arXiv:2602.02739v1 [cs.LG] 2 Feb 2026

TopoPrune：通过统一潜在空间拓扑实现的鲁棒数据剪枝 y

Arjun Roy [1] **Prajna G. Malettira** [1] **Manish Nagaraj** [1] **Kaushik Roy** [1]

摘要

几何数据剪枝方法虽然实用，能够充分利用预训练模型，但本质上并不稳定。由于依赖外在几何结构，这类方法对潜在空间中的扰动高度敏感，导致其在跨架构迁移或存在特征噪声时性能下降。我们提出 TopoPrune，这是一种利用拓扑捕捉数据稳定内在结构的框架，旨在解决这一挑战。TopoPrune 在两个尺度上运行：(1) 利用拓扑感知流形近似为数据集构建全局低维嵌入；随后，(2) 采用可微持久同调对流形嵌入进行局部拓扑优化，并依据样本的结构复杂度对其排序。我们证明，统一的双尺度拓扑方法能够确保较高的准确率和精确率，尤其是在数据集剪枝率较高时（例如 90%）。此外，得益于拓扑固有的稳定性，TopoPrune (a) 对潜在特征嵌入中的噪声扰动具有极强的鲁棒性，(b) 在不同网络架构之间展现出更强的迁移能力。本研究展示了一条颇具前景的路径，有望构建稳定且有原则的基于拓扑的框架，实现鲁棒的数据高效学习。

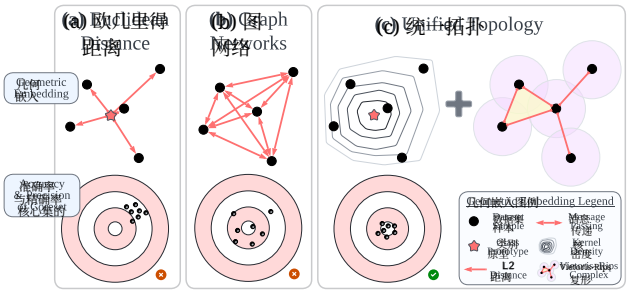


图*1*。 **拓扑数据选择可生成性能更高且更稳定的核心集。** 多次运行中的核心集性能揭示了常用方法的局限性。**(a)**基于欧几里得距离的选择较为稳定，但准确率较低。**(b)** 基于图的方法可实现更高的准确率，但波动性很大。**(c)** 我们提出的拓扑方法兼具高准确率和稳定性。

1. 引言

由于当代模型和数据集的规模不断扩大，训练现代深度学习系统所需的计算资源急剧增加。这种增长使训练和微调在计算上难以承受，因此亟需数据高效学习策略。数据剪枝便是其中一种策略，旨在从大型数据集中抽取一个较小且具有代表性的子集（或核心集），以保留完整数据集的基本学习特征。由此既能实现快速模型训练、高效微调并降低存储成本，又能最大限度地减少最终模型性能的下降。

总体而言，核心集选择方法可分为三大类。基于优化的方法会选择一个核心集，使其损失景观（Killamsetty 等人，2021b）或梯度动态（Mirza- soleiman 等人，2019；Killamsetty 等人，2021a；Tan 等人，2023）与整个数据集的情况最为一致，从而确保在该子集上训练的模型具有可比的泛化能力。尽管这类方法通常有效，但其实际应用受到显著限制，例如依赖计算成本高昂的二阶优化（Pooladzandi 等人，2022）或双层优化（Borsos 等人，2020）。基于分数的方法依据模型预测分数对训练样本进行排序和选择。这些分数可以来自训练动态（Toneva 等人，2019；Garg & Roy，2023；Zheng 等，2025）和不确定性估计（Paul 等人，2021；何恺明 等人，2023；2024；Cho 等人，2025）。然而，这些分数本质上依赖于模型和训练过程，反映的是特定网络在训练特定阶段所掌握的知识。因此，基于优化的方法和基于分数的方法不仅成本高昂——因为它们需要从头训练一段时间——而且还无法兼容规模庞大且不断发展的公开预训练模型生态系统，因为在这些模型中，通常只能获取最终权重。

[1]电气与计算机工程系，普渡大学，美国。通讯作者：Arjun Roy<roy208@purdue.edu>。

预印本。*2026*年*2*月*4*日。

为克服这一限制，基于几何学的核心集选择方法可以利用预训练模型生成的静态特征嵌入。这一领域的方法包括：基于倒数第二层特征嵌入空间表示样本（Xia等人，2023）；通过最优传输（Xiao 等人，2024）或 Wasserstein 距离（Xiong 等人，2024）衡量分布相似性；以及利用结合决策边界信息的样本几何重构误差（Yang 等人，2024）。尽管这些方法避免了代价高昂的训练分析，但一个显著限制是它们依赖于对特征嵌入空间的外在几何敏感的度量，我们将这种脆弱性称为“几何脆弱性”（Papillon 等人，2025）。这种脆弱性带来两个主要缺点：（1）倾向于优先选择密集区域中的样本，而牺牲分布稀疏尾部中具有信息量的样本（Zheng 等人，2023）；（2）在不同网络架构之间，或向嵌入中引入噪声时，性能不稳定。这一点在欧几里得距离度量（Xia 等人，2023）或消息传递图方法（Maharana 等人，2024；Xie 等，2025）中尤为明显，因为它们对特征嵌入空间的变化高度敏感（见图 1）。

在这项工作中，我们提出了 TopoPrune，这是一种新颖的框架，通过利用拓扑分析（Seifert 与 Threlfall，1980）解决几何脆弱性这一难题。对于不熟悉拓扑学的读者来说，拓扑学是几何学的一个分支，研究在拉伸和弯曲等连续变形下保持不变、但在撕裂时会改变的空间性质。一个经典例子是，从拓扑学的角度看，咖啡杯与甜甜圈是等价的，因为二者可以相互变形，同时保留它们共有的那个孔。我们关注这种稳定的内在结构（孔），而非短暂的外在几何度量（如距离或曲率），从而能够以稳定的拓扑度量分析数据集的特征嵌入。因此，面对由噪声特征或不同架构所产生的不同特征嵌入（Cohen-Steiner 等人，2005；Suresh 等人，2024）导致的特征嵌入空间中的轻微扰动，TopoPrune 仍能保持出色的稳定性。这使得我们能够利用代理模型（Coleman 等人，2020），或直接利用大量预训练模型和基础模型生成核心集，而无需从头开始重新训练特定模型。

我们的框架首先通过**拓扑感知流形近似**建立全局结构(McInnes 等人，2018a；Wang 等人，2021)，将高维特征投影到标准化的低维流形嵌入中。尽管这种全局结构能够将相似样本归为一组，但无法区分局部区域内应优先考虑哪些样本。现有方法通常在局部区域内采用随机采样（Zheng 等人，2023年）或使用消息传递等几何启发式方法（Maharana 等人，2024年）。为更好地以**局部结构**补充这一全局视角，我们接着采用可微持久同调(Scoccola等人，2024年；Carrière 等人，2024年；Mukherjee 等人，2024年)来评估样本相对于其直接邻居的结构相关性。持久同调追踪多个尺度（同调群的滤过）下拓扑结构的“诞生”与“消亡”（持久性）。在我们的应用中，我们进行优化，以最大化由流形投影 *Vietoris–Rips* 复形的滤过构造的局部拓扑特征的持久性（诞生与消亡时间）（Loiseaux等人，2023a）。这一过程通过迭代重新定位样本，使其达到最优配置，从而消除拓扑歧义并增强拓扑稳定性，直接衡量样本对其局部邻域结构复杂度的贡献。最后，将统一分数中的全局密度（流形嵌入的密度）与局部持久性相结合，以兼顾全局和局部拓扑结构。

我们的方法有以下贡献：

- 我们提出了 TopoPrune，这是一种新颖的核心集选择框架，通过双尺度拓扑分析来定义样本重要性。它将全局流形投影 与基于可微持久同调优化得出的局部持久性 分数相结合，从而比以往的核心集方法更准确、更稳定地识别出在结构上至关重要的样本。

- 我们证明，TopoPrune 具备稳健的跨架构迁移能力，无论迁移方向如何，都能持续生成高质量核心集：既适用于利用多样化代理嵌入（例如从 ResNet 迁移到 ViT）来训练固定目标模型，也适用于使用单一代理模型训练一组多样化的目标模型。这种稳定性使我们能够使用小型代理模型或现成的预训练模型来生成真正模型无关的核心集，而无需付出高昂的重新训练成本。

- 我们通过广泛的实证验证表明，TopoPrune 的表现显著优于当前最先进的几何方法、梯度方法和基于分数的方法。我们的框架生成的核心集具有更高的准确率和精确率，并且对噪声特征嵌入的鲁棒性显著更强，尤其是在数据剪枝率较高时。

通过依据拓扑稳定且内在的性质来定义样本重要性，TopoPrune 超越了脆弱的几何度量，构建出真正稳健的核心集框架。

2. 背景与相关工作

2.1. 双尺度拓扑

尽管许多现代拓扑算法本身能够同时对数据的全局和局部结构进行建模，但我们的工作将这两个概念解耦为两个不同阶段。本文将全局拓扑定义为整个数据集的流形结构，并通过低维嵌入来表征。接着，我们将局部拓扑定义为样本与其近邻交互所形成的精细结构，并使用持久同调进行分析。

低维流形近似。 高维分析中的一个关键步骤是创建低维数据表示。主成分分析（PCA）等线性方法（Pearson，1901年）效率很高，但只能保留全局方差，无法捕捉复杂的非线性结构。相比之下，t-分布随机邻居嵌入（t-SNE）等非线性技术（van der Maaten 和 Hinton，2008年）擅长保留细粒度的局部邻域，但这种关注往往会扭曲数据的整体全局结构。较新的技术则直接利用拓扑原理，创建更忠实的流形嵌入。统一流形近似与投影（UMAP）（McInnes 等人，2018a）和 PaCMAP（Wang 等人，2021）等先进方法将高维数据建模为模糊拓扑结构，以同时保留细粒度的局部连通性和大尺度的全局关系。

虽然基于学习的方法也能实现强大的全局拓扑保持，例如拓扑自编码器（TopoAE）（Moor 等人，2020年）和正则化拓扑数据分析（RTD）（Trofimov 等人，2023年），但它们需要耗费大量计算资源进行训练，而且在复杂数据集上往往无法实现鲜明的类别分离（详见B.1.2 节）。因此，鉴于 UMAP 等算法方法效率更高，也更符合我们轻量、免训练的目标，我们优先采用这类方法。这些算法方法已被证明在多个领域中，都能有效地解释深度模型学到的复杂表征，应用领域从单细胞基因组学（Becht 等人，2018）到字典学习中的聚类（Fel 等人，2023年；2024年）。正如（de Bodt 等人，2025年）所示，这些技术生成的嵌入具有紧凑且定义清晰的簇，从而提升下游分析的效果。尽管这类嵌入的定量保真度仍是当前研究的热点（Jeon 等人，2025年），但我们认为它们非常适合我们的任务。通过明确保留高维空间中的最近邻结构，UMAP 和 PaCMAP 等方法能够从本质上维持数据流形的密度景观。这直接保留了“典型性”（高密度区域中的样本）和“非典型性——”（稀疏区域中的样本），使嵌入成为后续样本重要性评分的可靠基础（有关定性解释，参见A.3 节）。

样本与其最近邻之间的交互。理解样本与其邻居之间的局部交互，对于确定样本的重要性至关重要。一种常见方法是使用图神经网络（图神经网络），在通常由最近邻关系定义的图中传播节点间的信息。在图神经网络中，样本的重要性通过学习到的消息传递来量化，该过程会聚合其局部邻域中的特征（Maharana等人，2024年）。其他方法则将图级结构熵与 Shapley 值和蓝噪声采样相结合（Chen等人，2014），以选出多样化的核心集（Xie等人，2025年）。然而，这些方法基于单一的固定图进行操作，且可能对构图时采用的几何超参数较为敏感。

相比之下，持久同调（Botnan 和 Lesnick，2022）提供了一种从根本上不同且更稳健的框架。它不分析单一图，而是研究单纯复形多尺度滤过过程中高阶拓扑结构（例如连通分支、环和空洞）的演变。这能够同时完整概括数据在所有尺度上的形状。持久同调的一项关键优势是其已得到证实的稳定性（Cohen-Steiner等人，2005年）。可以保证，数据集的持久图只会因输入数据的微小扰动而发生轻微变化，因此是局部结构的稳健描述符（Turkes等人，2022；Mishra 和 Motta，2023年）。在机器学习应用中，人们广泛使用这些稳健的几何描述符来理解特征嵌入，例如监测网络在训练过程中的泛化能力（Birdal等人，2021），以及探索潜在嵌入在网络各层中的拓扑结构（Naitzat等人，2020年）。有关单纯复形和持久同调底层构造的详细概述，请参阅部分 A.1。

虽然传统持久同调提供了强大的描述工具，但由于其本身不可微，因此在现代深度学习流程中的应用受到限制。可微持久同调领域的最新进展克服了这一障碍，使梯度能够从持久图反向传播至输入数据点的坐标（Carrière等人，2024年；Mukherjee等人，2024年）。这使得我们能够在基于梯度的框架中直接优化数据的拓扑特征。 (Scoccola等人，2024年)的研究为这些梯度提供了快速且稳定的计算框架，即使面对表达能力更强的多参数持久同调也不例外。借助这一方法，我们不再只是描述静态的局部拓扑，而是能够通过优化主动增强它。

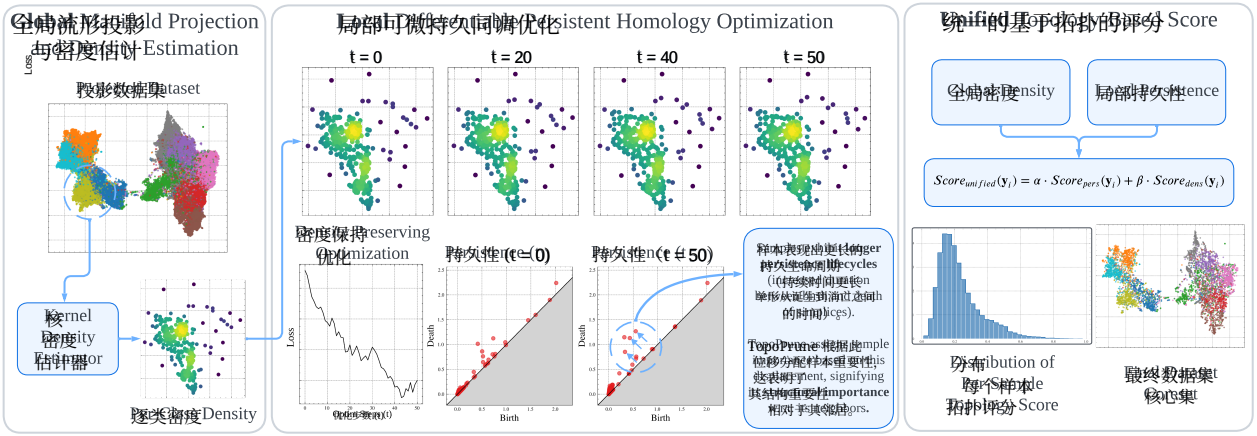


图2. **TopoPrune 概览。（左）**拓扑感知投影可视化了全局数据流形。**（中）**在每个类别中，通过保密度持久同调优化，为每个样本计算局部持久性分数。颜色映射表示从高（黄色）到低（蓝色）的密度。**（右）**通过对结合全局密度和局部持久性的统一分数进行分层抽样，构建最终核心集。这不仅优先选择拓扑信息最丰富的样本，还忠实地表示了原始数据集的密度分布。

