arXiv:2304.01222v1 [cs.HC] 1 Apr 2023

NEURODAVIS：一种用于数据可视化的神经网络模型

Chayan Maitra[1], Dibyendu B. Seal[2], 和 Rajat K. De[1],*

[1]印度统计研究所机器智能部，印度加尔各答巴拉克普尔干道 *203* 号，邮编 *700108*。[2]*Tatras Data Services Pvt. Ltd.*，印度新德里 *Vasant Vihar*，*Vasant Marg E64*，邮编 *110057*。*通讯作者：*Rajat K. De*，*rajat@isical.ac.in*

2023年4月5日

摘要

长期以来，高维数据集的降维与可视化一直是一项具有挑战性的问题。现代高通量技术产生了更新的高维数据集，这些数据集包含多个视图，并采用了相对新颖的数据类型。对这些数据集进行可视化，需要适当的方法，在不影响局部和全局结构的前提下揭示数据中的隐藏模式，并展现数据固有的非线性。然而，能够实现这一目标的方法仍然寥寥无几。在本研究中，我们提出了一种名为 NeuroDAVIS 的新型无监督深度神经网络模型，用于数据可视化。NeuroDAVIS 能够在不假设任何数据分布的情况下提取数据的重要特征，并在低维空间中有效地进行可视化。理论上已证明，高维空间中的数据邻域关系在低维空间中仍能得到保留。我们在各种合成和真实高维数据集上评估了 NeuroDAVIS 的性能，这些数据集涵盖数值型、文本、图像和生物数据。就可视化质量以及对数据规模、形状和局部与全局结构的保留而言，NeuroDAVIS 与 t-分布随机邻居嵌入（t-SNE）和统一流形近似与投影（UMAP）相比具有很强的竞争力。在大多数高维数据集上，它的表现也优于 t-SNE 的一种变体——基于快速插值的 t-SNE（Fit-SNE）。对于生物数据集，除 t-SNE、UMAP 和 Fit-SNE 外，与其他最先进算法相比，NeuroDAVIS 的表现也很出色，例如基于亲和力的轨迹嵌入的热扩散势（PHATE）以及基于孪生神经网络的方法 IVIS。下游分类和聚类分析也表明，NeuroDAVIS 生成的嵌入效果良好。

关键词：深度学习、无监督学习、形状保持、全局结构保持、单细胞转录组学

1 引言

基于机器学习的真实世界大型数据集分析，依托于降维（DR）方法，这些方法构成了数据集预处理和可视化的基础。常用的降维技术包括主成分分析（PCA） [1], 独立成分分析（ICA） [2], 多维尺度分析（MDS） [3], Isomap [4], NMF [5], SVD [6], t-分布随机邻居嵌入（t-SNE）[7] 和统一流形近似与投影（UMAP） [8]。PCA 将数据投影到一个新空间中，该空间由表示最大方差的向量张成；ICA 则从多变量信号中提取子成分。MDS 是一种降维和可视化技术，可提取数据结构中的差异性。Isomap 是一种非线性降维方法，结合了多种方法的优势。NMF 和 SVD 都是矩阵分解方法，广泛用于主题建模和信号处理。尽管 t-SNE 和 UMAP

是降维技术，但主要适用于可视化任务。上述方法都属于无监督降维。相反，线性判别分析（LDA） [9], 是一种用于降维和模式识别的监督方法。

任何降维技术都应兼顾保持数据的局部和全局结构。然而，我们更常见到的是方法偏重其中一方。主成分分析或 MDS 倾向于保留所有观测值之间的成对距离，而 t-SNE、Isomap 和 UMAP 则倾向于保留局部距离而非全局距离。对于上述其他降维技术在局部和全局形状保持方面的表现，相关研究仍然有限。

不过，t-SNE 这类非线性投影算法在提取数据中的复杂潜在结构方面，优于主成分分析这类线性算法 [10]。尽管如此，t-SNE 也有显著缺点。t-SNE 对噪声高度敏感，甚至可能将随机分布的点转换成伪聚类 [11]。此外，t-SNE 不擅长在保留全局距离的同时保留局部距离 [12, 13], , 这会妨碍我们从 t-SNE 可视化中得出符合实际的结论。基于快速插值的 t-SNE (Fit-SNE) [14] 是 t-SNE 的一种变体，可在大型数据集上加快处理速度。然而，要通过调整参数获得最佳嵌入，需要具备丰富的专业经验。另一方面，UMAP 使用流形学习技术降低数据维度，并且能够很好地扩展到大型数据集。不过，UMAP 假设数据中存在流形结构，而这并不总是符合实际情况。与 t-SNE 类似，它也优先保留局部关系而非全局关系 [8], 。基于亲和力的轨迹嵌入的热扩散势 (PHATE) [15] 是另一种基于流形学习的方法，近年来在生物学领域得到了广泛应用。不过，PHATE 在其他真实世界数据集上的有效性仍有待评估。

近十年来，基于神经网络（NN）的方法也一直被用作几乎所有领域中的非线性降维（DR）工具 [16, 17, 18]。无监督神经网络经过训练以学习非线性函数，而从具有相对较少节点的中间隐藏层中提取的特征，则作为数据的低维表示。神经网络方法最早用于投影可追溯至自组织映射（SOM），也称为 Kohonen 映射 [19]。SOM 是一种无监督方法，可在保留拓扑结构的同时将数据投影到更低维空间。近年来，基于自编码器（AE）的方法在降维领域占据主导地位 [20, 21, 22] ，人们已开发出多种 AE 变体，以解决包括计算机视觉和计算生物学在内的多个应用领域中的“维数灾难”问题。最近，一种采用三元组损失函数的孪生神经网络架构——称为 IVIS [23], ——已被开发用于高维生物数据集的数据可视化。所有这些基于神经网络的方法在降维和可视化任务方面都展现出巨大潜力。然而，并非所有方法都能同时保留数据的局部和全局结构。

受神经网络模型捕捉数据非线性的巨大潜力所吸引，我们在本文中提出了一种新颖的无监督深度学习模型，称为 NeuroDAVIS。该模型用于数据可视化，同时兼顾局部与全局结构保持。本文的主要贡献如下：NeuroDAVIS 是一种通用神经网络模型，可用于高维数据集的降维与可视化。它能够从数据中提取有意义的嵌入，从而捕捉表征数据内在非线性的重要特征。NeuroDAVIS 提取的特征还可用于数据重构。此外，NeuroDAVIS 不对数据分布作任何假设。我们在多种二维合成数据集和真实高维数据集上评估了 NeuroDAVIS 的性能，这些数据集涵盖数值型、文本、图像和生物数据。尽管上文讨论了 t-SNE 和 UMAP 的诸多局限，它们仍是当前降维或可视化方法中的先进代表。NeuroDAVIS 的表现可与 t-SNE 和 UMAP 相媲美，而且可以说，它比这两种方法更好地保留了数据形状、大小以及局部与全局关系。数学证明表明，高维空间中的局部邻居所对应的嵌入在低维空间中仍然彼此接近。NeuroDAVIS 适用于各种类型（数据模态）的数据集。此外，它无需任何预处理即可独立生成出色且可解释的可视化结果。

本文其余部分安排如下。第2节探讨 NeuroDAVIS 背后的动机并介绍其架构，随后在正确性证明（第3节）中对 NeuroDAVIS 的若干性质进行了数学论证。第4节介绍实验结果，包括对二维数据集的比较（第4.1节）、其嵌入、可视化与结构保持（第4.1.1节）、全局结构保持（第4.1.2节）、簇间距离保持（第4.1.3节）、簇大小保持（第4.1.4节），以及高维数据集的结果（第4.2节），其中包括数值型数据集（第4.2.1节）、文本数据集（第4.2.2节）、图像数据集（第4.2.3节）和生物数据集（第4.2.4节）。第5节讨论 NeuroDAVIS 的优缺点，并总结全文。

