，2004年，第31–页。[48] R. Beran，《参数模型的最小海灵格距离估计》，《统计年刊》，第5卷，第3期，第445–463页，1977年。[49] F. J. M. Jr.，《用于拟合优度检验的柯尔莫哥洛夫-斯米尔诺夫检验》，《美国统计学会杂志》，第46卷，78，1951 253，68期，第–页。[50] A. Gritsenko、T. Salimans、R. van den Berg、J. Snoek 和 N. Kalch[50] brenner，《用于并行语音合成的谱能量距离》，载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin 编，第33卷。Curran Associates, Inc.，2020年，第13 062–13 072页。[51] S. Verd´u，《全变差距离与相对信息分布》，载于2014年信息论与应用研讨会（ITA），2014年，第1–3页。[52] T. Bendokat、R. Zimmermann 和 P.-A. Absil，《格拉斯曼流形手册：基础几何与计算方面》，《计算数学进展》，第50卷，第1期，2024年1月。[53] N. Mankovich 和 T. Birdal，《旗流形上的弦平均及其应用》，2023年。[54] F. Li、L. Lai 和 S. Cui，《关于子空间学习的对抗鲁棒性》，《IEEE信号处理汇刊》，第68卷，第1470–1483页，2020年。[55] A. Ma´ckiewicz 和 W. Ratajczak，《主成分分析（PCA）》，《计算机与地球科学》，第19卷，第3期，第303–342页，1993年。[56] I. Gulrajani 和 D. Lopez-Paz，《寻找失落的领域泛化》，CoRR，卷abs/2007.01434，2020年。[57] M. A. Alvarez、L. Rosasco 和 N. D. Lawrence，《向量值函数的核：综述》，2012年。[58] S.-H. Cha 和 S. N. Srihari，《论直方图之间距离的度量》，《模式识别》，第35卷，第6期，第1355–1370页，2002年。[59] J. Guo、C.-K. Wen、S. Jin 和 G. Y. Li，《基于卷积神经网络的用于大规模MIMO CSI反馈的多速率压缩感知：设计、仿真与分析》，《IEEE无线通信汇刊》，第19卷，第4期，第2827–2840页，2020年。[60] S. Dasgupta 和 A. Gupta，《约翰逊-林登施特劳斯定理的初等证明》，《随机结构与算法》，第22卷，2003年。[61] A. M. Ikotun、A. E. Ezugwu、L. Abualigah、B. Abuhaija 和 J. Hem[61] ing，《K均值聚类算法：综合综述、变体分析及大数据时代的进展》，《信息科学》，第622卷

，第 178–210 页，2023年。

[62] V. Cohen-Addad、V. Kanade、F. Mallmann-Trenn 和 C. Mathieu，《层次聚类：目标函数与算法》，2017年。[63] M. Ester、H.-P. Kriegel、J. Sander 和 X. Xu，《一种用于在含噪声的大型空间数据库中发现簇的基于密度的算法》，载于第二届知识发现与数据挖掘国际会议论文集，丛书 KDD'96。AAAI Press，1996年，第226–231页。[64] A. Altmann、L. Tolosi、O. Sander 和 T. Lengauer，《置换重要性：一种经过校正的特征重要性度量》，《生物信息学》（牛津版），第26卷，第1340–7页，2010年4月。[65] C. Molnar、T. Freiesleben、G. K¨onig、J. Herbinger、T. Reisinger、G. Casalicchio、M. N. Wright 和 B. Bischl，关联部分依赖图和针对数据生成过程的置换特征重要性。Springer Nature Switzerland，2023年，第456–479页。[66] M. T. Ribeiro、S. Singh 和 C. Guestrin，“为什么我应该相信你？”：解释任何分类器的预测，2016年。[67] S. Lundberg 和 S.-I. Lee，“一种统一的模型预测解释方法”，2017年。[68] S. Alikhani、G. Charan 和 A. Alkhateeb，“大型无线模型（LWM）：面向无线信道的基础模型”，2025年。 [在线]。网址：https://arxiv.org/abs/ 2411 08872 。[69] A. C¸elik 和 A. Eltawil，“生成式人工智能时代的黎明：6G无线智能新前沿教程与综述”，*IEEE*通信学会开放期刊，2024年1月。