3. 方法

我们提出的方法TopoPrune通过分析数据在两个不同尺度上的拓扑结构来构建核心集。（1）全局流形嵌入将原始高维嵌入投影到标准化的低维空间中，从而确保能够稳定地全局观察数据的整体结构。（2）局部拓扑交互采用可微多参数持久同调，探查由样本及其最近邻形成的局部结构。结合这两个拓扑尺度，我们根据全局密度和局部持久同调，为每个样本计算重要性分数，从而形成统一的拓扑度量，用于筛选单个样本（见图 2）。

3.1. 全局结构：通过拓扑流形嵌入实现数据集表示

给定一个训练良好的深度模型，记为 $f(\cdot)$，我们可以将其表示为特征提取器 $h(\cdot)$ 与分类器 $g(\cdot)$ 的复合，即 $f(\cdot) = g(h(\cdot))$。其中，$h(\cdot)$ 表示截至倒数第二层的网络，它将输入数据点 x 映射为高维特征嵌入 $\mathbf{z} = h(\mathbf{x}) \in \mathbb{R}^D$。因此，完整数据集$\mathcal{D} = \{(\mathbf{x}1, y_1) \ldots, (\mathbf{x}_N, y_N)\}$可以转换为高维特征集 $Z = \{\mathbf{z}1, \ldots, \mathbf{z}_N\}$。尽管这个高维空间 Z 包含丰富的语义信息，但其外在几何结构往往复杂且依赖于网络架构。为获得稳定且标准化的表示，我们使用基于拓扑的流形近似与投影技术（McInnes 等人，2018a；Wang 等人，2021）将 Z 投影到低维流形上。

该过程包含两个主要阶段。首先，将高维数据构建为模糊单纯集的拓扑表示。该结构通过为每个点与其邻居之间可能存在的连接赋予隶属强度（p_{ij}），从而捕捉数据的形状；其中，“模糊”表示我们相信某个单纯形存在于真实的底层流形中。随后，学习一个低维嵌入 $Y = \{\mathbf{y}1, \ldots, \mathbf{y}_N\}$，其中 $\mathbf{y}_i \in \mathbb{R}^d$ 和 $d \ll D$，并以类似方式定义其自身的模糊单纯集（q_{ij}）。最终的低维表示 Y通过优化点 $\{\mathbf{y}_i\}$ 的位置得到，从而使高维（p_{ij}）与低维（q_{ij}）成对相似度之间的交叉熵损失最小：

$$\mathcal{L}_{\text{proj}}(Y) = \sum_{i \neq j} \left[p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + (1 - p_{ij}) \log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right) \right] \tag{1}$$

该过程生成了一个标准化的流形嵌入，同时保留数据的内在形状。通过对B.1 节中介绍的不同流形近似与投影技术进行详细研究，我们采用 UMAP（McInnes 等人，2018a），因为它能够在不同网络架构之间生成更加一致的流形嵌入（参见C.4 节，其中探讨了 UMAP超参数）。在这个低维流形上，我们使用核密度估计器（KDE）为每个样本计算一个**密度分数**，以捕捉其全局代表性。该方法通过累加邻近点的贡献来估计每个样本$\mathbf{y}_i$ 的概率密度 $\sum_{j=1}^{n} \mathbf{y}^j$:

$$\text{Score}_{\text{dens}}(\mathbf{y}_i) = \frac{1}{Nh} \sum_{j=1}^{N} K\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{h}\right) \tag{2}$$

其中，N 表示样本数量，K 表示高斯核，h 表示带宽参数。该分数使我们能够将在流形高密度（原型）区域中的样本与低密度（非典型）区域中的样本区分开来。

3.2. 局部结构：使用基于持久性的优化器优化样本邻域

全局流形嵌入提供了一种低维表示，能够忠实地保留数据流形的全局结构。尽管这确保了稳定的高层表示，但仅从全局视角出发不足以识别信息量最大的样本，因为这通常取决于样本与其最近邻之间复杂的局部交互。为捕捉这种细粒度结构，我们利用持久同调，将其视为动态拓扑优化过程，而非静态描述符。该过程的目标是迭代调整类别流形中每个点的位置，以最大化持久生命周期（延长单纯复形的诞生与消亡之间的持续时间）。此过程针对每个类别独立进行 $c \in \{1, \ldots, C\}$，以分析特定的类内结构。对于每个类别，我们从全局流形嵌入中获取其点云 $Y_c = \{\mathbf{y}_i \mid \text{标签}(\mathbf{y}_i) = c\}$，并在 Oudot，2015）的基础上构造 Vietoris-Rips 过滤，其计算可扩展性优于 Alpha 和 ˇ Cech 复形（Otter 等人，2017；Mishra 和 Motta，2023年）。

与（Scoccola 等，2024）的工作类似，我们定义一个可微损失函数，$\mathcal{L}_{\text{pers}}(Y_c)$，其负梯度 $-\nabla_{Y_c}\mathcal{L}_{\text{pers}}$ 指向能够最大程度延长样本持久生命周期的方向。该损失函数基于一个多参数滤过构造，考虑两个参数：(1) 类别流形Vietoris-Rips过滤（VR_{Y_c}）和 (2) 类别流形核密度估计器（$f = KDE_{Y_c}$）。这一双参数滤过的持久性通过同调次数为 1 的希尔伯特分解有符号测度（H_1）进行概括，记作 $\mu_{H_1}^{Hil}(_{VR_{Y_c}}, \hat{}_f)$（Loiseaux等人，2023b）。该描述符将持久图表示为参数空间（距离、密度）中的有限个正点质量（表示特征诞生）和负点质量（表示特征消亡）。我们的目标是延长持久生命周期，这可以通过最大化该有符号测度与零测度 0 之间的最优传输（OT）距离来实现（Carriere等人，2021）。因此，对于给定类别 c，可微损失函数定义如下：

$$\mathcal{L}_{\text{pers}}(Y_c) = \text{OT}(\mu_{H_1(VR_{Y_c}, \hat{f})}^{Hil}, \mathbf{0}) \tag{3}$$

优化的目标是寻找新的点配置 Y_c'，使该损失最小，并通过梯度下降迭代求解（参见部分 C.1，其中探讨了优化步数）。这一特定形式确保优化增强拓扑稳定性，同时保留类别流形的原始密度，因为每个 epoch 都会重新计算密度，且密度是损失计算不可或缺的一部分。随后，我们将每个属于类别 c 的样本**的持久性分数**$\mathbf{y}_i$定义为其在针对该类别进行优化过程中总位移的幅度，其中$\mathbf{y}_i$ 是初始位置，$\mathbf{y'}_i$ 是最终优化后的位置。：

$$\text{Score}_{\text{pers}}(\mathbf{y}_i) = \|\mathbf{y}_i - \mathbf{y'}_i\|_2, \quad \text{for } \mathbf{y}_i \in Y_c, \mathbf{y'}_i \in Y_c' \tag{4}$$

解读局部数据集结构这一概念。较高的持久性分数衡量了样本在其所属类别流形中引入拓扑不稳定性的程度。关键在于，我们的优化过程经过设计，能够保持密度不变：它增强局部拓扑特征，同时不改变类别流形的整体密度分布。这对核心集选择至关重要，因为它确保我们在寻找结构上重要的样本时，不会扭曲数据的全局代表性。优化过程会重新定位这些点，以明确潜在的类内结构并延长其持久生命周期。因此，这种校正位移的幅度直接、动态地衡量了样本对其所属类别拓扑复杂性的贡献，这一衡量源自流形中所有点的集体相互作用。

3.3. 综合分数：结合全局和局部数据集结构

为构建全面的样本重要性指标，我们制定了一个最终分数，将全局密度和局部持久性信息加权结合为这两项度量的组合：

$$\text{Score}_{\text{unified}}(\mathbf{y}_i) = \alpha \cdot \text{Score}_{\text{pers}}(\mathbf{y}_i) + \beta \cdot \text{Score}_{\text{dens}}(\mathbf{y}_i) \tag{5}$$

其中，超参数 α、$\beta \in [0, 1]$ 用于调节局部拓扑复杂性（持久性分数）与全局分布稀有性（密度分数）的影响。通过在 α 和 β 的不同取值范围内进行探索（见 C.2 节），我们发现 $\alpha = 0.5$ 和 $\beta = 0.5$ 能够较好地平衡全局与局部信息。这样，我们的框架便能构建一个核心集，其中不仅包含丰富的、有挑战性的边界定义样本，还能忠实保留完整数据集的底层分布。

3.4. 无需训练的错误标记样本代理指标

受 CCS 研究发现的启发（Zheng 等人，2023年），我们确保核心集不会受到噪声数据或错误标注数据的污染，因为这类数据可能获得较高的重要性分数，却会降低模型性能（Swayamdipta 等人，2020年）。我们加入一个筛选步骤来创建干净数据集 $\mathcal{D}_{clean} = \mathcal{D} \setminus \mathcal{I}_{mis}$，其中 $\mathcal{I}_{mis}$ 表示错误标注样本索引。虽然大多数

近期方法，包括 CCS 和 D2（Maharana 等人，2024年），采用边际面积（AUM）等训练动态度量（Pleiss 等人，2020年）来识别错误标注样本，但这与我们的最终目标背道而驰：仅基于预训练模型的静态嵌入进行操作。

为解决这一问题，我们提出一种免训练的 AUM 代理指标，称为邻域标签纯度分数（NLPS）。对于每个样本，我们计算其潜在空间中 k 个最近邻里与其类别标签相同的邻居所占比例。NLPS 较低表明样本位于混合标签邻域中，这意味着它可能处于噪声决策边界上，或被错误标注；这与 AUM 识别出的“翻转”候选样本类似。我们通过探索多种 AUM 代理指标验证了这一方法，如C.3节所示，并遵循Zheng 等人 （2023年）确立的标准误标比例，如附录表 13所示。

3.5. 拓扑分层与核心集构建

在最后阶段，我们使用剩余的“干净”数据集（$\mathcal{D}_{clean}$）构建核心集。我们根据计算得出的 $\text{Score}_{\text{unified}}$ 进行分层抽样，以选出能够严格保持原始类别分布的样本，达到预期剪枝率，如Zheng 等人 （2023年）所示。重要的是，TopoPrune 生成的是不平衡的核心集，而不是强制每个类别的样本数量相同；我们尊重原始数据集结构中固有的类别不平衡。这一过程生成的核心集在拓扑结构上信息丰富，并能代表整体数据。

总而言之，TopoPrune 包含三个不同阶段：（1）双尺度拓扑评分（投影全局流形并优化局部持久性）；（2）错误标注样本过滤（创建一个移除了错误标注样本的“干净”数据集）；（3）拓扑选择（基于统一拓扑分数在“干净”数据集上进行分层抽样）。有关实现细节，我们提供了完整的伪代码，见B.2节，并在A.4节中给出了选择过程的示例。

4. 结果

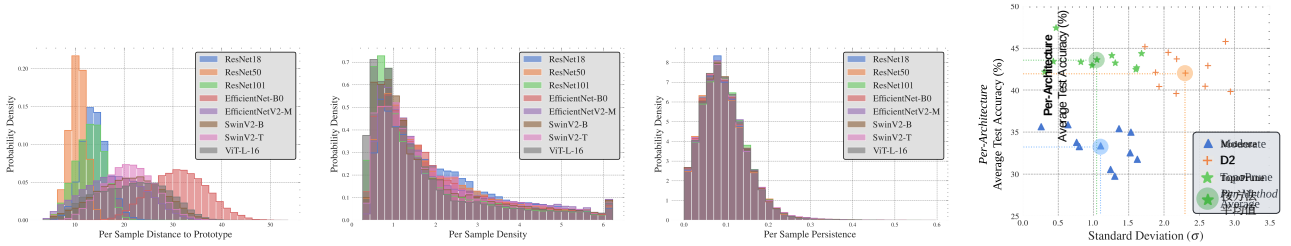
4.1. 实验设置

TopoPrune 使用了多种工具和框架。流形投影通过 `UMAP` (McInnes 等人，2018年) 完成；`multipers` (Loiseaux 与 Schreiber，2024年) 用于实现微分持久同调，后者以GUDHI C++库为后端（Maria 等人，2025年）；此外，`DeepCore`(Guo 等人，2022) 用于统一不同方法中的核心集选择和训练流程。

为确保实验中的公平比较，我们评估了框架的两个版本。在与其他免训练方法进行基准测试时，我们使用*TopoPrune*（NLPS），其利用 NLPS 识别错误标注样本。在与需要训练时信息的方法进行比较时，我们采用使用原始 AUM 分数的 TopoPrune。我们将 TopoPrune（NLPS）与几种静态的基于几何的核心集选择方法进行比较：A）随机选择。B）Moderate（Xia等人，2023年）选取距离类别原型（点质量分布的重心）接近中位数的样本。C）FDMat（Xiao 等人，2024年）通过最优传输，使数据集与核心集的数据分布相匹配。我们还将 TopoPrune 与几种需要训练时信息的基于几何、分数和优化的方法进行比较：D）Moderate（AUM），将 Moderate 与基于 AUM 的错误标注样本剔除相结合。E）遗忘（Toneva 等人，2019年）依据训练过程中样本在先前被正确分类后被错误分类的次数进行筛选。F）Glister（Killamsetty 等人，2021b）采用双层优化。G）LCMat-S（Shin 等人，2023年）使数据集与核心集的损失曲率相匹配。H）CCS（Zheng 等人，2023年）对难度分数（如 AUM 或遗忘）进行分层抽样，并在层内随机抽样。H）D2（Ma-

表 *1.* **各核心集选择方法的准确率**，涵盖 CIFAR-10、CIFAR-100 和 ImageNet-1K。TopoPrune 的优势（准确率与精确率）在 ImageNet-1K 等具有挑战性的数据集以及高剪枝率（例如 90%）下最为显著。