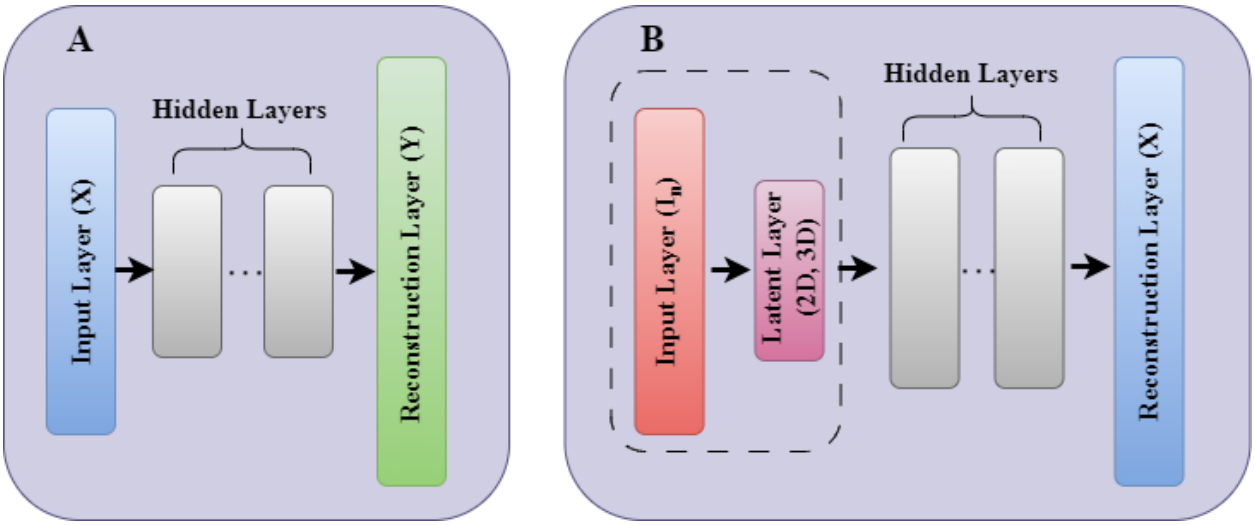


图 1：多层神经网络（A）和 NeuroDAVIS（B）的框图。（B）中的虚线部分表示经过修改的架构，使回归器能够得到更新。

2 方法

本节介绍所提出的无监督神经网络模型 NeuroDAVIS，该模型用于高维数据集的可视化。NeuroDAVIS 是一种非循环前馈神经网络，能够提取低维嵌入，以捕捉相关特征并实现高维数据集的高效可视化。它能够保持数据中各簇之间的局部与全局关系。下文将介绍该模型背后的动机，随后阐述其架构。

2.1 动机

保留局部与全局形状是可视化面临的两大主要挑战。迄今已有多种方法被提出，但没有一种能够妥善解决这两个问题。大多数方法试图保留局部形状而非全局形状，而另一些技术则做到了相反。因此，开发一种能将数据映射到低维空间，同时成功保留局部与全局距离的方法，是我们的主要目标。

NeuroDAVIS 背后的基本数学概念源自广为研究的回归问题。在回归问题中，使用一组自变量（回归器）来预测一个或多个因变量。与回归不同，可视化问题中只有一份数据。为了可视化手头的数据，可以通过无监督方式学习一组随机回归器。这种放宽可能使回归任务等同于可视化任务。

通过回归进行可视化

设 $\mathbf{X}_1 = \{\mathbf{x}1,_i : \mathbf{x}1,_i \in \mathbb{R}^{d_1}\}^n{}_{i=1}$ 和 $\mathbf{X}_2 = \{\mathbf{x}2,_i : \mathbf{x}2,_i \in \mathbb{R}^{d_2}\}^n{}_{i=1}$ 是两个数据集，分别包含 n 个样本，其特征维数分别为 d_1 和 d_2。回归的任务是构造一个从 $\mathbb{R}^{d_1}$ 到 $\mathbb{R}^{d_2}$ 的连续函数 f，使重构损失（通常为 $\sum^n{}_{i=1} \|f(\mathbf{x}1,_i) - \mathbf{x}2,_i\|^2$）最小化。为此，可以使用多层神经网络模型。为确保输入数据能恰当地输入模型，输入层应包含 d_1 个神经元。随后，输入会经过若干非线性激活层处理，最终由包含 d_2 个神经元的输出层输出。然后根据预期输出计算重构损失，并使用传统的反向传播学习更新权重和偏置。这种学习方式的关键在于，它不支持更新回归器。因此，多层神经网络能够学习到足够复杂的函数，从而生成接近真实值的预测。

在所提出的方法中，首先准备随机的二维或三维数据，以对原始数据进行回归。现在，存在多个目标，并作出了一些放宽。在整个学习过程中，不仅需要学习一个能够有效生成数据的合适连续函数，还需要学习回归器。所学习函数的连续性

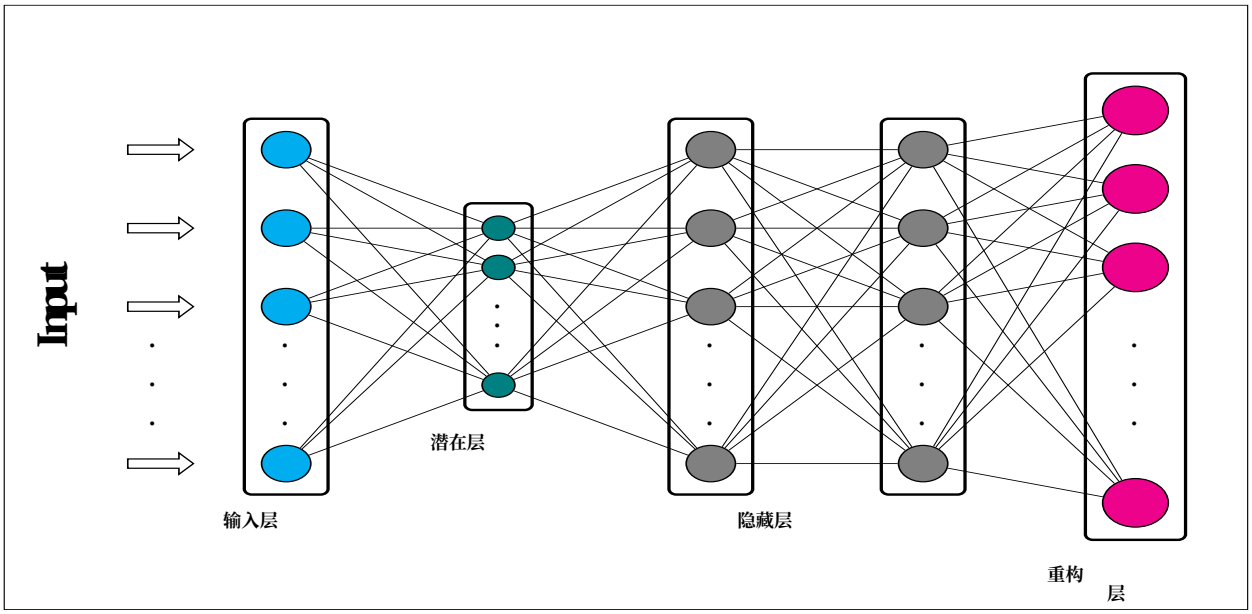


图2：NeuroDAVIS 架构。

及其较低的复杂度确保高维空间中的局部邻居在投影到低维空间时得以保留（见第3节中的定理1）。

如 NeuroDAVIS 的框图（图1B）所示，虚线部分用于控制回归器。NeuroDAVIS 的输入是一个大小为 $n \times n$ 的单位矩阵，其中 n 表示样本数量。在前向传播过程中，单位矩阵的每一列都会馈入输入层，从而确保只有特定的权重值（随机初始化）作为输入馈入潜在层。利用前两层生成的这些随机点用作回归器。与多层神经网络类似，NeuroDAVIS 也会尝试在重构层重构数据。随后计算重构损失，并使用标准反向传播算法更新相应的权重和偏置。因此，在每个训练轮次中，随着第一层的权重值更新，回归器也会随之调整，使它们与各自对应的预测保持等距。详细的学习过程已在第2.4节中说明。

2.2 架构

如图 2 所示，NeuroDAVIS 的架构是一种新颖的神经网络模型。NeuroDAVIS 网络由不同类型的层组成，即输入层、潜在层、一个或多个隐藏层 以及重构层。设$\mathbf{X} = \{\mathbf{x}_i : \mathbf{x}_i \in \mathbb{R}^d\}^n{}_{i=1}$ 为一个包含n 个样本的数据集，每个样本由d 个特征表征。那么，NeuroDAVIS 中的输入层 包含n 个神经元。潜在层 中的神经元数量为k，其中k表示用于可视化的维度数。换言之，k 表示提取低维嵌入时所采用的维度数。通常，在实际应用中，可以进行二维或三维可视化。输入层用于在潜在层创建n个样本（观测值）的随机低维嵌入，该嵌入可视为n 个原始观测值的初始代表。潜在层 中的低维嵌入通过一个或多个隐藏层投影到重构层 。重构层 尝试根据n个样本的随机低维嵌入，重构其原始d 维空间。因此，重构层 中的神经元数量为d。在本研究中，我们仅使用了两个隐藏层。隐藏层 尝试捕捉数据中的非线性，并将所学信息传递给序列中的下一层。因此，可以根据数据的复杂程度使用多个隐藏层 。不过，为避免过拟合，隐藏层的数量不宜过多。