	剪枝率（→）	30%	50%	70%	80%	90%
		CIFAR-10（ResNet-18）				
No 训练动态	随机	94.5±0.1	93.5±0.1	90.8±0.2	86.6±0.3	76.7±0.9
	Moderate	94.2±0.1	93.1±0.1	89.9±0.2	87.2±0.2	76.9±1.0
	FDMat	94.7±0.1	93.6±0.2	**90.8±0.2**	87.3±0.4	74.4±0.7
	TopoPrune (NLPS)	**94.8±0.1**	**93.6±0.2**	90.3±0.2	**87.3±0.3**	**77.1±0.6**
With 训练动态	Moderate (AUM)	93.9±0.2	93.1±0.2	90.1±0.2	87.1±0.2	79.9±0.3
	遗忘	94.5±0.2	92.6±0.1	89.8±0.2	85.6±0.3	67.6±0.4
	Glister	94.4±0.2	93.8±0.2	90.8±0.4	85.1±0.6	66.8±1.3
	LCMat-S	94.5±0.2	93.3±0.2	90.5±0.2	86.9±0.2	75.1±0.8
	CCS	95.5±0.1	94.8±0.2	93.0±0.2	**90.7±0.2**	81.9±0.7
	D2	**95.6±0.1**	**94.8±0.1**	**93.1±0.1**	89.2±0.2	80.9±1.5
	TopoPrune	94.7±0.2	93.7±0.2	91.6±0.1	88.7±0.4	**82.1±0.3**
		CIFAR-100（ResNet-18）				
No 训练动态	随机	75.3±0.2	71.6±0.1	63.7±0.5	55.9±1.0	34.0±1.1
	Moderate	74.9±0.3	70.1±0.3	63.7±0.2	56.1±0.5	34.9±2.1
	FDMat	75.4±0.2	71.9±0.3	64.0±0.6	56.1±1.5	37.5±1.6
	TopoPrune (NLPS)	**75.6±0.2**	**71.9±0.2**	**65.3±0.4**	**56.7±0.4**	**41.6±0.8**
With 训练动态	Moderate (AUM)	75.9±0.3	72.4±0.2	66.7±0.3	60.2±0.8	40.0±1.2
	遗忘	74.8±0.2	67.2±0.9	50.6±0.7	32.3±0.9	24.3±1.4
	Glister	75.8±0.3	70.7±0.7	66.1±1.2	54.7±1.6	38.4±1.7
	LCMat-S	75.3±0.2	71.1±0.2	62.5±0.8	52.1±2.0	36.1±1.7
	CCS	**76.9±0.3**	**73.8±0.3**	**67.8±0.7**	60.7±0.6	45.2±2.4
	D2	75.1±0.5	71.2±0.2	67.8±0.9	61.1±1.4	44.3±2.6
	TopoPrune	75.9±0.4	72.8±0.3	66.9±0.5	**61.9±0.6**	**45.8±0.7**
		ImageNet-1K（ResNet-50）				
No 训练动态	随机	69.8±0.5	68.4±0.5	65.1±0.4	61.9±0.5	52.5±0.6
	Moderate	69.5±0.2	65.8±0.4	60.5±0.1	57.7±0.2	50.0±0.4
	FDMat	**70.8±0.3**	68.7±0.5	65.5±0.7	62.0±0.3	51.9±0.3
	TopoPrune (NLPS)	70.7±0.4	**69.9±0.2**	**66.4±0.2**	**63.2±0.3**	**53.9±0.2**
With 训练动态	Moderate（AUM）	69.6±0.4	67.2±0.6	63.9±0.8	60.4±0.6	52.7±0.3
	遗忘	69.9±0.2	66.8±0.6	60.2±0.5	59.1±0.4	50.0±0.5
	Glister	66.3±0.4	63.5±0.3	59.3±0.5	56.5±0.3	49.3±0.8
	LCMat-S	69.8±0.4	67.5±0.5	62.2±0.3	59.7±0.5	48.8±0.6
	CCS	70.1±0.5	69.1±0.3	65.7±0.3	62.6±0.6	55.2±0.7
	D2	69.5±0.3	67.1±0.5	65.7±0.4	62.7±0.9	55.5±1.3
	TopoPrune	**70.8±0.2**	**69.5±0.2**	**66.2±0.1**	**63.1±0.3**	**56.1±0.2**



(a) 跨数据的欧几里得距离网络　*(b)*全局流形核跨网络密度　*(c)* 跨网络局部持久性网络　*(d)* 多样的可迁移性嵌入 → 固定目标

图*3*。 **拓扑度量在不同网络间更加一致**。这直接转化为更优的核心集性能。随着我们从**(a)**不稳定的欧几里得距离，转向**(b)** 基于全局拓扑投影的密度估计，最终到**(c)** 局部持久性，度量分布逐渐变得更加均匀。这种增强的度量稳定性使 TopoPrune 能够持续优于基于几何的基线方法**(d)**：在 CIFAR-100 上以 90% 的高剪枝率对 10 种不同架构进行测试时，TopoPrune 同时实现了更高的平均准确率和更低的标准差；图中左上角（标准差低、准确率高）为最佳表现。

harana 等人，2024年）采用消息传递图网络，同时也利用 AUM 识别错误标注样本。所有报告的准确率和标准差均根据五次独立训练运行计算得出。

4.2. 使用 TopoPrune 构建高性能且稳定的核心集

我们的实验详见表 1，结果表明，在 CIFAR-10、CIFAR-100 和 ImageNet-1K 上，TopoPrune 生成的核心集质量始终高于现有基线方法。具体而言，在大多数剪枝场景中，TopoPrune（NLPS）的表现优于所有其他无训练动态方法。值得注意的是，这一性能优势会随着任务难度增加而扩大。虽然在较简单的数据集上具有竞争力，但在具有挑战性的 ImageNet-1K 基准测试以及极高剪枝率（例如 90%）下，我们的方法优势最为显著，并始终名列前茅。除准确率外，TopoPrune 还表现出卓越的稳定性。表 1中的标准差指标以及D.1 部分的统计显著性检验结果均证实，与外在几何基线方法相比，我们的方法具有显著更高的精确率。例如，在 ImageNet-1K 上进行 90% 剪枝时，TopoPrune 将方差最多降低 $6.5\times$，相比之下，D2 使用的是基于图的启发式方法。这种稳定性具有实际价值，可确保数据选择可靠，无需多次进行代价高昂的试验来减轻从头训练的随机性。

4.3. 不同架构间的稳定性和可迁移性

我们研究了几何度量在各种网络架构中的局限性，发现随着转向基于拓扑的度量，其稳定性和可迁移性均显著提升。(1) 度量稳定性：首先，我们分析样本与其类别原型之间的欧几里得距离。事实证明，这一度量在不同架构之间高度不一致，表现出“几何脆弱性”（图 3a）。相比之下，我们的拓扑度量表现出显著的稳定性。全局密度分数（图 3b）提升了均匀性，而局部持久性分数（图 3c）在所有测试架构之间几乎完全一致。(2) 多样化嵌入 → 固定目标：利用这种稳定性，我们评估了从多样化特征嵌入到单一目标模型的可迁移性。如图 3d所示，TopoPrune 在多种代理模型（ResNet、EfficientNet、Swin、ViT、OpenCLIP）之间始终能取得更高的准确率和更低的标准差，以训练固定的 ResNet-18 目标模型，显著优于几何基线方法。详细数值请参见附录表10 ，进一步的理论论证请参见A.2节 。(3)固定嵌入 → 多样化目标：最后，我们验证了从静态特征嵌入向多种目标模型迁移的效果。我们使用标准 ResNet 代理模型选择核心集，用于训练 EfficientNet 和 Swin Transformer。结果发现，TopoPrune 的性能可与“预言机”选择相媲美，在某些情况下甚至超过后者；“预言机”选择即由目标模型自行选择核心集（见表2）。这证实了从标准代理嵌入中得出的拓扑重要性具有很强的泛化能力，使用户能够选出单一的高质量核心集，并使其适用于各种下游架构。

影响。 我们的方法具有更优异的架构可迁移性，这除了能用于小型代理模型之外（Coleman等人，2020年），还具有实际意义。这是迈向构建可重复使用的模型无关核心集的一步：可以先用高效模型整理数据集，再将其用于训练或评测多种不同的大型架构。这种方法与近期利用拓扑测度（如贝蒂数）来研究和比较不同网络嵌入空间复杂性的工作相契合（Suresh等人，2024年）。

表2. **固定嵌入的**可迁移性 → 多样化目标。我们比较“预言机”性能（使用目标模型自身的嵌入从 TopoPrune 中选出的核心集）与使用较小代理模型的嵌入从 TopoPrune 中选出的核心集（CIFAR-100 使用 ResNet-18，ImageNet-1K 使用 ResNet-50），并用这些核心集训练各类模型。

剪枝率（→）	80%			90%		
CIFAR-100	预言机	ResNet-18	Δ	预言机	ResNet-18	Δ
ResNet-50	57.0±0.6	56.1±0.7	-0.9	38.7±1.4	39.6±1.6	+0.9
EfficientNet-B0	55.0±2.1	54.4±1.4	-0.6	39.8±1.9	39.7±3.3	-0.1
SwinV2-T	38.1±0.7	39.9±0.9	+1.8	27.6±1.2	29.6±1.5	+2.0

剪枝率 (→)	80%			90%		
ImageNet-1K	预言机	ResNet-50	Δ	预言机	ResNet-50	Δ
EfficientNetV2-M	39.1±1.4	40.8±0.3	+1.7	35.9±0.3	37.1±1.3	+1.2
SwinV2-T	57.8±0.7	59.0±1.1	+1.2	38.3±3.4	38.0±4.1	-0.3
SwinV2-B	44.7±1.2	45.8±2.1	+1.1	44.4±2.1	45.1±1.1	+0.7

4.4. 噪声特征嵌入的稳定性

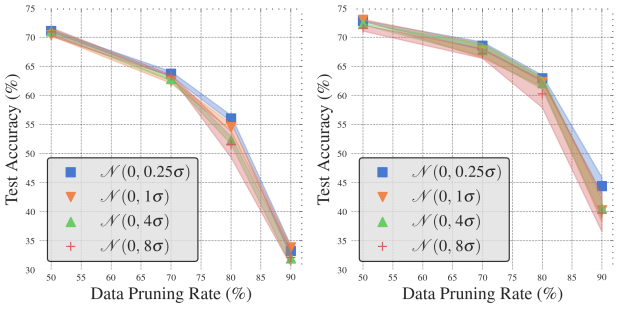
为评估我们的方法对潜在空间中扰动的鲁棒性，我们向 CIFAR-100 上 ResNet-18 模型倒数第二层的嵌入注入高斯噪声。我们将 TopoPrune 与基于几何学的方法基线 Moderate（Xia等人，2023年）和 D2（Maharana等人，2024年）进行比较；这些方法对这类扰动较为敏感。通过扰动每个样本的特征向量 $\mathbf{z} \in \mathbb{R}^D$，生成“含噪”版本的特征嵌入。具体而言，首先计算每个样本特征向量的标准差 $\sigma_{\mathbf{z}}$，然后将噪声向量 $\epsilon \sim \mathcal{N}(0,\ \sigma_{\mathbf{z}})$ 添加到原始特征 $\mathbf{z}' = \mathbf{z} + \epsilon$。

如图 4所示，TopoPrune 在所有测试噪声水平下都表现出更强的抗噪能力。随着扰动增大，基线方法的性能会下降，而我们的拓扑方法则能保持较高的准确率和稳定性。在高剪枝率下，这种鲁棒性尤为明显，凸显了我们的方法即使在噪声特征空间中也能识别出稳定核心集的能力。详细的定量结果请参见附录中的表9。

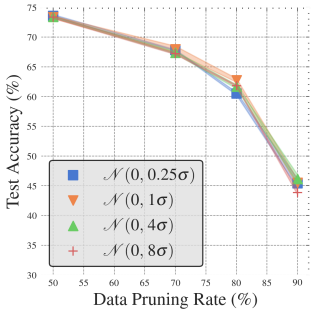
意义。 这种对噪声特征的鲁棒性表明，我们的方法并不过度依赖经过完美优化的源模型。这说明，即使使用部分训练的模型、为边缘设备量化的模型，或应用于轻微分布外数据的模型所生成的嵌入，也能构建有效的核心集，这一点与（Turkes 等人，2022）的结果类似。这种灵活性拓宽了我们的方法在更多实际场景中的潜在适用范围。

5. 结论

在本研究中，我们针对几何核心集选择方法中的关键不稳定性问题展开研究；这类不稳定性源于其对外在度量的依赖。我们提出了 TopoPrune，这是一种通过利用拓扑来捕捉数据内在结构、克服这种“几何脆弱性”的新型框架——



*(a)*Moderate（Xia等人，2023年）*(b)*D2（Maharana等人，2024年）



(c) TopoPrune（本文方法）

图4。**Moderate、D2 和 TopoPrune 在 CIFAR-100 上的噪声特征嵌入影响。(a)** Moderate 的精确率尚可（阴影区域的紧密程度表明了这一点），但平均准确率较低。**(b)** D2 的准确率更高，但精确率明显较低，尤其是在剪枝率较高且 ϵ-噪声较大的情况下。**(c)** TopoPrune 在所有剪枝率和 ϵ-噪声水平下均实现了较高的准确率和更小的标准差。

性”。我们的双尺度拓扑方法结合了全局拓扑感知的流形投影，以及从可微持久同调中导出的局部重要性分数。TopoPrune 具有以下几个关键优势：**(1)** 能够生成具有可证明更高准确率和稳定性的核心集；**(2)** 在广泛的网络架构中表现出更高的稳定性；**(3)** 对噪声特征嵌入具有很强的鲁棒性。

TopoPrune 的稳定性使其在实际应用中颇具价值。在大型基础模型的训练中，它通过确保数据子集始终保持高质量，遏制“随机种子彩票”现象，降低不稳定性的风险。在自动化流水线中，其更具确定性的特性为构建可信、易维护且安全关键的人工智能系统提供了必不可少的可审计性和可靠性。TopoPrune 兼具准确率与稳定性，标志着迈向可行且有原则的、基于拓扑的数据高效学习框架的一步。

影响声明

本文的研究通过确立微分持久同调作为一种严谨工具，用于探究神经表示的内在结构，推进了我们对深度学习的理解。TopoPrune 将全局流形几何学与局部拓扑相互作用相结合，为理解深度模型如何组织和分离数据提供了新颖视角。除直接应用于高效核心集选择外，该框架还提供了一种稳健的免训练机制，用于量化样本重要性并过滤标签噪声。归根结底，这项工作为未来的拓扑可解释性工具奠定了基础，在提供有关复杂模型动态的可扩展洞见的同时，也能在资源受限的环境中实现高性能训练。

致谢

本项目部分经费由普渡大学安全微电子生态系统中心（CSME#210205）和认知系统协同设计中心（CoCoSys）提供；后者是美国国防高级研究计划局资助的 JUMP 2.0 中心。

参考文献

Bauer, U., Masood, T. B., Giunti, B., Houry, G., Kerber, M., 和 Rathod, A. 保持稀疏：重新审视持久同调的计算。*arXiv*，2022年。Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I., Ng, L. G., Ginhoux, F., 和 Newell, E. W. 使用 UMAP 降维以可视化单细胞数据。自然生物技术，2018年。Birdal, T., Lou, A., Guibas, L., 和 Simsekli, U. 神经网络中的内在维度、持久同调与泛化。载于神经信息处理系统进展，2021。Boissonnat, J.-D. 和 Maria, C. 单纯形树：一种适用于一般单纯复形的高效数据结构。*Algorithmica*，2014年。Borsos, Z., Mutny, M., 和 Krause, A. 通过双层优化构建核心集，以用于持续学习和流式处理。神经信息处理系统进展，2020年。Botnan, M. B. 和 Lesnick, M. 多参数持久性的导论。*arXiv*，2022年。Carriere, M., Chazal, F., Glisse, M., Ike, Y., Kannan, H., 和 Umeda, Y. 优化基于持久同调的函数。载于国际机器学习会议。PMLR，2021年。Carrière, M., Theveneau, M. 和 Lacombe, T. 用于高效持久性拓扑优化的微分同胚插值。载于神经信息处理系统进展，2024年。Chazal, F., Cohen-Steiner, D., Guibas, L. J., Mémoli, F. 和 Oudot, S. Y. 利用持久性为形状构建 Gromov–Hausdorff 稳定签名。计算机图形论坛，2009年。Chen, H., Chen, W., Mei, H., Liu, Z., Zhou, K., Chen, W., Gu, W. 和 Ma, K.-L. 多类别散点图的视觉抽象与探索。*IEEE* 可视化与计算机图形学汇刊，2014年。Cho, Y., Shin, B., Kang, C. 和 Yun, C. 基于样本难度和预测不确定性、无需完整训练的轻量级数据集剪枝。载于国际机器学习会议，2025年。Cohen-Steiner, D., Edelsbrunner, H. 和 Harer, J. 持久图的稳定性。载于计算几何研讨会，2005年。Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J. 和 Zaharia, M. 通过代理进行选择：深度学习中的高效数据选择。载于国际学习表征会议，2020年。de Bodt, C., Diaz-Papkovich, A., Bleher, M., Bunte, K., Coupette, C., Damrich, S., Sanmartin, E. F., Hamprecht, F. A., Horvát, E.-Á., Kohli, D. 等。高维数据的低维嵌入。*arXiv*，2025年。Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., 何恺明, Gelly, S. 等。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于国际学习表征会议，2021。Edelsbrunner, H. 和 Harer, J. L. 计算拓扑学：导论。美国数学会，2010。Fel, T., Boutin, V., Béthune, L., Cadene, R., Moayeri, M., Andéol, L., Chalvidal, M. 和 Serre, T. 统一自动概念提取与概念重要性估计的整体方法。载于神经信息处理系统进展，2023年。Fel, T., Béthune, L., Lampinen, A. K., Serre, T. 和 Hermann, K. 从复杂性的视角理解视觉特征依赖。载于神经信息处理系统进展，2024年。Feldman, V. 学习是否需要记忆？一个关于长尾的简短故事。载于*ACM SIGACT* 计算理论研讨会，2020年。

Feldman, V. 和 Zhang, C. 神经网络记忆什么以及为何记忆：通过影响力估计发现长尾。神经信息处理系统进展，2020年。

Garg, I. 和 Roy, K. 低损失曲率样本可提高数据效率。载于*IEEE/CVF*计算机视觉与模式识别会议，2023年。

Garg, I.、Ravikumar, D. 和 Roy, K. 从样本周围损失函数曲率的视角看记忆。载于国际机器学习会议，2024年。

Guo, C.、Zhao, B. 和 Bai, Y. DeepCore：用于深度学习核心集选择的综合库。载于数据库与专家系统应用国际会议，第181–195页。Springer，2022年。何恺明、Zhang, X.、Ren, S. 和 Sun, J. 用于图像识别的深度残差学习。载于*IEEE/CVF*计算机视觉与模式识别会议，2016年。He, M.、Yang, S.、Huang, T. 和 Zhao, B. 利用动态不确定性进行大规模数据集剪枝。载于*IEEE/CVF*计算机视觉与模式识别会议，2024年。He, Y.、Xiao, L. 和 Zhou, J. T. 只浓缩一次：剪枝浓缩数据集的两条规则。载于神经信息处理系统进展，2023年。Jeon, H.、Park, J.、Shin, S. 和 Seo, J. 停止误用 t-SNE 和 UMAP 进行可视分析。*arXiv*，2025年。Killamsetty, K.、Ramakrishnan, G.、De, A. 和 Iyer, R. Grad-Match：基于梯度匹配的数据子集选择，用于高效深度模型训练。载于国际机器学习会议，2021a。Killamsetty, K.、Sivasubramanian, D.、Ramakrishnan, G. 和 Iyer, R. Glister：基于泛化的数据子集选择，实现高效且鲁棒的学习。载于人工智能 *AAAI* 会议，2021b。Lesnick, M. 和 Wright, M. 计算二参数持久同调的最小表示与双分次贝蒂数。*SIAM*应用代数与几何杂志，2022年。Liu, Z.、Hu, H.、Lin, Y.、Yao, Z.、Xie, Z.、Wei, Y.、Ning, J.、Cao, Y.、Zhang, Z.、Dong, L. 等。Swin Transformer V2：扩展容量与分辨率。载于*IEEE/CVF*计算机视觉与模式识别会议，2022年。Loiseaux, D. 和 Schreiber, H. Multipers：面向机器学习的多参数持久性。开源软件期刊，9(103):6773，2024年11月。ISSN 2475-9066。doi：10.21105/joss.06773。

Loiseaux, D.、Carrière, M. 和 Blumberg, A. 多参数持久同调分解的快速稳定表示框架。载于神经信息处理系统进展，2023年。Loiseaux, D.、Scoccola, L.、Carrière, M.、Botnan, M. B. 和 Oudot, S. 使用带符号条形码作为测度实现多参数持久同调的稳定向量化。神经信息处理系统进展，2023年。Maharana, A.、Yadav, P. 和 Bansal, M. D2 剪枝：用于平衡数据剪枝中多样性与难度的消息传递。载于国际学习表征会议，2024年。Maria, C.、Dlotko, P.、Rouvreau, V. 和 Glisse, M. Rips 复形。载于*GUDHI* 用户与参考手册。GUDHI 编辑委员会，第 3.11.0 版，2025年。McInnes, L.、Healy, J. 和 Melville, J. UMAP：用于降维的统一流形近似与投影。*arXiv*，2018a。McInnes, L.、Healy, J.、Saul, N. 和 Grossberger, L. UMAP：统一流形近似与投影。《开放源代码软件杂志》，3(29):861，2018年。Mindermann, S.、Brauner, J. M.、Razzak, M. T.、Sharma, M.、Kirsch, A.、Xu, W.、Höltgen, B.、Gomez, A. N.、Morisot, A.、Farquhar, S. 等。针对可学习、值得学习且尚未学会的样本进行优先训练。载于国际机器学习会议，2022。Mirzasoleiman, B.、Bilmes, J. A. 和 Leskovec, J. 面向数据高效机器学习模型训练的核心集。载于国际机器学习会议，2019年。Mishra, A. 和 Motta, F. C. 使用 Delaunay-Rips 复形的持久同调稳定性及其机器学习应用。《应用数学与统计学前沿》，2023年。Moor, M.、Horn, M.、Rieck, B. 和 Borgwardt, K. 拓扑自编码器。载于国际机器学习会议，2020年。Mukherjee, S.、Samaga, S. N.、Xin, C.、Oudot, S. 和 Dey, T. K. D-gril：基于 2-参数持久性的端到端拓扑学习。*arXiv*，2024年。Nagaraj, M.、Ravikumar, D. 和 Roy, K. 轨迹核心集：通过损失差异的相关性选择数据。《机器学习研究汇刊》，2025年。Naitzat, G.、Zhitnikov, A. 和 Lim, L.-H. 深度神经网络拓扑。《机器学习研究杂志》，2020年。

Otter, N.、Porter, M. A.、Tillmann, U.、Grindrod, P. 和 Harrington, H. A. 持久同调计算路线图。*EPJ* 数据科学，2017年。

Oudot, S. 持久性理论：从箭图表示到数据分析。数学调查与专著。2015年。ISBN 9781470425456。

Papillon, M., Sanborn, S., Mathe, J., Cornelis, L., Bertics, A., Buracas, D., Lillemark, H. J., Shewmake, C., Dinc, F., Pennec, X., 等。《超越欧几里得：几何、拓扑与代数结构视角下的现代机器学习图解指南》。机器学习：科学与技术，2025年。

Paul, M., Ganguli, S., 和 Dziugaite, G. K.《数据节食下的深度学习：尽早找出重要样本》。神经信息处理系统进展，2021。Pearson, K.《空间中点系的最近点拟合直线与平面，第三篇》。伦敦、爱丁堡和都柏林哲学杂志与科学杂志，1901年。Pleiss, G., Zhang, T., Elenberg, E., 和 Weinberger, K. Q.《利用间隔排序曲线下面积识别错误标注数据》。神经信息处理系统进展，2020年。Pooladzandi, O., Davini, D., 和 Mirzasoleiman, B.《面向数据高效机器学习的自适应二阶核心集》。载于国际机器学习会议，2022。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., 和 Sutskever, I.《通过自然语言监督学习可迁移视觉模型》。载于国际机器学习会议，2021。Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., 和 Jitsev, J.《LAION-5b：用于训练新一代图文模型的开放式大规模数据集》。载于神经信息处理系统进展数据集与基准专题，2022。

Scoccola, L., Setlur, S., Loiseaux, D., Carrière, M., 和 Oudot, S.《多参数持久同调的可微性与优化》。载于国际机器学习会议，2024年。Seifert, H. 和 Threlfall, W. 拓扑学教材。学术出版社，1980年。ISBN 0-12-634850-2。

Shin, S., Bae, H., Shin, D., Joo, W. 和 Moon, I.-C. 用于数据集选择和凝缩的损失-曲率匹配。载于人工智能与统计国际会议，2023年。Suresh, S., Das, B., Abrol, V. 和 Roy, S. D. 使用代数拓扑刻画神经网络嵌入空间的演化。《模式识别快报》，2024年。

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A. 和 Choi, Y. 数据集制图：利用训练动态对数据集进行映射和诊断。*arXiv*，2020年。Tan, H., Wu, S., Du, F., Chen, Y., Wang, Z., Wang, F. 和 Qi, X. 通过移除单个样本进行数据剪枝。载于神经信息处理系统进展，2023年。

Tan, M. 和 Le, Q. Efficientnet：重新思考卷积神经网络的模型缩放。载于国际机器学习会议，2019年。

Tan, M. 和 Le, Q. Efficientnetv2：更小的模型和更快的训练。载于国际机器学习会议，2021。

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y. 和 Gordon, G. J. 深度神经网络学习过程中样本遗忘的实证研究。载于国际学习表征会议，2019年。Trofimov, I., Cherniavskii, D., Tulchinskii, E., Balabin, N., Burnaev, E. 和 Barannikov, S. 学习保持拓扑的数据表示。载于国际学习表征会议，2023年。Turkes, R., Montufar, G. F. 和 Otter, N. 关于持久同调的有效性。神经信息处理系统进展，2022年。van der Maaten, L. 和 Hinton, G. 使用 t-SNE 进行数据可视化。《机器学习研究杂志》，2008年。Wang, Y., Huang, H., Rudin, C. 和 Shaposhnik, Y. 理解降维工具的工作原理：解析 t-SNE、UMAP、TriMAP 和 PaCMAP 数据可视化功能的实证方法。《机器学习研究杂志》，2021。Xia, X., Liu, J., Yu, J., Shen, X., Han, B. 和 Liu, T. 适度核心集：一种适用于真实世界数据高效深度学习的数据选择通用方法。载于国际学习表征会议，2023年。

Xiao, W., Chen, Y., Shan, Q., Wang, Y. 和 Su, J. 通过最优传输进行特征分布匹配，以实现有效且稳健的核心集选择。*AAAI* 人工智能会议，2024年。

Xie, T., Zhu, J., Ma, G., Lin, M., Chen, W., Yang, W., and Liu, S. 基于结构熵的样本选择，以实现高效且有效的学习。在国际学习表征会议，2025年。

Xiong, Z., Dalmasso, N., Sharma, S., Lecue, F., Maga-zzeni, D., Potluru, V. K., Balch, T., and Veloso, M. 公平的Wasserstein核心集。在神经信息处理系统进展，2024年。

Yang, S., Cao, Z., Guo, S., Zhang, R., Luo, P., Zhang, S., 和 Nie, L. 注意决策边界：通过重构决策边界进行核心集选择。载于国际机器学习会议，2024年。Zheng, H., Liu, R., Lai, F., 和 Prakash, A. 面向高剪枝率的以覆盖率为中心的核心集选择。载于国际学习表征会议，2023年。Zheng, H., Tsai, E., Lu, Y., Sun, J., Bartoldson, B. R., Kailkhura, B., 和 Prakash, A. ELFS：基于代理训练动态的无标签核心集选择。载于国际学习表征会议，2025年。

附录

目录

概念框架与直觉13
A.1 单纯复形与持久同调概述 .13A.2 理论论证：拓扑特征与欧几里得特征的可迁移性 14
A.2.1 欧几里得距离对原型的不稳定性 .15A.2.2 持久同调的稳定性保证 .15A.3 拓扑密度与样本记忆的几何 .16A.4 核心集构建示例 .16

B 实现与计算复杂度分析 17
B.1 流形投影技术的比较分析 .17B.1.1 标准线性与非线性流形近似的评估 .17B.1.2 深度拓扑自编码器的评估 .17B.2 TopoPrune 伪代码 .20B.3 计算复杂度分析 .21

C 消融研究 22
C.1 可微持久性优化步骤 .22C.2 局部持久性（α）和全局密度（β）的敏感性 .22C.3 用于错误标签检测的间隔下方面积（AUM）免训练代理指标 .23C.4 对 UMAP 流形投影超参数的敏感性 .23

D 扩展实证结果 23
D.1 稳定性提升的统计显著性 .23D.2 详细实验结果路线图 .24

A. 概念框架与直觉

A.1. 单纯复形与持久同调概述

持久同调用于刻画有限度量空间形状在多个尺度下的拓扑变化。从较高层面来看，这可以描述为拓扑结构（由同调群定义）的“诞生”与“消亡”（持久性）。这一过程始于构造一个单纯复形，它由点（0-单形）、边（1-单形）、三角形（2-单形）及其更高维对应物组成，用于表示数据的结构（Boissonnat & Maria，2014）。为了分析该结构如何随尺度变化，我们构造一个滤过（见图 5a）。这是一个嵌套的单纯复形序列，$K_{r_1} \subseteq K_{r_2} \subseteq \cdots \subseteq K_{r_n}$由非递减的尺度参数 r索引。对于滤过中的每个复形 K_r，我们都可以计算其同调群，$H_k(K_r)$，它们是以代数方式刻画其 k维特征的向量空间。