2.3 前向传播

如前所述，我们考虑一组n 个样本$\mathbf{X} = \{\mathbf{x}_i : \mathbf{x}_i \in \mathbb{R}^d\}^n{}_{i=1}$。NeuroDAVIS 的输入是一个$\mathbf{I}$阶$n \times n$单位矩阵。对于第$i$个样本$\mathbf{x}_i$，我们考虑$I$的第$i$ 列向量$\mathbf{e}_i$。也就是说，当输入

$\mathbf{e}_i$到输入层后，重构层会重构出 $\mathbf{x}_i$ 的近似版本 $\tilde{\mathbf{x}}_i$。令 $\mathbf{a}^{ji}$ 和 $\mathbf{h}^{ji}$分别表示输入第 i^{th} 个样本时第 j 层的输入和输出；$\mathbf{W}^j$为第 $(j-1)^{th}$ 层与第 j 层之间的权重矩阵（$j=1,2,\cdots,(l+2)$）；$\mathbf{b}^j$为第 j 层节点的偏置项。第 j 层可以是输入层（$j=0$）、潜在层（$j=1$）、隐藏层（$j=2,3,...,(l+1)$）或重构层（$j=(l+2)$）。因此，对于输入层，

$$\begin{cases}\mathbf{a}_{0i}=\mathbf{e}_i,\\ \mathbf{h}_{0i}=\mathbf{e}_i,\end{cases}\qquad \forall i=1,2,\cdots,n \tag{1}$$

对于潜在层，我们有

$$\begin{cases}\mathbf{a}_{1i}=\mathbf{W}_1\mathbf{e}_i+\mathbf{b}_1,\\ \mathbf{h}_{1i}=\mathbf{a}_{1i},\end{cases}\qquad \forall i=1,2,\cdots,n \tag{2}$$

此处，$\mathbf{e}_i$控制在潜在层获得的i^{th}样本的低维嵌入的权重参数。它确保在呈现i^{th}样本时，只有连接到输入层 的i^{th}神经元的连接才会激活潜在层 中的神经元。对于l 隐藏层，我们有

$$\begin{cases}\mathbf{a}_{ji}=\mathbf{W}_j\mathbf{h}_{(j-1)i}+\mathbf{b}_j,\\ \mathbf{h}_{ji}=ReLU(\mathbf{a}_{ji}),\end{cases}\qquad \forall j=2,3,\cdots,(l+1),\forall i=1,2,\cdots,n \tag{3}$$

其中$ReLU(\mathbf{y})$ = 最大值$(\mathbf{0},\mathbf{y})$；最大值 （最大值）按元素计算。

在称为重构层的最终层执行原始数据的重构。对于重构层，我们有

$$\begin{cases}\mathbf{a}_{(l+2)i}=\mathbf{W}_{l+2}\mathbf{h}_{(l+1)i}+\mathbf{b}_{l+2},\\ \mathbf{h}_{(l+2)i}=\mathbf{a}_{(l+2)i},\end{cases}\qquad \forall i=1,2,\cdots,n \tag{4}$$

因此，NeuroDAVIS 将在潜在层 获得的样本的潜在嵌入，通过隐藏层投影到d维空间中。

2.4 学习

如前所述，NeuroDAVIS 已用于高维数据的降维和可视化。对于每个样本 $\mathbf{x}$，NeuroDAVIS 都会通过最小化重构误差 $\|\mathbf{x}-\tilde{\mathbf{x}}\|$，尝试找到输入数据的最优重构 $\tilde{\mathbf{x}}$。为避免过拟合，节点活动和边权重采用了$L2$ 正则化。使用$L2$ 正则化可确保模型复杂度最小化，也有助于将输出值和权重值拉向零。节点活动和权重的正则化分别通过正则化参数 α 和 β 控制。因此，目标函数变为

$$\mathscr{L}_{NeuroDAVIS}=\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\tilde{\mathbf{x}}_i\|^2+\alpha\sum_{j=1}^{l+1}\sum_{i=1}^{n}\|\mathbf{h}_{ji}\|_2+\beta\sum_{j=1}^{l+1}\|\mathbf{W}_j\|_F \tag{5}$$

NeuroDAVIS 使用 Adam 优化器 [24]进行训练。 α 和 β 的值针对每个数据集通过实验设定。此外，还仔细观察了$\mathscr{L}_{NeuroDAVIS}$达到饱和时的训练轮数，以确定收敛所需的最佳训练轮数。

在前向传播开始时，输入层 与潜在层 之间的权重值会随机初始化。NeuroDAVIS 网络的学习由一个作为输入馈入输入层的单位矩阵$\mathbf{I}$ 控制。使用单位矩阵可确保仅更新与当前批次中的样本相关联的权重值。因此，最初会在潜在层创建由n 个样本（观测值）构成的随机低维嵌入。随后，NeuroDAVIS 会将潜在层 中的这一低维表示投影到重构层 。在重构层，使用公式5计算误差值。接着，该误差会反向传播，并在下一次前向传播中更新权重和偏置值，以便更好地重构当前批次中的样本。此过程会持续进行，直至收敛。训练阶段完成后，从潜在层 提取的变换后的特征集将作为数据的低维嵌入，并用于可视化$k=2$ 或3。

3 正确性证明

本节从数学上证明了，高维空间中的邻居在投影到低维空间后仍能保持不变。

引理1。对于　任意弗罗贝尼乌斯范数小于或等于 *1* 的实矩阵$\mathbf{W}$，以及 $\eta \leq 1, \|\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T\|_2 \leq 1$　　.

证明。众所周知，

$$\|\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T\|_2^2 = \lambda_{max}(\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T)^2,$$

其中 $\lambda_{max}(A)$ 表示A的最大特征值。
现在，$\mathbf{W}\mathbf{W}^T$是一个半正定矩阵。因此，$\mathbf{W}\mathbf{W}^T$的所有特征值均非负。此外，

$$\lambda_{max}(\mathbf{W}\mathbf{W}^T) \leq \|\mathbf{W}\|_F \leq 1$$

因此，$\mathbf{W}\mathbf{W}^T$的所有特征值都位于 $[0, 1]$ 和 $\eta \leq 1$ 中，这意味着 $\lambda_{max}(\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T)^2 \leq 1$ □

定理1. 设$\mathbf{x}_i$和$\mathbf{x}j$是 d维空间中的两个点$(\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d)$，且它们属于一个 δ-球，即$\|\mathbf{x}_i - \mathbf{x}_j\| < \delta$，其中$_j\| < \delta$是预先设定的一个较小正数。在训练过程中，如果最终层的权重矩阵的弗罗贝尼乌斯范数小于1，那么 *NeuroDAVIS* 生成的对应低维嵌入将在每次迭代中彼此更加接近。

证明. 考虑一个没有隐藏层且没有正则化的简单 NeuroDAVIS 模型。设$\mathbf{y}_i$和$\mathbf{y}j$分别是$\mathbf{x}_i$和$\mathbf{x}j$对应的初始低维嵌入。那么

$$\mathbf{y}_i = \mathbf{W}_1\mathbf{e}_i + \mathbf{b}_1 = \mathbf{W}_{1,i} + \mathbf{b}_1,$$

$\mathbf{e}_i$是i^{th}列$\mathbf{I}$。在每次前向传播过程中，NeuroDAVIS 都会尝试重构原始数据点。令˜$\mathbf{x}_i$为$\mathbf{x}_i$的重构，即 ˜$\mathbf{x}_i = (\mathbf{W}1,_i + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$。Th
erefore, the loss functi$_{\text{on}}$ 能　be written 为

$$\mathscr{L}_{NeuroDAVIS} = \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2,$$

相应的权重和偏置值通过 $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta\nabla\mathbf{W}^{(t)}$ 进行更新。现在，

$$\frac{\partial\mathscr{L}}{\partial\mathbf{W}_{1,i}} = (\mathbf{x}_i - \tilde{\mathbf{x}}_i)(-\frac{\partial\tilde{\mathbf{x}}_i}{\partial\mathbf{W}_{1,i}}) = -(\mathbf{x}_i - \tilde{\mathbf{x}}_i)\frac{\partial}{\partial\mathbf{W}_{1,i}}(\mathbf{W}_{1,i}\mathbf{W}_2) = -(\mathbf{x}_i - \tilde{\mathbf{x}}_i)\mathbf{W}_2^T$$