这个群的秩被称为 k阶 贝蒂数 ($\beta_k = \text{秩}(H_k(K_r))$，用于统计这些特征：$\beta_0$ 统计连通分支，β_1 统计环或隧道，β_2 统计空洞，依此类推。这些数值对于理解对象的外在 几何学与其 内在 拓扑属性之间的区别至关重要。一个说明这种差异的经典例子是咖啡杯和环面（甜甜圈）。这两个对象在拓扑上是等价的，因为它们具有相同的贝蒂数（见 图 5b）。尽管它们的外在几何学（包括其形状、曲率，以及嵌入三维空间后的距离）大不相同，但它们的内在拓扑却完全一致。这是因为可以在不撕裂或粘合的情况下，将其中一个连续变形为另一个，同时保留定义二者的那个单一孔洞。

持久同调并不局限于任何一种度量构造方式；事实上，你可以在纯粹的抽象单纯复形及其上的任意滤过上进行持久性分析。然而，针对我们的应用，我们将其应用于点云$P = \{\mathbf{x}_i\}$，采用常见的方法构建滤过，即使用 *Vietoris-Rips*（*VR*）复形 (见 图 5c）。对于给定的尺度 $r \geq 0$，复形 $VR(P, r)$ 包含所有满足以下条件的单形：$\sigma \subseteq P$任意两点之间的欧几里得距离

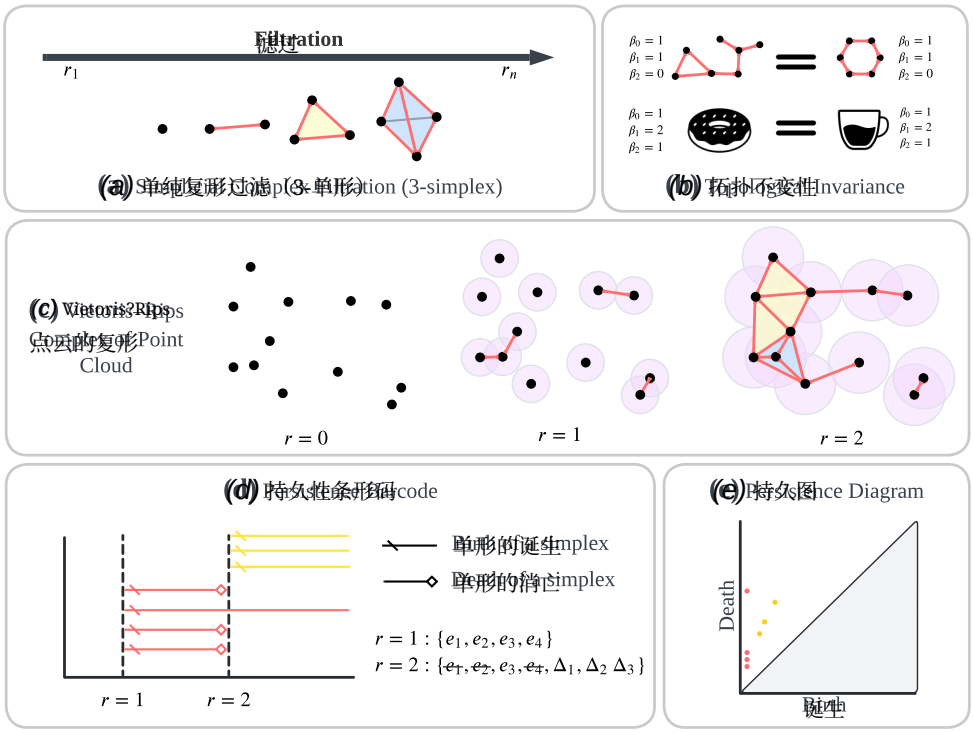


图 5. 单纯复形与持久同调概述

中的点 σ 至多为 $2r$。随着 r 增大，单形会被添加到复形中，导致新的连通分量与较早出现的连通分量合并。持久同调会追踪整个滤过过程中这些拓扑特征的诞生与消亡。 包含映射 $K_{r_i} \hookrightarrow K_{r_j}$ 对于 $r_i \leq r_j$ 在同调群之间诱导出一个同态， $H_k(K_{r_i}) \to H_k(K_{r_j})$。当某个特征首次出现时，其尺度称为“诞生”尺度 r_{birth} ；当它与较早出现的特征合并时，其尺度称为“消亡”尺度 r_{death} ，这一过程由持久性条形码表示（图 5d）。

定义 A.1 (Vietoris-Rips 过滤)。 对于点云 $P \subset \mathbb{R}^n$ 和尺度参数 $r \geq 0$， *Vietoris-Rips*复形$VR(P, r)$ 是这样一种单纯复形：其顶点是 P 中的点，单形则是 P 中直径至多为 $2r$的所有有限子集。滤过是复形 $\{VR(P, r)\}_{r \geq 0}$构成的嵌套序列。

这一过程的输出可概括为一个持久图Dgm(P)，它是平面上的一个点多重集，其中每个点对应一个拓扑特征，并以其(诞生，消亡) $\to (b,\ d)$ 坐标绘制（见图 5e）。特征的持久性定义为其生命周期， $d - b$。图中距离对角线 $y = x$较远的点代表数据中稳健且具有结构意义的特征，而靠近对角线的点则被视为生命周期较短的拓扑噪声。这为数据的内在形状提供了稳定的多尺度签名。

定义 A.2（持久图）。将同调函子 $H_k(\cdot)$ （对于固定维度 k，例如 $k = 0$ 连通分支）应用于一个滤过，可得到一组诞生-消亡对 (b, d) ，表示拓扑特征出现和消失时对应的尺度。这些配对构成一个多重集，即持久图，记作 $Dgm(P)$。特征 (b, d) 的持久性定义为 $d - b$。

请注意，为便于在本概述部分中清晰呈现并易于可视化，我们展示的是单参数持久性分析。但需要指出的是，我们的方法 TopoPrune 采用的是多参数持久模，其可视化更为复杂，却能更丰富地描述数据的拓扑。

A.2. 理论论证：论拓扑特征与欧几里得特征的可迁移性 s

我们从理论上论证了，基于持久同调提取的拓扑特征相较于传统欧几里得度量，在不同神经网络架构之间具有更强的可迁移性。我们证明，持久同调固有的稳定性保证了其输出能够抵御不同网络

嵌入之间常见的几何变化。相反，我们表明，基于欧氏距离的度量（例如到类别原型的距离）本质上容易受到这些变化的影响（Papillon 等人，2025年）。

预备知识与符号。 令 X 为输入数据空间，$Y = \{1, \ldots, K\}$ 为 K 个类别标签组成的集合。神经网络架构是一个函数 $f: X \to \mathbb{R}^n$，它将输入数据映射到一个 n维嵌入空间。令 f_A 和 f_B 表示两个不同的网络架构（例如 ResNet18 和 ViT-L-16）。对于整个数据集 X，这些网络的输出分别是点云 $X_A = f_A(X)$ 和 $X_B = f_B(X)$，位于各自的嵌入空间中。我们在这些嵌入空间上赋予标准欧氏度量 d_E。

定义 A.3 (瓶颈距离)。 两个持久图 Dgm_1 和 Dgm_2 之间的相似性通过瓶颈距离 $d_B(Dgm_1, Dgm_2)$衡量，其定义为对所有双射 $\eta : Dgm_1 \to Dgm_2$ 取下确界，即匹配点之间距离的上确界；其中 $p \in Dgm$ 表示持久图中的一个点，对应于其生灭对 $p = (b, d)$：

$$d_B(Dgm_1, Dgm_2) = \inf_{\eta} \sup_{p \in Dgm_1} \|p - \eta(p)\|_\infty$$

定义 A.4 （格罗莫夫-豪斯多夫距离）。 两个度量空间（M）配备距离函数（d）；（M_1, d_1）和（M_2, d_2）之间的距离由格罗莫夫-豪斯多夫距离 d_{GH}（M_1, M_2）度量，它是将两个空间等距嵌入到同一个度量空间的所有可能方式所对应距离的下确界。它量化了两个空间之间的“度量不相似性”。

A.2.1. 欧几里得距离对原型的不稳定性

现在，我们将形式化阐述欧几里得距离缺乏这种稳定性的情况。

定义 A.5 （类别原型与距离分布）。 对于嵌入 f（X）和类别 $k \in Y$，类别原型（质心）为 $c_k = \frac{1}{|X_k|} \sum_{x \in X_k} f$（$x$），其中 X_k 是类别 k的样本。到原型的距离集合为S_k（f）$= \{d_E$（f（x） c_k）$|$标签（x）$= k\}$。令 P（S_k（f））为这些距离的概率分布。

命题 A.6 （对缩放的敏感性）。 设 f_A 是一个网络嵌入。考虑通过简单的各向同性缩放变换定义的新嵌入 f_B： f_B（x）$= \alpha f_A$（x），其中 $\alpha > 0$、$\alpha \neq 1$ 为某个标量。那么，到原型的距离分布也会相应缩放： P（S_k（f_B）$= \alpha P$（S_k（f_A）。

证明。 在嵌入 f_B 下，新的类别原型 c'_k 为：

$$c'_k = \frac{1}{|X_k|} \sum_{x \in X_k} f_B(x) = \frac{1}{|X_k|} \sum_{x \in X_k} \alpha f_A(x) = \alpha \left(\frac{1}{|X_k|} \sum_{x \in X_k} f_A(x) \right) = \alpha c_k$$

类别 k 中任意样本 x 到新原型的距离为：

$$\begin{aligned} d_E(f_B(x), c'_k) &= d_E(\alpha f_A(x), \alpha c_k) \\ &= \|\alpha f_A(x) - \alpha c_k\|_2 \\ &= |\alpha| \cdot \|f_A(x) - c_k\|_2 = \alpha \cdot d_E(f_A(x), c_k) \end{aligned}$$

因此，集合 S_k（f_A）中的每个距离值都乘以 α，从而得到集合 S_k（f_B）。因此，这些距离的概率分布是原始分布的缩放版本。 □

A.2.2. 持久同调的稳定性保证

基于持久性的特征具有可迁移性，这直接源于拓扑数据分析的基本稳定性定理。

命题 A.7 （持久同调的不变性与稳定性）.

1. 等距不变性*(Edelsbrunner & Harer, 2010)*：设 $P \subset \mathbb{R}^n$ 是一个点云，且 g： $\mathbb{R}^n \to \mathbb{R}^n$ 是一个欧几里得等距变换（平移、旋转、反射）。那么，持久图保持不变：$Dgm(P) = Dgm(g(P))$。

2. 稳定性 (*Chazal etal.*，*2009年*；*Cohen-Steineret al.*，*2005年*)：设 X_A 和 X_B 是 $\mathbb{R}^n$ 中的两个点云。它们各自的持久图之间的瓶颈距离，有界于其度量空间之间的格罗莫夫-豪斯多夫距离：

$$d_B(Dgm(X_A), Dgm(X_B)) \leq d_{GH}((X_A, d_E), (X_B, d_E))$$

注： *(1)* 等距变换 g 保留所有两两欧几里得距离。由于 *Vietoris-Rips* 过滤仅基于这些距离构造，因此过滤 $\{VR(P,r)\}_{r\geq 0}$ 与 $\{VR(g(P)\ r)\}_{r\geq 0}$ 完全相同。对相同的过滤应用同调函子，会得到相同的持久图。*(2)* 该证明是拓扑数据分析中的一项基石性结果。它将这一直觉形式化：如果两个空间在度量上相似（d_{GH} 较小），那么由持久同调捕获的拓扑特征也必然相似（d_B 较小）。

A.3. 拓扑密度与样本记忆的几何

我们定性考察投影流形的类内密度与已有样本记忆概念之间的联系（Feldman，2020年；Feldman 和 Zhang，2020年），并通过输入曲率分数进行度量（Garg 等人，2024年）。我们的发现揭示了明确的对应关系：高密度、原型样本始终表现出低输入曲率（这是未被记忆样本的典型特征）；而低密度、非典型样本则表现出高输入曲率（这是记忆的一个关键指标）。这种一致性表明，我们的拓扑流形投影有效保留了区分原型样本与非典型样本的全局结构。



图6。 **原型样本：** 对于 CIFAR-100 中的五个类别，展示同一类别中曲率最低的前 10 个样本（左）与密度最高的样本（右）。

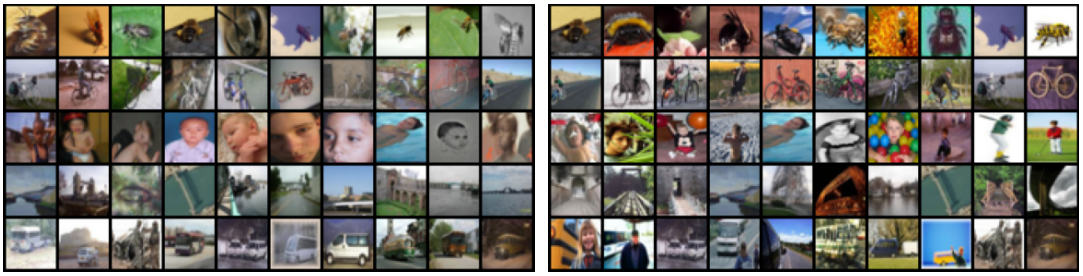


图7。 **非典型样本：** 对于 CIFAR-100 中的五个类别，展示同一类别中曲率最高的前 10 个样本（左）与密度最低的样本（右）。

A.4. 核心集构建示例

CIFAR-100 中“蝴蝶”类别在不同剪枝率（70%、50%、30%、20% 和 10%）下 TopoPrune 的定性可视化（见图 8）。可视化结果显示，在类别流形的局部区域内，持久性分数的方差较大，体现了该方法对细粒度局部结构的敏感性，以及区分相邻样本的能力。尽管侧重于局部复杂性，最终的核心集仍能保持密度，其整体分布与完整数据集非常接近。这表明，TopoPrune 成功平衡了对拓扑关键局部样本的选择与对全局数据结构的保留。

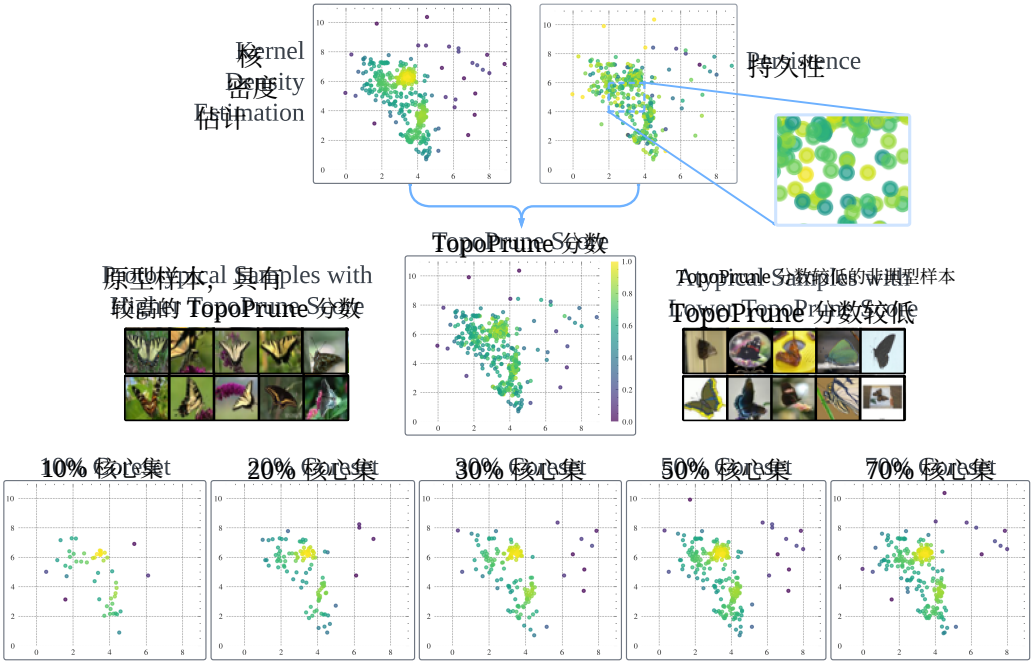


图*8.* CIFAR-100 中“蝴蝶”类别的 TopoPrune。

B. 实现与计算分析

B.1. 流形投影技术的比较分析

我们的全局流形投影对于在不同神经网络架构之间实现度量稳定性至关重要。高维嵌入的外在几何可能各不相同，但它们具有共同的内在拓扑。UMAP 利用这一共有结构构建新的低维流形，使其不仅忠实于拓扑，也实现了几何标准化。这一过程的关键结果是，所得低维嵌入在不同架构之间均能保持密度。这种标准化确保全局密度分数（样本重要性计算的核心组成部分）无论来源网络为何，都是稳定可靠的度量。

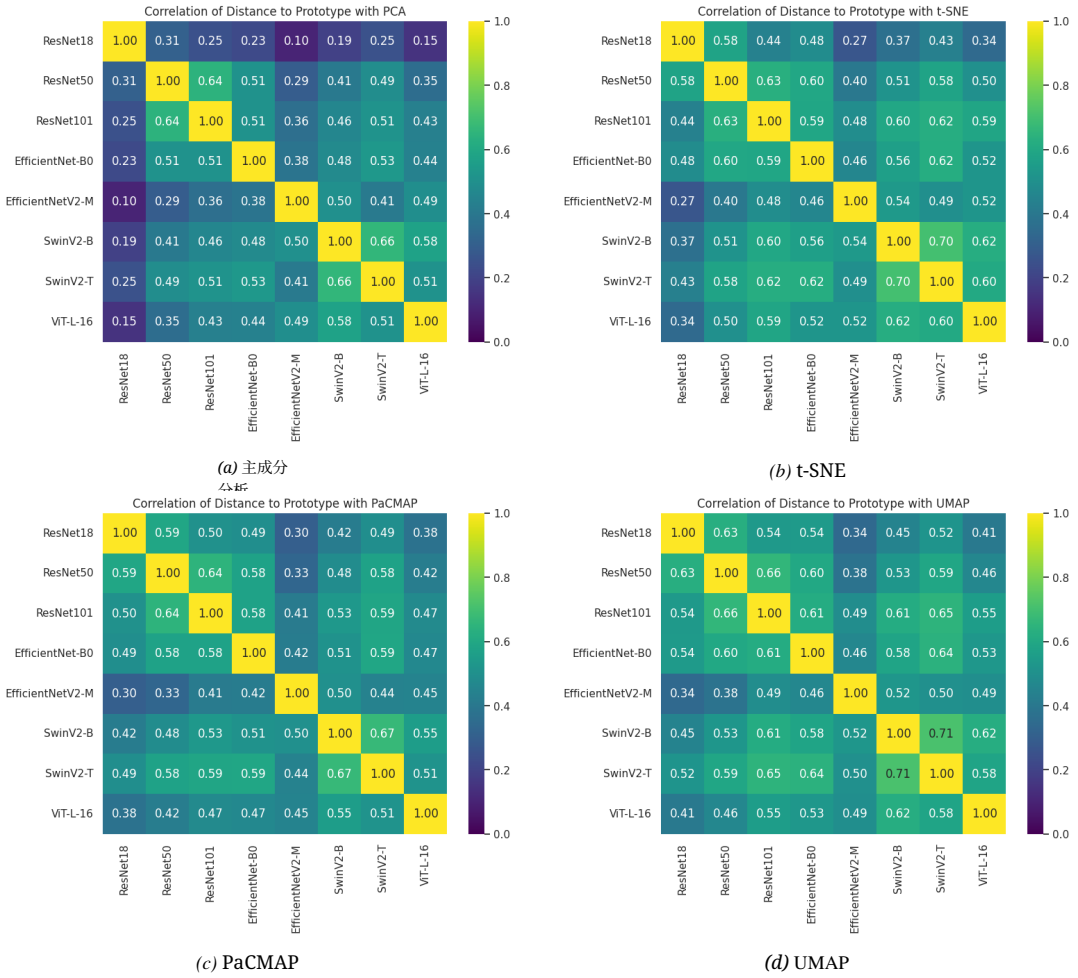
B.1.1. 标准线性与非线性流形近似的评估

为了进一步阐述如何在嵌入空间中的扰动下，标准化基于拓扑的流形近似与投影，我们考察了不同流形投影和特征降维技术下每个样本到原型的距离相关性（图 9）与分布（图 10）：(a) PCA（Pearson，1901年）、(b) t-SNE（van der Maaten & Hinton，2008年）、(c) PaCMAP（Wang 等人，2021）和 (d) UMAP（McInnes 等人，2018a）。我们发现，与线性 PCA 或更关注局部结构的 t-SNE 相比，基于拓扑的方法 UMAP 和 PaCMAP 的相关性显著更高，因此在不同架构间具有更好的可迁移性。值得注意的是，UMAP 的可迁移性略优于 PaCMAP，这进一步印证了我们在框架中选择 UMAP 的合理性。较小模型（例如 ResNet-18）与较大模型之间的高相关性尤其有价值，因为这验证了使用计算成本较低的网络生成流形嵌入的可行性，而这些嵌入仍能有效地用于规模大得多的模型上的数据选择。

B.1.2. 深度拓扑自编码器的评估

我们还考虑了拓扑自编码器（TopoAE）（Moor 等人，2020年）和正则化拓扑数据分析（RTD）（Trofimov 等人，2023年），用于流形嵌入。虽然这些方法在保留全局拓扑方面功能强大，但出于计算效率、领域适用性以及与核心集选择目标的一致性考虑，我们选择了 UMAP。

计算成本与实证验证。TopoPrune 旨在成为一种轻量级、免训练且适用于冻结特征的方法，而 UMAP 与这一要求完美契合。相比之下，训练拓扑自编码器作为预处理步骤的计算成本高得难以承受；在 CIFAR-100 上训练 RTD 需要 ∼8 小时，而 UMAP 只需 ∼22 秒，速度提升了



(a) 主成分分析　*(b)* t-SNE

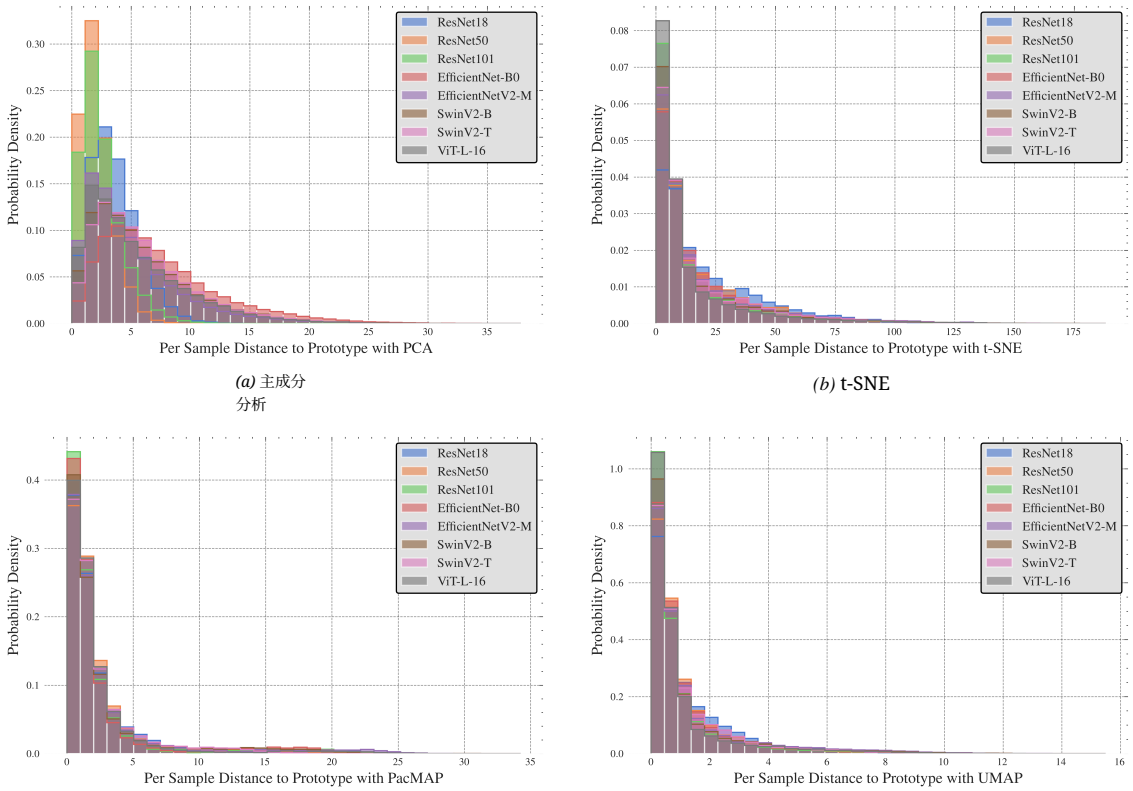
(c) PaCMAP　*(d)* UMAP

图*9.* 使用不同线性和非线性流形投影技术时，不同架构下每个样本到原型的距离相关性。

倍（见表3a）。为实证验证我们的选择，我们在 CIFAR-10/100 上分别训练了 TopoAE 和 RTD 模型，并在流程中用它们生成的嵌入替代 UMAP。如表3b所示，无需承担如此巨大的训练开销，UMAP 始终能取得更高或相当的准确率。

领域适用性与结构对齐。使用拓扑自编码器时，核心集准确率出现下降，这很可能源于目标之间根本性的不匹配。首先，就潜在空间质量而言，TopoAE 难以为复杂数据集生成清晰、彼此分离的表示。正如 TopoAE论文所指出的，CIFAR-10“很难以纯无监督的方式进行嵌入”（何恺明等人在Moor 等人（2020年）的第5.2.2节），这往往导致潜在空间中的类别混杂在一起，而非清晰分离（见Moor 等人（2020年）的图4）。这种缺乏分离性的情况严重削弱了逐类密度估计的效果，与 UMAP 实现的清晰簇划分形成鲜明对比。此外，TopoAE 和 RTD 等方法优先保留全局结构相似性（例如，维持图1中猛犸象“头部”和“脚部”簇之间的相对距离；见Trofimov 等人（2023年）），使这一问题更加突出。尽管这种全局约束对可视化很有价值，但对于核心集选择而言相关性较低，因为我们会将数据划分为基于类别的流形。

总而言之，尽管拓扑自编码器是用于流形学习的稳健工具，但对于我们特定的多尺度框架而言，UMAP 提供了更高效、更适合该领域且性能更优的基础。我们鼓励未来的研究探索能够灵活平衡全局与局部结构优先级的拓扑自编码器，并针对点云稀疏化任务进行专门优化。



(a) 主成分分析

(b) t-SNE

(c) PaCMAP

(d) UMAP

图*10*。 pe 的分布　　采用不同线性和非线性方法时，不同架构中的 r-样本到原型的距离　　r
流形投影技术　　es.

表*3.* **使用拓扑自编码器时的吞吐量和准确率。(a)** 与 UMAP 的算法投影相比，拓扑自编码器需要耗费大量成本进行训练。**(b)** 以 TopoAE 或 RTD 替代 UMAP 来进行全局流形嵌入会导致性能下降。

方法	CIFAR-10	CIFAR-100
TopoAE (Moor 等人，2020年)	14,847.79	15,248.11
RTD (Trofimov 等人，2023年)	26,622.89	28,661.29
UMAP (McInnes 等人，2018a)	**22.42**	**22.74**

*(a)*延迟（秒）

方法	CIFAR-10	CIFAR-100
TopoAE（Moor 等人，2020年）	75.0±0.3	41.2±1.0
RTD (Trofimov 等人, 2023年)	78.0±1.7	**46.4±0.4**
UMAP(McInnes 等人,2018a)	**82.1±0.3**	45.8±0.7

(b) 90% 剪枝率下的准确率（3 次运行的平均值）

B.2. TopoPrune 伪代码

算法1 使用 TopoPrune 进行核心集选择

输入： 数据集 $\mathcal{D} = \{(x_i,\ y_i)\}_{i=1}^{N}$、倒数第二层编码器 h_θ、剪枝率 p、误标比例 γ、权重 $\alpha,\ \beta$, 持久性优化步数 T
可调用项： UMAP、核密度估计，K近邻 ▷见超参数于表 13
输出： 选定索引 $\mathcal{S}$

双尺度拓扑评分

提取嵌入： $Z \leftarrow h_\theta(X)$
全局流形投影： $Y \leftarrow \text{UMAP}(Z)$
初始化评分向量 S_{pers} $S_{dens} \leftarrow \mathbf{0}^N$
对于每个类别 $c \in \{1, \ldots, C\}$ **执行**
  获取索引 $\mathcal{I}_c \leftarrow \{i \mid y_i = c\}$ 和子集 $Y_c \leftarrow Y[\mathcal{I}_c]$
  $S_{dens}[\mathcal{I}_c] \leftarrow$ 核密度估计(Y_c) ▷ 全局密度分数
  初始化可优化输入 $Y_c' \leftarrow Y_c$
  对于 $t = 1$ **到** T，**执行**
    计算希尔伯特分解有符号测度： μ^{Hil} $H_1(VR_{Y_c'}, \hat{f})$
    计算拓扑损失： $\mathcal{L}_{\text{pers}}(Y_c') \leftarrow \text{OT}(\mu^{Hil}$ $H_1(VR_{Y_c'}, \hat{f})}, \mathbf{0})$
    更新坐标： $Y_c' \leftarrow Y_c' + \eta \nabla_{Y_c'} \mathcal{L}_{pers}$
  结束循环
  $S_{pers}[\mathcal{I}_c] \leftarrow ||Y_c - Y_c'||_2$ ▷局部持久性得分
结束循环