在t^{th}次迭代中，权重更新如下：$\mathbf{W}_1^{(t+1)}{}_{,i} = \mathbf{W}_1^{(t)}{}_{,i} + \eta(\mathbf{x}_i - \tilde{}\mathbf{x}_i)\mathbf{W}_2^{(t)T}$因此，

$$\begin{aligned}
\mathbf{y}_i^{(t+1)} - \mathbf{y}_j^{(t+1)} &= \mathbf{W}_{1,i}^{(t+1)} - \mathbf{W}_{1,j}^{(t+1)} \\
&= [\mathbf{W}_{1,i}^{(t)} + \eta(\mathbf{x}_i - \tilde{\mathbf{x}}_i)\mathbf{W}_2^{(t)T}] - [\mathbf{W}_{1,j}^{(t)} + \eta(\mathbf{x}_j - \tilde{\mathbf{x}}_j)\mathbf{W}_2^{(t)T}] \\
&= [\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}] + \eta[(\mathbf{x}_i - \mathbf{x}_j) - (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)]\mathbf{W}_2^{(t)T} \\
&\approx [\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}] - \eta(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)\mathbf{W}_2^{(t)T} \qquad [\text{as } \|\mathbf{x}_i - \mathbf{x}_j\| < \delta] \\
&\approx [\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}] - \eta[\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}]\mathbf{W}_2^{(t)}\mathbf{W}_2^{(t)T} \\
&\approx [\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}][\mathbf{I} - \eta\mathbf{W}_2^{(t)}\mathbf{W}_2^{(t)T}] \\
\|\mathbf{y}_i^{(t+1)} - \mathbf{y}_j^{(t+1)}\| &\approx \|[\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}][\mathbf{I} - \eta\mathbf{W}_2^{(t)}\mathbf{W}_2^{(t)T}]\| \\
&\leq \|\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}\|\|\mathbf{I} - \eta\mathbf{W}_2^{(t)}\mathbf{W}_2^{(t)T}\| \\
&\leq \|\mathbf{W}_{1,i}^{(t)} - \mathbf{W}_{1,j}^{(t)}\| \qquad [\text{By Lemma 1}]
\end{aligned}$$

因此，

$$\|\mathbf{y}_i^{(t+1)} - \mathbf{y}_j^{(t+1)}\| \leq \|\mathbf{y}_i^{(t)} - \mathbf{y}_j^{(t)}\|$$

□

表1：用于评估 NeuroDAVIS 的数据集描述

类别		Name	#样本	#特征	#类别	来源
2D		EllipticRing	1100	2	3	合成数据
		奥林匹克	2500	2	5	
		螺旋	312	2	3	
		形状	2000	2	5	
		世界地图	2843	2	5	
HD	数值型	乳腺癌	569	30	2	[25]
		Wine	178	13	3	[26]
	Text	Spam	5572	513	2	[27]
	图像	Coil20	1440	16385	20	[28]
		Fashion MNIST	60000	784	10	[29]
	单细胞RNA测序	Usoskin	622	25334	13	[30]
		Jurkat	3388	32738	11	[31]

4 结果

我们在多种数据集上评估了 NeuroDAVIS 的性能，包括二维合成数据集，以及不同数据模态的高维（HD）数据集，如数值型、文本、图像和生物数据（单细胞RNA测序（scRNA-seq）数据）。用于评估的数据集已在表 1 中说明。本节介绍了本研究为评估 NeuroDAVIS 性能并分析结果而开展的实验。我们重点考察了 NeuroDAVIS 在低维嵌入、可视化、结构保持、簇间距离保持和簇大小保持方面的能力。

4.1 合成数据2维数据集上的性能评估

为了展示 NeuroDAVIS 的有效性，我们首先将 NeuroDAVIS 应用于四个合成二维数据集，即椭圆环、奥林匹克、螺旋和形状。并将结果与使用 t-SNE 和 UMAP 得到的结果进行了比较。

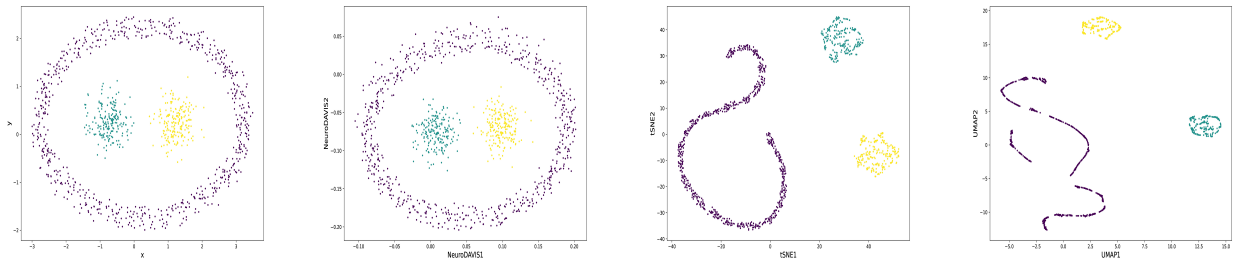
4.1.1 嵌入、可视化与结构保持

椭圆环数据集（图 3a）由外层高斯椭圆环内的两个小高斯球组成；奥林匹克数据集（图 3e）包含代表奥林匹克标志的五个圆环；螺旋数据集（图 3i）包含三条同心螺旋，而形状数据集（图 3m）由代表字符 ‘S’、‘H’、‘A’、‘P’ 和 ‘E’ 的点组成。

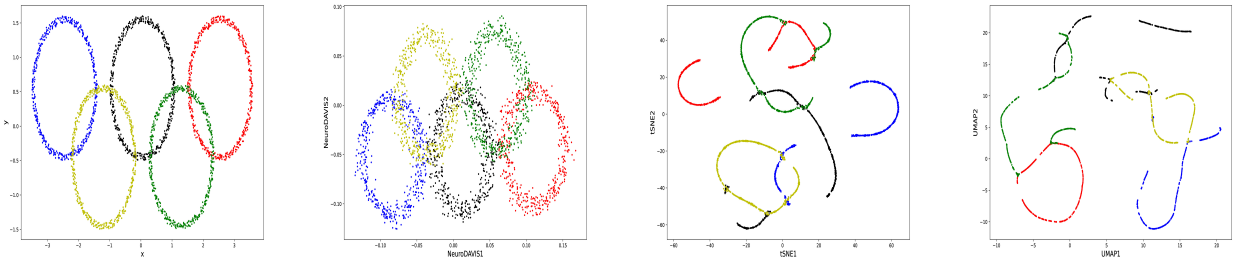
我们观察到，NeuroDAVIS 生成的椭圆环数据集嵌入（图 3b）中，外层高斯环内包含两个高斯球，与原始数据相似。同样，奥林匹克数据集中各圆环之间的距离、螺旋数据集中各同心螺旋之间的距离及其形状，都在 NeuroDAVIS 生成的嵌入中得到了保留（图 3f 和图 3j）。此外，对于形状数据集，NeuroDAVIS 保留了字符 ‘S’、‘H’、‘A’、‘P’ 和 ‘E’ 的可读性（图 3n）；而在 t-SNE 和 UMAP 的嵌入中，这些字符完全无法辨认。因此，可以说 NeuroDAVIS 能够以与原始分布相似的方式表示数据中的簇，同时保留其形状和大小。然而，t-SNE 和 UMAP 均未能保留原始分布，而是生成了紧凑的簇，破坏了数据中的局部和全局结构。

我们进一步使用斯皮尔曼秩相关比较了原始分布中的成对距离与 t-SNE、UMAP 和 NeuroDAVIS 生成的嵌入中的成对距离。图 4 展示了同一实验进行十次的结果。我们观察到，对于所有数据集，NeuroDAVIS 得到的相关系数均优于 t-SNE 和 UMAP。

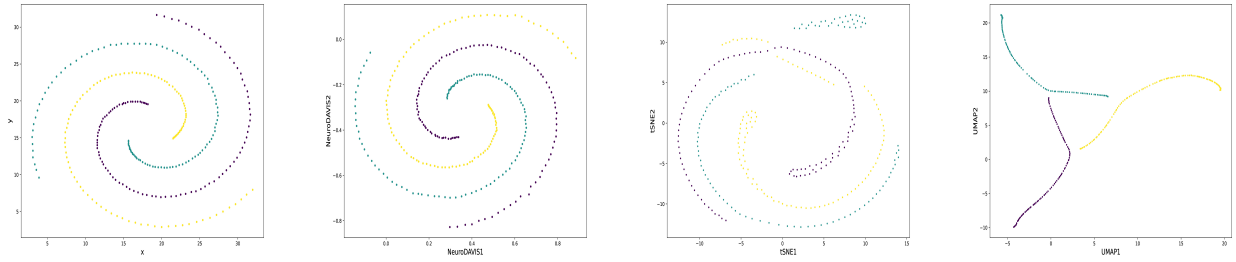
随后，我们在这些合成数据集上进行了另一项实验，使用 [10], 中采用的变换将二维空间投影到九维空间（$x + y, x - y, xy, x^2, y^2, x^2y, xy^2, x^3, y^3$），并应用 NeuroDAVIS 将数据投影回二维空间。我们将结果与使用相同变换得到的 t-SNE 和 UMAP 结果进行了比较。图 S1（见补充材料）表明，与 t-SNE 和 UMAP 相比，NeuroDAVIS 再次生成了更好的嵌入，同时保留了簇的形状和大小。