误标检测与过滤

如果方法为 NLPS**则**
  对于每个样本 $i \in \{1, \ldots, N\}$ **执行**
    计算 k 最近邻： $\mathcal{N}_k(z_i) \leftarrow$ K近邻(Z)
    统计不匹配邻居： $m_i \leftarrow \sum_{j \in \mathcal{N}_k(z_i)} \mathbb{I}(y_j \neq y_i)$
    计算标签纯度比： $S^{(i)}$ $_{mis} \leftarrow m_i / k$ ▷ 免训练
  end for
否则如果方法为 AUM**则**
  获取预计算分数： $S_{mis} \leftarrow \text{AUM}(Z)$ ▷训练动态
结束 if
确定错误标注样本索引： $\mathcal{I}_{mis} \leftarrow \text{TopK}(S_{mis}, \gamma \cdot N)$
定义干净候选集： $\mathcal{D}_{clean} \leftarrow \mathcal{D} \setminus \mathcal{I}_{mis}$

基于统一分数的分层抽样

$S_{unified}^{(i)} \leftarrow \alpha \cdot S_{pers}^{(i)} + \beta \cdot S_{dens}^{(i)}$
$\mathcal{S} \leftarrow$ 分层抽样$(\mathcal{D}_{clean}, S_{unified}, p)$ ▷见Zheng 等人（2023年）中的算法1
返回 $\mathcal{S}$

B.3. 计算复杂度分析

相对于几何基线的理论复杂度。我们的分析遵循Nagaraj 等人（2025年）提出的核心集选择复杂度评估框架。TopoPrune 的主要计算开销来自局部拓扑优化。不过，我们通过利用多参数持续同调的高效约简，确保了计算的可处理性。

- 关键的是，我们采用希尔伯特分解有符号测度，它将多参数问题简化为网格上的单参数持久性切片。如Loiseaux等人(2023b的附录 D.1 所详述，对于网格大小为 + 的双参数滤流（Rips m），该算法执行 m 次单参数持久性优化。

- 因此，持久同调优化的成本相当于在 N_c 个点上进行单参数优化的成本。虽然持久性在理论上的最坏情况复杂度为立方级（Lesnick与Wright，2022），但在单参数持久性情形下，计算在实证中呈线性复杂度（Bauer等人，2022）。

- 此外，损失函数 $\mathcal{L}_{\text{pers}}$ (式（3）)的梯度计算可简化为对特征持续性求和（参见Scoccola等人(2024）中的推论 E.2）。该操作受一个常数 K 约束，该常数由单纯复形 $\mathcal{K}$导出，因此反向传播 $\mathcal{O}(K)$的复杂度为严格的 $\mathcal{O}(m \cdot N_c \log N_c)$。

表4. **基于几何的方法的计算复杂度分析**。 N 是数据集大小，其中包含 C 个类别、每类 $N_c \approx N/C$ 个样本、 d 维，以及 k 个邻居。对于 TopoPrune，开销主要由 T 优化步数和 m 网格分辨率决定。

	计算复杂度	说明
Moderate（Xia等人，2023年）	$\mathcal{O}(Nd) + C * \mathcal{O}(N_c \log N_c)$	距离计算。 (Nd) + 原型排序
D2 (Maharana等人，2024年)	$\mathcal{O}(Nkd) + \mathcal{O}(T \cdot Nk)$	k近邻图 (Nkd) + 消息传递
TopoPrune	$\mathcal{O}(N \log N) + C * \mathcal{O}(Tm \cdot N_c \log N_c)$	全局 UMAP + 局部持久性

实测墙钟时间分析。 我们在 AMD EPYC 7502（32 核）中央处理器上，对 TopoPrune 与基线方法的延迟和资源利用率进行了基准测试。尽管延迟较高，我们的分析显示，TopoPrune 对可用硬件的利用率明显不足（利用率为 16%，而基线方法为 82%）。这表明瓶颈源于实现本身，而非算法存在根本缺陷，因为我们的拓扑后端（multipers）目前尚不支持多线程处理和 GPU 加速。我们预计，随着拓扑库日趋成熟，这一墙钟时间差距将显著缩小。无论当前实现受到何种限制，TopoPrune 在高风险数据选择中都具有更优的成本效益特征，原因如下。(1) TopoPrune 比“训练动态型”核心集方法（如 Glister、遗忘）快几个数量级；这类方法需要从头训练代理模型（耗时数小时），而我们只需对冻结嵌入进行探测（耗时数分钟）。(2) 与快速几何启发式方法（如 D2、Moderate）不同，后者存在较高的方差且精确率较低（如表 1所示），TopoPrune 接受较高的前期计算成本，以保证获得稳定、高保真的核心集。这样就无需为了降低随机性而反复运行数据选择。

表5。 **延迟和利用率** 在 CIFAR-100 上执行筛选时的情况。为确保单线程和多线程实现之间的公平比较，延迟按中央处理器利用率进行归一化。

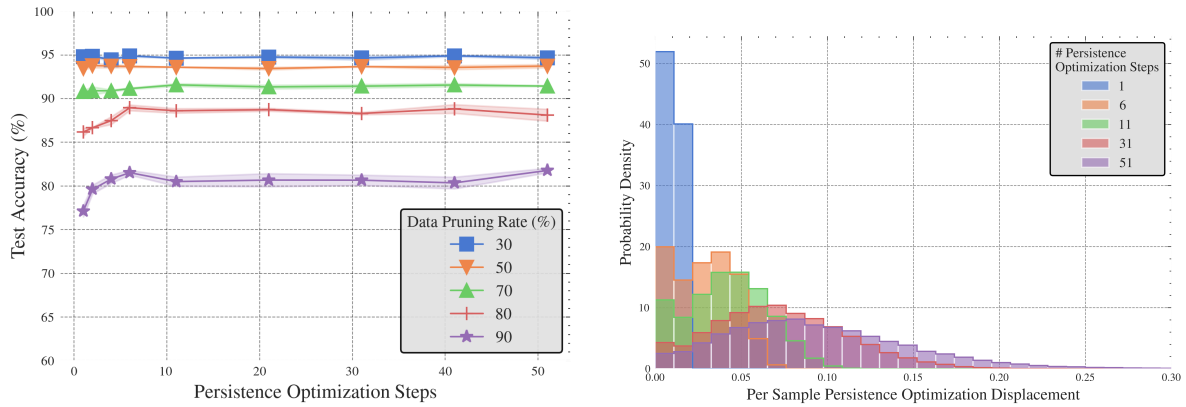
	全局（秒）	局部（秒）	总计（秒）	中央处理器最大利用率
D2	-	-	86.05	82%
TopoPrune	22.74	146.16	168.90	16%

TopoPrune之所以仍然易于处理，是因为它采用了一种构造，将复杂的二参数持久性问题归约为一系列 m 标准一参数计算。通过逐类别应用该优化（其中 N_c 相对较小）并利用这一高效归约策略，我们避免了多参数拓扑通常带来的高昂成本。随着拓扑软件基础设施的改进，我们预期TopoPrune能够以可忽略的延迟权衡，提供拓扑方法所具有的更优稳定性。

C. 消融研究

C.1. 可微持久性优化步骤

我们研究优化步数对多参数持久同调的影响（见图11）。所需的持久性优化步数与最终核心集大小呈负相关。在选择较大核心集时（例如剪枝率为30%），选择过程较为稳健，即使仅需少量优化步数（1-2）也足以识别出高质量子集。然而，在高剪枝率下（例如90%），区分最关键样本的任务变得更加敏感，因此需要更多优化步数（≥6），以使点的位置收敛，并准确揭示结构上最重要的样本。

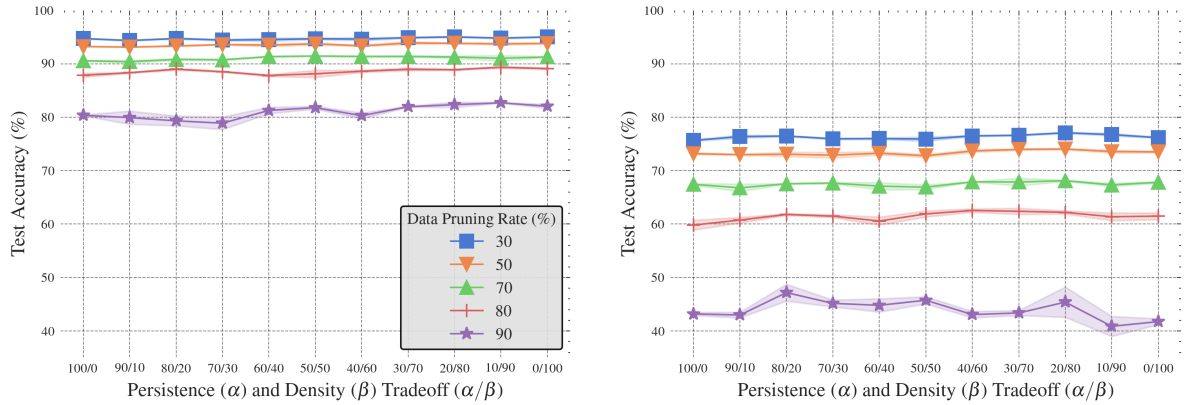


*(a)*CIFAR-10 上的准确率　　*(b)* 每样本位移的分布

图*11.* 核心集越小，容错空间就越小，因为每个入选样本的重要性会被放大。因此，需要更多优化步数，才能准确区分最关键的样本。相比之下，较大的核心集容错性更强，只需更少的优化步数便能取得高质量结果。

C.2. 敏感　性：局部持久性（α）与全局密　度（β）

我们研究了超参数 α 和 β 在公式 (5)中的影响，这些参数用于平衡全局密度和局部持久性分数的影响（见图 12）。分析表明，核心集质量在一系列(α, β) 取值范围内总体稳定，但同时结合两种度量始终能带来最佳表现。仅使用密度或持久性都能提供合理的基线表现，但在高剪枝率（例如 90%）下，同时结合两者尤为关键；与单独使用任一度量相比，平衡后的分数可将准确率提高最多 5.4%。这表明全局和局部拓扑对于实现最佳选择都至关重要，也证明了我们在所有实验中采用固定且平衡的配置$(50/50)$ 的合理性，从而减少了大量超参数调优的需要。



(a) CIFAR-10　　*(b)* CIFAR-100

图*12.* 拓扑超参数在全部数据剪枝率范围内的表现，涵盖**(a)** CIFAR-10与**(b)** CIFAR-100。

C.3. 用于错误标签检测的间隔下方面积（AUM）无需训练的代理指标

我们评估了多种免训练方法，将其作为边际面积（AUM）的代理指标。这些代理指标依据不同的几何准则识别可能含噪声的样本：例如，根据样本到类别原型的距离剔除样本，或利用相邻距离比值，移除那些更靠近另一类别原型的点。其他启发式方法包括剔除密度分数最低的样本，或使用我们提出的邻域标签纯度分数（*NLPS*），通过计算同标签最近邻所占比例（在 20 个最近邻中）来识别标签混杂区域中的点。结果表明，在所有无需训练的代理指标中，NLPS 的核心集准确率最高，且在数据剪枝率较高时优势尤为明显。虽然其表现未能完全达到使用真实 AUM 时的水平，但 NLPS 是一种简单有效的免训练代理指标。

表*6.* **间隔下方面积（AUM）的无需训练的代理指标** (Pleiss 等人，2020年) 在 CIFAR-100 上。我们看到，邻域标签纯度分数（NLPS）的表现最接近 AUM。

剪枝率（→）	30%	50%	70%	80%	90%
距离	75.4±0.4	71.3±0.3	63.2±0.1	56.1±0.6	37.8±0.3
相邻距离	75.5±0.3	71.8±0.2	62.8±1.0	57.1±0.3	38.1±1.5
密度	75.6±0.1	71.4±0.2	64.3±0.3	52.4±0.3	38.0±0.7
NLPS	**75.6±0.2**	**71.9±0.2**	**65.3±0.4**	**56.7±0.4**	**41.6±1.0**
AUM	75.9±0.4	72.8±0.3	66.9±0.5	61.9±0.6	45.7±0.7

C.4. 对 UMAP 流形投影超参数的敏感性

为验证 TopoPrune 全局流形嵌入阶段的稳健性，我们对 UMAP 进行了全面的超参数扫描。我们重点考察两个关键参数：(1) *n_邻居*，它控制局部与全局几何结构保留之间的平衡；(2) 最小值_距离，它决定样本在低维流形中的紧密程度。 表 7 中的结果表明，我们选择*n_邻居=15*和最小值_距离=*0.1*能够在高准确率和低方差之间取得良好平衡。

表 7。**TopoPrune 对 UMAP 超参数的敏感性。(a)** CIFAR-100 上的准确率（剪枝率为 90%），以及**(b)** 调节*n_邻居* 和最小值_距离 UMAP 超参数时，相对于默认配置的变化。

最小值 距离_	n_邻居 5	15	50	100	各行平均值
0.05	43.9±2.2	43.0±0.9	44.4±1.7	44.9±1.6	44.1±1.6
0.1	43.1±1.8	**45.8±0.7**	45.7±0.8	43.6±0.7	44.6±0.9
0.5	46.2±1.0	43.8±0.7	44.3±2.2	43.9±0.4	44.6±1.0
1.0	44.3±0.9	45.4±0.6	45.0±1.1	43.1±1.3	44.4±1.0
列平均值	44.4±1.5	44.5±0.7	44.8±1.4	43.9±1.0	-

*(a)*准确率（5次运行）

min_dist	n_邻居 5	15	50	100
0.05	-1.9	-2.8	-1.4	-0.9
0.1	-2.7	★	-0.1	-2.2
0.5	+0.4	-2.0	-1.5	-1.9
1.0	-1.5	-0.4	-0.8	-2.7

(b) 准确率 Δ

D. 扩展实证结果

D.1. 稳定性改进的统计显著性

为验证 TopoPrune 的稳定性优势（具体而言，即独立运行之间最终模型准确率的波动性降低），我们开展了统计检验，重点关注高剪枝率（例如 90%）的情形，因为此时核心集选择的方差通常最为显著。我们将 TopoPrune 与表现最佳的基线方法 D2（Maharana 等人，2024年）和 CCS（Zheng 等人，2023年）进行比较。我们的分析采用了两种互补的统计方法：(1) 单侧F检验，用于检验 TopoPrune 的方差是否低于基线方法（$\sigma^2_{\text{TopoPrune}} < \sigma^2_{\text{baseline}}$）；(2) 对最终准确率的标准差计算自助法95%置信区间，采用有放回抽样，重复抽样 10,000 次。

如表8所详述，在较简单的数据集上，各方法之间的方差差异在统计上并不显著。然而，对于更具挑战性的基准，TopoPrune 相较于 D2 和 CCS，展现出统计上显著的 p值（$p < 0.05$，以粗体标出），其方差更低。此外，标准差的自助法置信区间

TopoPrune 的标准差置信区间严格低于基线方法，在某些情况下甚至不重叠。这些发现证实，与以往的几何方法相比，我们的拓扑选择机制具有更优越的内在稳定性。

表8。 **方差的统计显著性** 在高剪枝率（90%）下，使用**(a, b, c)** F检验表明，TopoPrune在较难的数据集上的较低方差具有统计显著性。**(d)** TopoPrune更窄的标准差95%置信区间进一步验证了这一点。

p_值	CCS	D2	TopoPrune
CCS	0.500	∼	∼
D2	0.128	0.500	∼
TopoPrune	0.422	0.094	0.500

(a) CIFAR-10上的F检验

p_值	CCS	D2	TopoPrune
CCS	0.500	∼	∼
D2	0.437	0.500	∼
TopoPrune	**0.015**	**0.011**	0.500

(b) CIFAR-100 上的 F检验

p_值	CCS	D2	TopoPrune
CCS	0.500	∼	∼
D2	0.059	0.500	∼
TopoPrune	**0.012**	**0.001**	0.500

(c) ImageNet-1K 上的 F 检验

95%置信区间	CIFAR-10	CIFAR-100	ImageNet-1K
CCS	[0.05, 0.91]	[0.85, 3.10]	[0.07, 0.75]
D2	[0.38, 1.77]	[0.26, 2.69]	[0.3, 2.12]
TopoPrune	[0.08, 0.35]	[0.1, 0.87]	[0.04, 0.23]

(d) 标准差的自助法95%置信区间

D.2. 详细实验结果路线图

实验 / 消融主题	主要部分	完整数据表
基线与数据集上的总体性能	第4.2部分	表 1
可迁移性：多样化嵌入 →固定目标	第4.3节	表10
可迁移性固定嵌入 →多样化目标	第4.3节	表2
噪声特征嵌入的稳定性	第4.4节	表9
深度拓扑自编码器的评估	B.1.2 节	表3
计算复杂度	第B.3节	表4，表5
消融实验：可微持久性优化步骤	C.1节	表11
消融实验：（α 与 β）的敏感性	C.2 节	表12
消融实验：用于错误标签检测的无需训练的代理指标	C.3节	表6
消融实验：对 UMAP 超参数的敏感性	C.4 节	表 7
实现超参数	∼	表 13

表9。 **CIFAR-100 上特征嵌入噪声条件下的核心集性能。** 该图展示了三种几何方法在噪声（ϵ）和剪枝率不断增加时的平均准确率与标准差（基于 5 次运行）。Moderate 的精确率较高（方差较低），但准确率较低；D2 的准确率较高，但精确率较低（方差较高）；TopoPrune 则始终兼具较高的准确率和精确率。TopoPrune 的优势在剪枝率和噪声水平最高时最为明显，凸显了其鲁棒性。最佳准确率和标准差值以粗体显示。如果多个方法的值相同，则第二个值加下划线。

噪声（→）	$\epsilon \sim \mathcal{N}(0, 0.25\sigma)$				$\epsilon \sim \mathcal{N}(0, \sigma)$			
剪枝率 (→)	50%	70%	80%	90%	50%	70%	80%	90%
Moderate (Xia等人,2023年)	71.1±0.2	63.7±0.4	56.0±0.6	33.2±0.9	70.4±0.1	62.5±**0.4**	54.6±0.6	33.9±**0.2**
D2(Maharana等人,2024年)	72.8±**0.1**	**68.5**±0.7	**63.0**±0.5	44.4±1.5	73.0±**0.1**	67.7±1.2	62.3±0.8	40.2±2.0
TopoPrune	**73.5**±0.3	67.8±**0.1**	60.5±**0.2**	**45.4**±**0.8**	**73.3**±0.2	**68.0**±0.5	**62.7**±**0.3**	**45.5**±0.6

噪声（→）	$\epsilon \sim \mathcal{N}(0, 4\sigma)$				$\epsilon \sim \mathcal{N}(0, 8\sigma)$			
剪枝率（→）	50%	70%	80%	90%	50%	70%	80%	90%
Moderate(Xia等人，2023年)	71.0±0.4	62.9±0.3	52.3±0.6	32.0±1.2	70.9±0.2	63.4±0.2	51.6±2.2	32.1±1.4
D2(Maharana等人，2024年)	72.3±0.1	**67.8**±1.1	**62.1**±1.0	40.5±1.7	71.6±0.5	67.2±0.8	60.3±2.4	39.8±3.2
TopoPrune	**73.4**±**0.1**	67.4±**0.2**	61.7±**0.3**	**46.1**±**0.7**	**73.4**±**0.2**	**67.2**±**0.1**	**61.8**±**0.1**	**43.9**±**0.4**

表*10*。 **多样化嵌入的**可迁移性 → 固定目标。利用不同架构的特征在 CIFAR-100 上训练 ResNet-18 模型。大多数模型取自 `torchvision` 预训练库，并基于 ImageNet-1K 进行微调。我们还研究了在 LAOIN-2b 数据集上训练的更大型 OpenCLIP 基础模型的特征可迁移性（Schuhmann 等人，2022）。

剪枝率（→）	50%			70%		
	Moderate	D2	**TopoPrune**	Moderate	D2	**TopoPrune**
ResNet-18（He 等人，2016）	70.9±0.4	73.0±0.8	73.6±0.2	62.9±0.2	67.9±0.3	68.1±0.2
ResNet-50（He 等人，2016）	71.1±0.1	73.0±0.2	73.7±0.2	63.3±0.4	67.7±0.4	68.0±0.1
ResNet-101(He 等人, 2016)	70.0±0.5	73.2±0.2	73.5±0.2	62.8±0.4	66.9±0.9	68.0±0.3
EfficientNet-B0（Tan 和 Le，2019年）	71.6±0.3	73.2±0.2	72.9±0.2	62.9±0.1	67.5±0.4	67.5±0.2
EfficientNetV2-M(Tan 和 Le，2021)	69.6±0.3	72.8±0.7	73.4±0.2	61.0±0.2	67.1±0.8	67.1±0.1
SwinV2-T（Liu 等人，2022）	70.5±0.1	73.6±0.1	73.6±0.2	60.4±0.6	66.9±1.0	67.6±0.1
SwinV2-B（Liu 等人，2022）	69.9±0.3	73.5±0.4	73.6±0.2	61.9±0.4	67.1±0.4	68.1±0.2
ViT-L-16(Dosovitskiy 等人，2021)	69.8±0.3	73.3±0.3	73.6±0.1	61.5±0.3	67.2±0.6	68.0±0.1
OpenCLIP ViT-L-14 (Radford 等,2021)	71.2±0.4	73.2±0.2	73.3±0.1	63.5±0.4	66.8±0.3	67.9±0.3
OpenCLIP ViT-H-14 (Radford 等，2021) (Schuhmann 等人，2022)	70.9±0.3	73.0±0.2	73.1±0.4	62.4±0.3	66.6±1.3	67.7±0.2
总体平均值	70.6±0.3	73.2±0.3	**73.4±0.2**	62.3±0.3	67.2±0.7	**67.8±0.2**

剪枝率（→）	80%			90%		
	Moderate	D2	**TopoPrune**	Moderate	D2	**TopoPrune**
ResNet-18（He 等人，2016）	54.8±0.2	60.3±1.9	60.2±0.2	33.8±0.8	42.5±1.9	43.4±0.4
ResNet-50（He 等人，2016）	55.9±0.6	60.8±1.0	61.3±0.7	31.8±1.6	44.5±1.7	47.4±0.5
ResNet-101（He 等人，2016）	54.5±0.4	60.4±0.2	60.0±0.6	35.9±0.6	41.7±2.6	43.2±1.3
EfficientNet-B0（Tan 和 Le，2019年）	54.8±1.1	60.5±0.8	62.0±0.4	29.8±1.3	42.0±2.2	42.2±0.3
EfficientNetV2-M(Tan 和 Le, 2021)	53.1±0.5	60.2±0.4	60.1±1.6	33.3±0.8	41.3±1.9	44.4±1.7
SwinV2-T(Liu 等人,2022)	53.3±1.2	59.3±2.0	61.8±0.4	32.5±1.5	41.4±2.2	43.4±0.8
SwinV2-B(Liu 等人，2022)	53.5±0.2	60.2±1.5	61.1±0.6	35.7±0.3	42.8±2.9	42.7±1.6
ViT-L-16(Dosovitskiy 等人,2021)	53.9±0.9	59.1±1.1	61.1±0.2	30.6±1.2	40.9±2.6	44.1±1.3
OpenCLIP ViT-L-14 (Radford 等，2021)	53.9±0.8	60.4±1.0	61.7±0.8	35.0±1.5	40.7±2.9	42.5±1.6
OpenCLIP ViT-H-14 (Radford 等, 2021) (Schuhmann 等人, 2022)	54.0±0.8	61.1±0.4	61.8±0.8	35.4±1.4	42.3±2.1	42.9±1.0
总体平均值	54.2±0.7	60.2±1.0	**61.1±0.6**	33.4±1.1	42.0±2.2	**43.6±1.1**

表*11.* **CIFAR-10 上持久同调的优化步数**。最优结果以粗体标出，并列结果加下划线。

	剪枝率（%）				
# 步数	30%	50%	70%	80%	90%
1	94.8±0.1	93.5±0.2	90.8±0.2	86.2±0.4	77.1±0.6
2	94.9±0.1	**93.8±0.3**	90.9±0.5	86.7±0.1	79.6±0.5
4	94.5±0.2	93.7±0.2	90.9±0.1	87.5±0.5	80.8±0.5
6	**94.9±0.1**	93.7±0.1	91.2±0.1	**89.0±0.3**	81.5±0.3
11	94.6±0.1	93.6±0.1	**91.6±0.2**	88.6±0.3	80.5±0.5
21	94.8±0.1	93.5±0.2	91.3±0.2	88.7±0.2	80.7±0.8
31	94.7±0.2	93.7±0.1	91.4±0.2	88.3±0.1	80.7±0.6
41	94.9±0.1	93.6±0.2	91.6±0.2	88.8±0.5	80.4±0.7
51	94.7±0.2	93.7±0.2	91.4±0.1	88.1±0.7	**81.8±0.3**

表*12*。 **局部持久性（α）和全局密度（β）的超参数**。 虽然固定的50/50划分带来了出色且稳定的性能，但结果表明，通过任务特定调优仍有可能进一步提高准确率。我们观察到一种趋势：在更具挑战性的数据集（例如 CIFAR-100 相较于 CIFAR-10）上，以及在更高的数据剪枝率下，最优平衡越来越依赖持久性分数（α）。最优结果以粗体显示，并列结果以下划线标出。

CIFAR-10	剪枝比例（%）				
α/β	30%	50%	70%	80%	90%
100/0	94.7±0.1	93.2±0.1	90.6±0.1	87.9±0.4	80.4±0.1
90/10	94.3±0.1	93.1±0.1	90.4±0.3	88.3±0.1	79.9±1.2
80/20	94.7±0.1	93.3±0.1	90.8±0.2	89.0±0.2	79.3±0.9
70/30	94.4±0.2	93.6±0.1	90.7±0.2	88.5±0.1	78.9±1.2
60/40	94.5±0.3	93.5±0.3	91.3±0.1	87.8±0.2	81.3±0.6
50/50	94.7±0.2	93.7±0.2	**91.6±0.1**	88.7±0.7	82.1±0.3
40/60	94.6±0.3	93.3±0.1	91.4±0.2	88.6±0.2	80.3±0.5
30/70	94.9±0.1	**93.9±0.3**	91.4±0.2	89.0±0.3	82.0±0.1
20/80	**95.0±0.1**	93.8±0.1	91.3±0.2	88.9±0.1	82.3±0.5
10/90	94.8±0.1	93.7±0.2	91.0±0.5	**89.3±0.2**	**82.7±0.1**
0/100	94.9±0.1	93.8±0.1	91.3±0.1	89.1±0.2	82.2±0.4

CIFAR-100	剪枝比例 (%)				
α/β	30%	50%	70%	80%	90%
100/0	75.7±0.2	73.2±0.2	67.4±0.2	59.8±0.9	43.2±0.3
90/10	76.4±0.3	73.0±0.1	66.8±0.7	60.8±0.5	43.0±0.5
80/20	76.5±0.1	73.1±0.4	67.5±0.1	61.8±0.2	**47.2±1.6**
70/30	76.0±0.2	72.9±0.6	67.7±0.2	61.5±0.3	45.2±0.7
60/40	76.0±0.2	73.3±0.3	67.1±0.7	60.6±0.8	44.8±1.2
50/50	75.9±0.4	72.8±0.3	66.9±0.5	61.9±0.6	45.8±0.7
40/60	76.5±0.1	73.7±0.3	67.9±0.1	**62.5±0.3**	43.1±0.6
30/70	76.6±0.1	**74.0±0.2**	67.9±0.7	62.4±0.6	43.4±0.5
20/80	**77.1±0.1**	74.0±0.2	**68.1±0.1**	62.2±0.3	45.4±2.8
10/90	76.7±0.2	73.6±0.3	67.3±0.3	61.4±0.7	40.9±1.9
0/100	76.2±0.1	73.5±0.2	67.8±0.2	61.5±0.7	41.8±0.6

表 *13.* **(a)**训练和拓扑超参数。**(b)** 数据集错误标注比例。

部分	超参数	CIFAR-10 CIFAR-100	ImageNet
训练 (DeepCore)	轮数	200	60
	批大小	256	128
	优化器	SGD	SGD
	动量	0.9	0.9
	学习率	1e-1	1e-1
	权重衰减	5e-4	5e-4
	调度器	余弦退火	余弦退火
全局流形 投影（UMAP）	邻居数量	15	15
	最小距离	0.1	0.1
	度量	余弦	余弦
	维度	2	2
核密度 估计 （sklearn）	带宽	0.4	0.4
	Theta（密度带宽）	0.4	0.4
局部持久 同调 (multipers)	函数/核	高斯	高斯
	复形	弱Delaunay	弱Delaunay
	同调次数	1	1
	优化步数	6	6
拓扑评分	全局密度（α）	0.5	0.5
	局部持久性（β）	0.5	0.5
NLPS （K近邻 sklearn）	邻居数量	20	20

*(a)*训练和拓扑超参数。

剪枝	误标比例（%）		
	C-10	C-100	ImageNet
30%	0%	10%	0%
50%	0%	20%	10%
70%	10%	20%	20%
80%	10%	40%	20%
90%	30%	50%	30%

(b) 数据集错误标注比例。与Zheng 等人(2023年)中的类似。