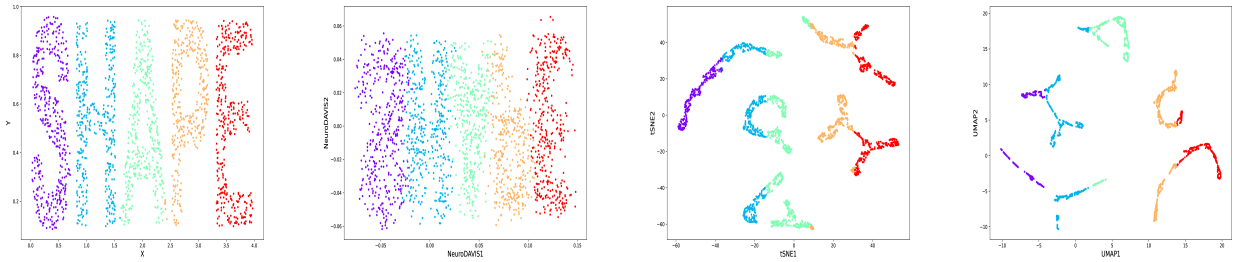
(a) 原始 (b) NeuroDAVIS (c) t-SNE (d) UMAP



(e) 原始 (f) NeuroDAVIS (g) t-SNE (h) UMAP



(i) 原始 (j) NeuroDAVIS (k) t-SNE (l) UMAP



(m) 原始 (n) NeuroDAVIS (o) t-SNE (p) UMAP

图 3：Elliptic Ring（(a)-(d)）、奥林匹克（(e)-(h)）、螺旋（(i)-(l)）和形状（(m)-(p)）数据集的原始分布，以及 NeuroDAVIS、t-SNE 和 UMAP 生成的嵌入。

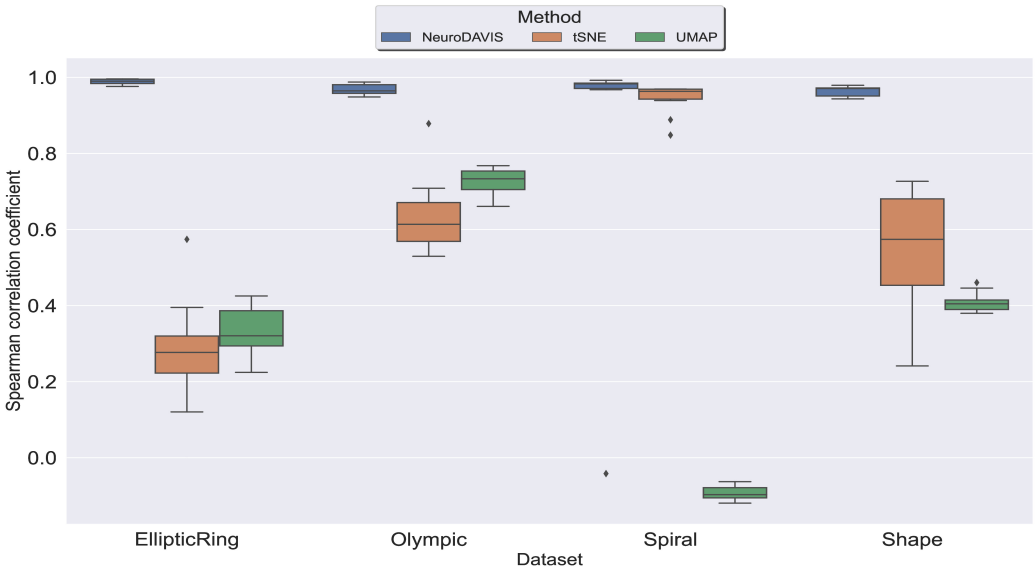


图 4：原始分布中的成对距离与 t-SNE、UMAP 和 NeuroDAVIS 对二维合成数据集椭圆环、奥林匹克、螺旋和形状生成的嵌入中的成对距离之间的斯皮尔曼秩相关。对于椭圆环数据集，得到的中位数相关系数值为 0.98（NeuroDAVIS）、0.27（t-SNE）和 0.32（UMAP）。对于奥林匹克数据集，得到的中位数相关系数值为 0.96（NeuroDAVIS）、0.61（t-SNE）和 0.73（UMAP）。对于螺旋数据集，得到的中位数相关系数值为 0.98（NeuroDAVIS）、0.96（t-SNE）和 −0.09（UMAP）。对于形状数据集，得到的中位数相关系数值为 0.97（NeuroDAVIS）、0.57（t-SNE）和 0.40（UMAP）。使用曼-惠特尼U检验得到的相应 p 值为：椭圆环、奥林匹克和形状数据集的 NeuroDAVIS-t-SNE 和 NeuroDAVIS-UMAP 均为 0.0001。对于螺旋数据集，得到的 p 值为 0.009（NeuroDAVIS-t-SNE）和 0.001（NeuroDAVIS-UMAP）。

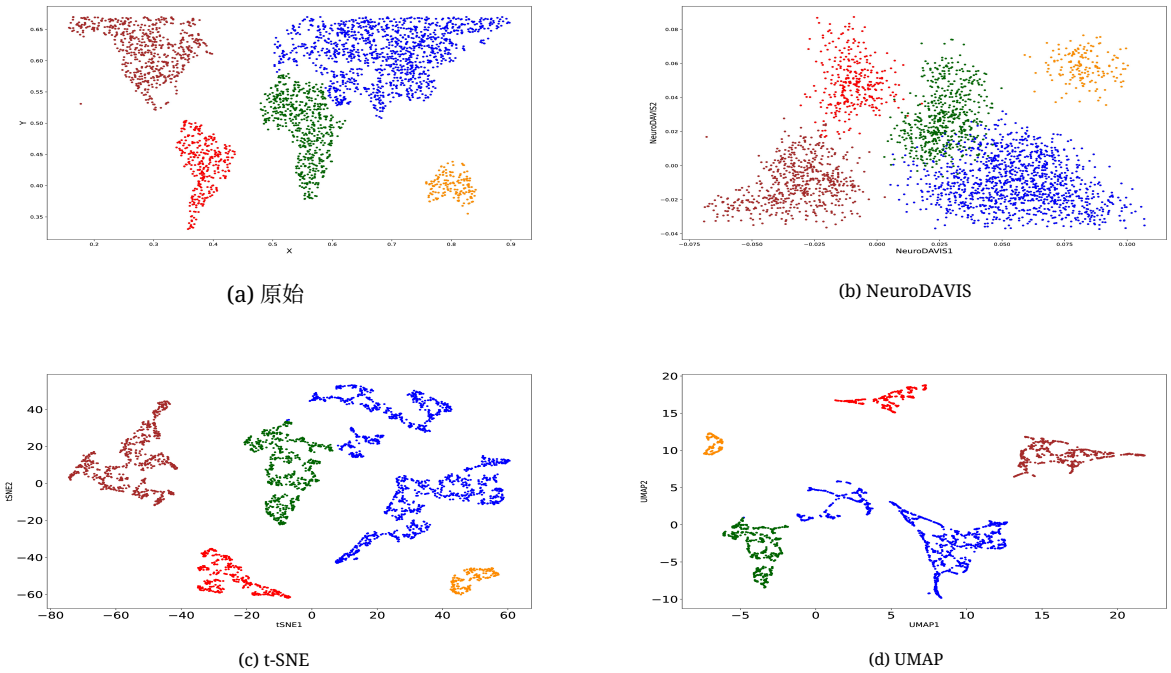
4.1.2 全局结构保持

为了评估降维/可视化方法的性能，我们需要分析降维后的嵌入，考察其在多大程度上能够以与高维空间中相似的分布方式表示各个簇。为此，研究数据中的簇间分离度至关重要。因此，我们又进行了一些实验，以确定 NeuroDAVIS 在低维空间中保持簇间距离的效果。真实标签信息对于此类分析十分重要。为此，我们按照教程（`https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17`）创建了一个表示世界地图（已知结构）的二维合成数据集，其中包含五个簇/大洲，即欧亚大陆、澳大利亚、北美洲、南美洲和非洲（不含南极洲）（图5a）。我们将 NeuroDAVIS 为世界地图数据集生成的嵌入与其原始数据分布以及 t-SNE 和 UMAP 生成的相应嵌入进行了比较（图5）。如图5所示，与 t-SNE 和 UMAP 不同，NeuroDAVIS 生成的嵌入中各簇的形状和大小与原始数据中的十分相似。

4.1.3 簇间距离保持

为了检查五大洲之间的簇间距离是否得到保持，我们计算了 NeuroDAVIS 得出的五个簇质心之间距离的斯皮尔曼秩相关系数，并将结果与使用 t-SNE 和 UMAP 得出的结果进行了比较。如图 6 所示，同一实验的多次重复运行证实，在距离保持方面，NeuroDAVIS 的表现远优于 t-SNE 和 UMAP。

我们进一步评估了世界地图数据集的 NeuroDAVIS嵌入，考察其保持某个簇中的点与其余簇中的点之间两两欧几里得距离的能力。图 7 展示了该实验多次重复运行的结果，表明对于欧亚大陆、澳大利亚、北美洲、南美洲和非洲这几个簇，NeuroDAVIS 保持原始成对距离的效果优于 t-SNE 和 UMAP。



(a) 原始

(b) NeuroDAVIS

(c) t-SNE

(d) UMAP

图 5：世界地图数据集的原始分布以及 NeuroDAVIS、t-SNE 和 UMAP 生成的嵌入。

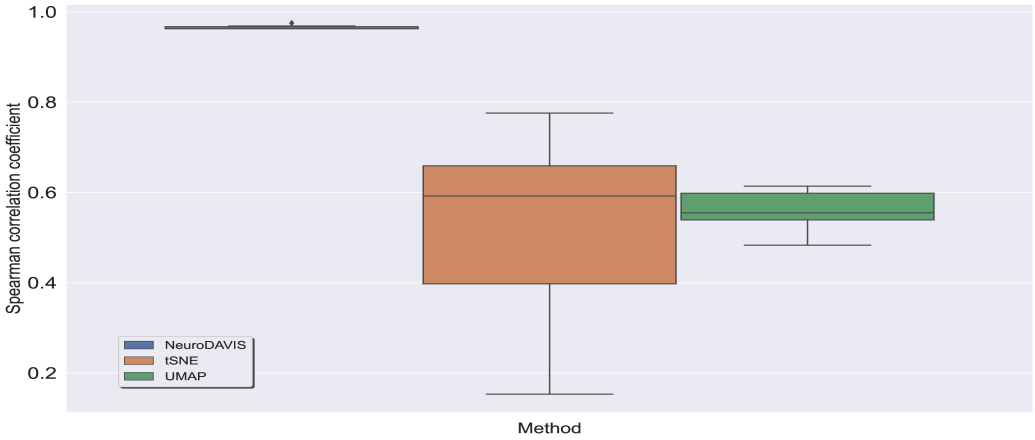
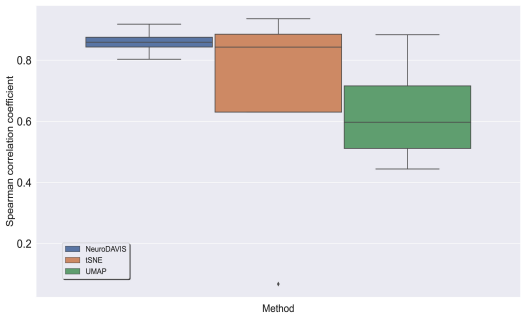
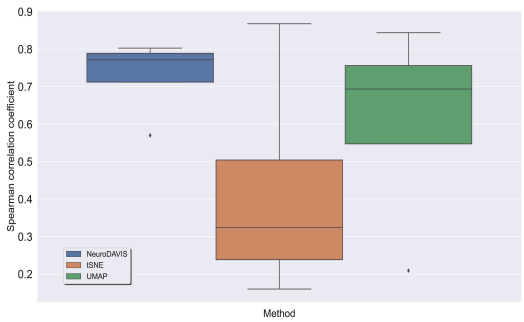


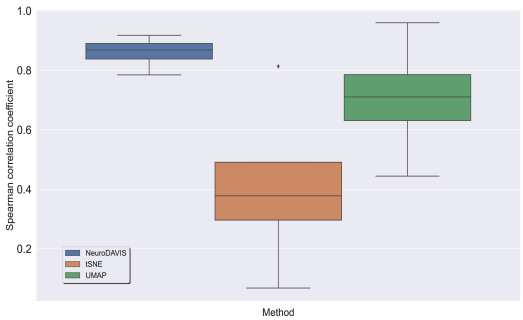
图6：原始数据集中簇质心之间的原始距离与 NeuroDAVIS、t-SNE 和 UMAP 生成的嵌入中簇质心之间距离的斯皮尔曼秩相关。对于世界地图数据集，中位数相关系数分别为 0.96（NeuroDAVIS）、0.59（t-SNE）和 0.55（UMAP）。使用曼-惠特尼U检验得到的相应 p 值为 0.0001（NeuroDAVIS-t-SNE 和 NeuroDAVIS-UMAP 均为 0.0001）。



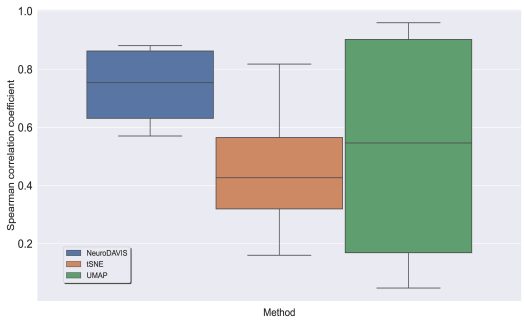
（a）欧亚大陆与其他地区之间的距离



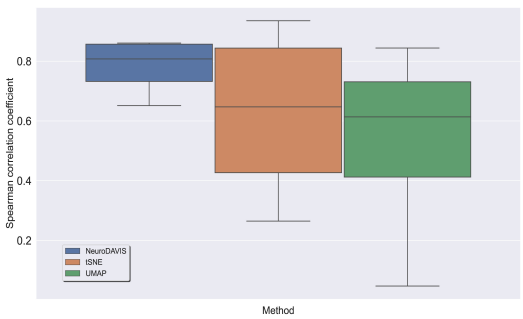
（b）澳大利亚与其他地区之间的距离



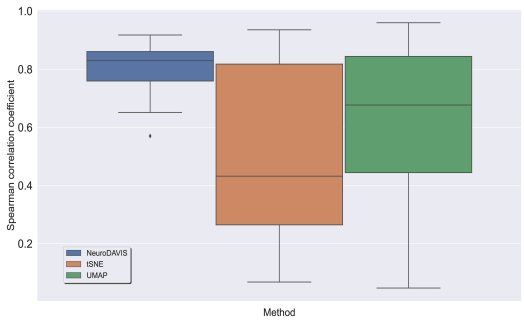
（c）北美洲与其他地区之间的距离



（d）南美洲与其他地区之间的距离



（e）非洲与其他地区之间的距离



（f）所有大洲之间的距离

图 7：NeuroDAVIS、t-SNE 和 UMAP 对世界地图数据集中各簇（大洲）之间成对距离的保持情况。

4.1.4 簇大小保持

我们想知道原始数据中的簇大小是否在 NeuroDAVIS 生成的嵌入中得到了保留。因此，为确保簇大小得到保留，我们在这个世界地图数据集上开展了几项额外实验。对于数据集中的每个簇，我们估算了簇边界最小矩形的面积。具体而言，我们考察了每个簇在两个方向上的延展程度，如图5a所示。随后，我们计算了簇在原始数据中的面积与其重构后对应面积之间的皮尔逊相关系数。通过多次重复这一实验，我们得出结论：与 t-SNE 和 UMAP 相比，NeuroDAVIS 更好地保留了簇的原始大小，如补充材料中的图S2所示。上述所有实验表明，在全局形状保持方面，NeuroDAVIS 优于 t-SNE 和 UMAP。

4.2 高维数据集上的性能评估

随后，我们在数值型、文本、图像和生物数据等多类高维数据集上验证了 NeuroDAVIS 的有效性。评估所采用的总体步骤如下。

我们首先将每个数据集降至 NeuroDAVIS 的两个维度。对于每个数据集，我们从两个不同方面，将 NeuroDAVIS 嵌入的质量与 t-SNE、UMAP 和 Fit-SNE 生成的嵌入进行了比较。首先，我们计算了数据点在原始空间中的成对距离与 NeuroDAVIS 、 t-SNE 、 UMAP 和 Fit-SNE 在潜在空间中生成的成对距离之间的斯皮尔曼相关系数。最后，对于除生物数据集之外的所有高维数据集，我们评估了 NeuroDAVIS 生成的嵌入的分类性能。我们按 80:20 的比例划分训练集与测试集，从 NeuroDAVIS 生成的嵌入中构建训练集和测试集。我们训练了两种分类器，即 k近邻和随机森林（RF），并使用留出的测试数据集评估分类器的性能。我们将结果与使用相同分类器处理 t-SNE、UMAP 和 Fit-SNE 嵌入所得的结果进行了比较。对于单细胞RNA测序（生物）数据集，我们在 NeuroDAVIS 嵌入上使用 k 均值和层次凝聚聚类进行细胞类型聚类，并将结果与在 t-SNE、UMAP 和 Fit-SNE 嵌入上得到的结果进行了比较。我们还将所得结果与两种近期经过基准测试的方法 PHATE [15] 和 IVIS [23], 进行了比较，这两种方法专为单细胞RNA测序数据集开发。为了衡量 NeuroDAVIS 和其他方法的性能，我们采用了 [32]中使用的一些指标。非生物数据集的分类性能采用准确率和 F1 分数衡量；对于单细胞RNA测序数据集，则采用两个外部指标——调整兰德指数（ARI）和 Fowlkes–Mallows指数（FMI）——衡量所生成簇的质量。

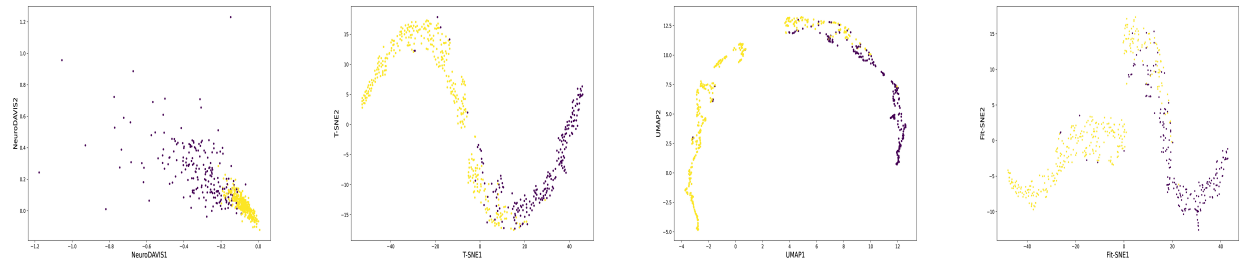
4.2.1 数值数据集

为评估 NeuroDAVIS 的性能，我们从 [33] 下载了两个数值数据集：乳腺癌和葡萄酒；这两个数据集最初可从 `https://archive.ics.uci.edu/ml/datasets.php` 获取。图 8a、8b、8c 和 8d 展示了 NeuroDAVIS 为乳腺癌数据集生成的嵌入，以及使用 t-SNE、UMAP 和 Fit-SNE 得到的嵌入；图 8e、8f、8g 和 8h 则展示了葡萄酒数据集对应的嵌入。如图 9 所示，相关系数值表明，高维空间中的成对距离与 NeuroDAVIS 嵌入中的距离相关性更高（中位数相关系数为 $= 0.94$（乳腺癌）、0.92（葡萄酒）），高于 t-SNE（中位数相关系数为 $= 0.76$（乳腺癌）、0.91（葡萄酒））、UMAP（中位数相关系数为 $= 0.77$（乳腺癌）、0.82（葡萄酒））和 Fit-SNE 嵌入（中位数相关系数为 $= 0.77$（乳腺癌）、0.92（葡萄酒））。乳腺癌数据集使用曼-惠特尼U检验得到的相应 p 值为 0.0001（所有配对比较均如此，即 NeuroDAVIS-t-SNE、NeuroDAVIS-UMAP 和 NeuroDAVIS-Fit-SNE）；葡萄酒数据集的相应 p 值为 0.0001（NeuroDAVIS-UMAP）、0.733（NeuroDAVIS-t-SNE）和 1.000（NeuroDAVIS-Fit-SNE）。

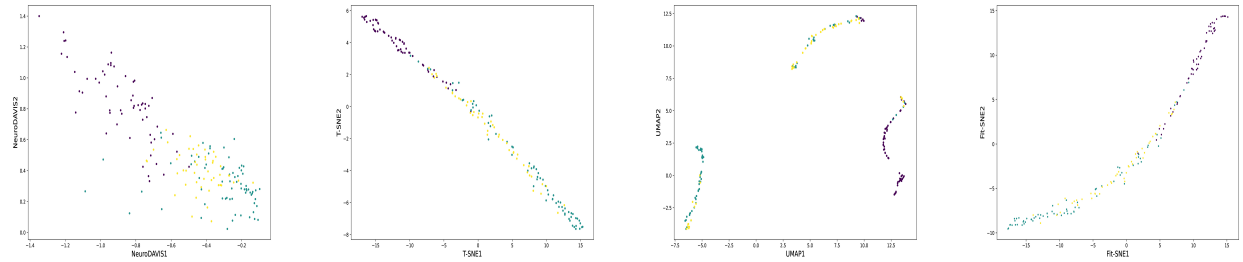
图 S3（见补充材料）展示了使用 k近邻和随机森林对 NeuroDAVIS嵌入进行分类的结果（准确率和 F1分数）。这些结果也表明，对于葡萄酒数据集，NeuroDAVIS 的表现优于 t-SNE、UMAP 和 Fit-SNE；对于乳腺癌数据集，NeuroDAVIS嵌入的分类性能则与 t-SNE、UMAP 和 Fit-SNE 嵌入相当。每项实验均重复多次，并记录结果。

4.2.2 文本数据集

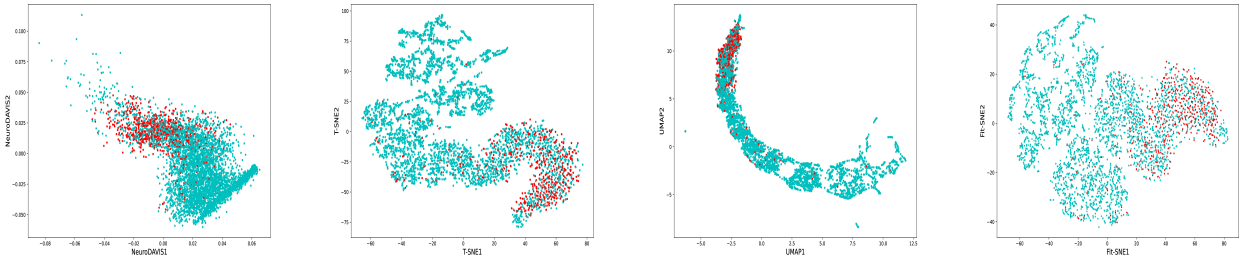
接下来，我们使用一个文本数据集来展示 NeuroDAVIS 的有效性。垃圾短信数据集最初可从 `https://archive.ics.uci.edu/ml/datasets.php` 获取，已从 `https://www.kaggle.com/datasets/team-ai/spam-text-message-classification` 下载。该数据集共包含 5572 条消息，其中 747 条消



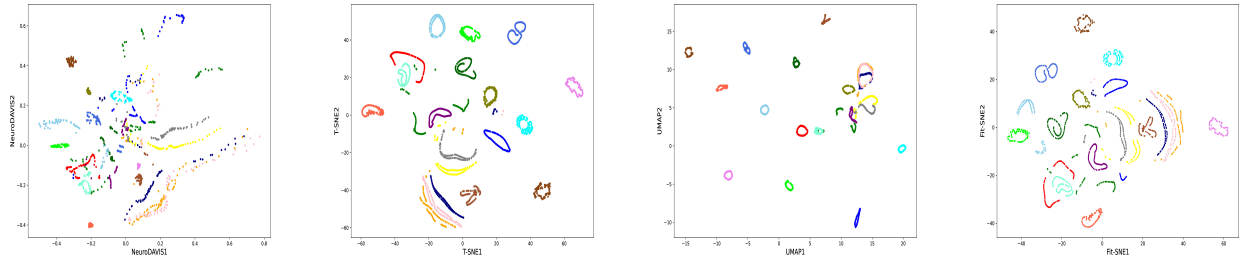
(a) NeuroDAVIS (b) t-SNE (c) UMAP (d) Fit-SNE



(e) NeuroDAVIS (f) t-SNE (g) UMAP (h) Fit-SNE



(i) NeuroDAVIS (j) t-SNE (k) UMAP (l) Fit-SNE



(m) NeuroDAVIS (n) t-SNE (o) UMAP (p) Fit-SNE

图 8：NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE 为乳腺癌（(a)–(d)）、葡萄酒（(e)–(h)）、垃圾邮件（(i)–(l)）和 Coil20（(m)–(p)）数据集生成的嵌入



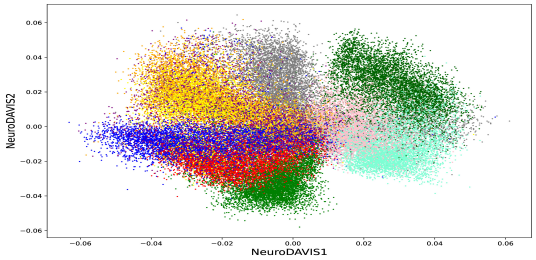
图 9：乳腺癌、葡萄酒、垃圾短信和 Coil20 数据集的原始分布中成对距离与 NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE 嵌入中成对距离之间的斯皮尔曼秩相关系数。乳腺癌数据集的中位数相关系数分别为 0.94（NeuroDAVIS）、0.76（t-SNE）、0.77（UMAP）和 0.77（Fit-SNE）。葡萄酒数据集的中位数相关系数分别为 0.92（NeuroDAVIS）、0.91（t-SNE）、0.82（UMAP）和 0.92（Fit-SNE）。垃圾短信数据集的中位数相关系数分别为 0.77（NeuroDAVIS）、0.59（t-SNE）、0.57（UMAP）和 0.61（Fit-SNE）；而 Coil20 数据集的对应数值分别为 0.61（NeuroDAVIS）、0.53（t-SNE）、0.15（UMAP）和 0.56（Fit-SNE）。使用曼-惠特尼U检验得到的相应 p值，对于乳腺癌和垃圾短信数据集，NeuroDAVIS-t-SNE 和 NeuroDAVIS-UMAP 的检验结果均为 0.0001。葡萄酒数据集的曼-惠特尼U检验 p值分别为 0.73（NeuroDAVIS-t-SNE）、0.002（NeuroDAVIS-UMAP）和 1.00（NeuroDAVIS-Fit-SNE）；而 Coil20 数据集的曼-惠特尼U检验 p值分别为 0.0007（NeuroDAVIS-t-SNE）、0.0001（NeuroDAVIS-UMAP）和 0.002（NeuroDAVIS-Fit-SNE）。

息被分类为垃圾邮件。该数据集已采用文本数据集常用的标准流程进行预处理，主要步骤包括：1. 移除 URL；2. 转换为小写；3. 移除标点符号；4. 移除多余空白字符；5. 移除停用词；6. 词形还原；7. 分词；8. 最小-最大缩放。

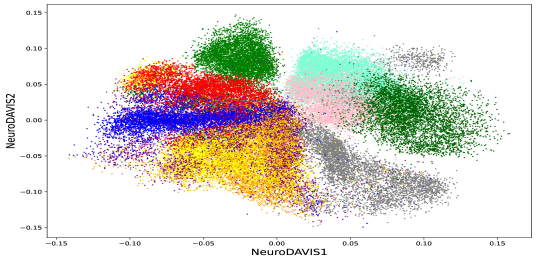
NeuroDAVIS 对预处理后的垃圾短信数据集生成的二维嵌入如图 8i 所示。同一数据集对应的 t-SNE、UMAP 和 Fit-SNE 嵌入分别如图 8j、8k 和 8l 所示。尽管这些嵌入看起来相当相似，但仔细观察可以发现，与 NeuroDAVIS 和 UMAP 的结果截然不同，t-SNE 和 Fit-SNE 嵌入中的非垃圾邮件簇都呈现出多个子簇，这并不符合实际情况。图 9 还显示，NeuroDAVIS 嵌入的相关系数值（中位数相关系数 = 0.77）远优于 t-SNE（中位数相关系数 = 0.59）、UMAP（中位数相关系数 = 0.57）和 Fit-SNE（中位数相关系数 = 0.61）嵌入的结果。使用曼-惠特尼U检验得到的相应 p 值为 0.0001（针对 NeuroDAVIS-t-SNE、NeuroDAVIS-UMAP 和 NeuroDAVIS-Fit-SNE）。然而，如补充材料中的图 S3 所示，垃圾短信数据集的 NeuroDAVIS 嵌入分类性能不如 t-SNE、UMAP 或 Fit-SNE 嵌入。

4.2.3 图像数据集

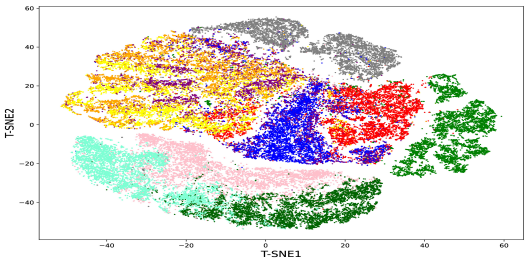
随后，我们在图像数据集 Coil20 上评估了 NeuroDAVIS。Coil20 数据集 [28] 包含 20 种不同物体的图像，且已去除背景。图 8m、8n、8o 和 8p 分别展示了在该数据集上获得的 NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE 嵌入。我们观察到，与数值型和文本数据集的情况类似，NeuroDAVIS 保留高维空间中物体间距离的效果优于 t-SNE 和 UMAP。NeuroDAVIS 报告的中位数相关系数为 0.61，而 t-SNE、UMAP 和 Fit-SNE 嵌入的中位数相关系数分别为 0.53、0.15 和 0.56。图 9 展示了多次执行该实验所得到的相关系数值分布。相应的 p-



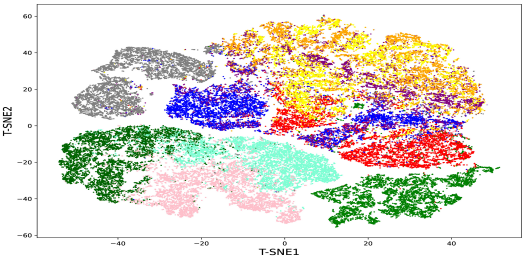
(a) 原始数据的 NeuroDAVIS 嵌入



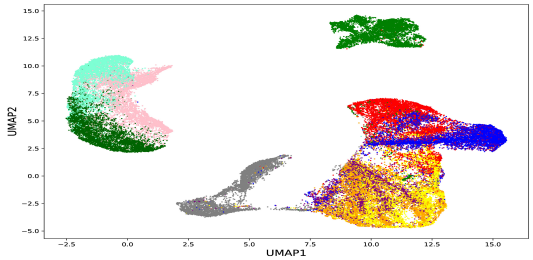
(b) NeuroDAVIS 之前的主成分分析



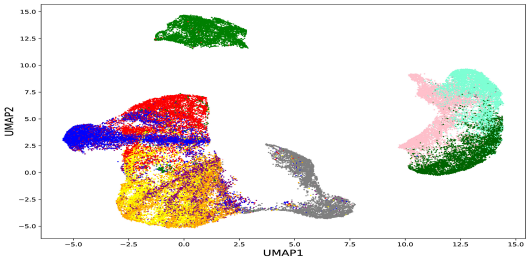
(c) 原始数据的 t-SNE 嵌入



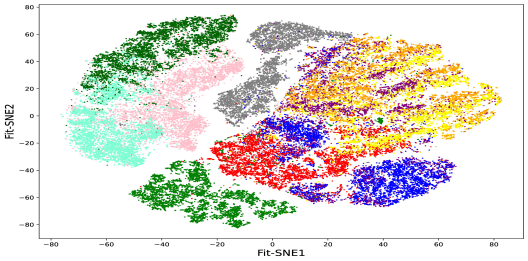
(d) t-SNE 之前的主成分分析



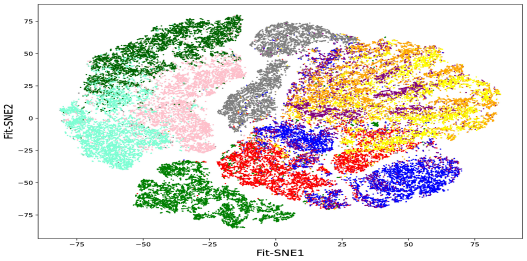
(e) 原始数据的 UMAP 嵌入



（f）在 UMAP 之前进行主成分分析



（g）原始数据的 Fit-SNE 嵌入



（h）在 Fit-SNE 之前进行主成分分析

图 10：（a）在 Fashion MNIST 数据集上获得的 NeuroDAVIS 嵌入，（c）t-SNE 嵌入，（e）UMAP 嵌入，（g）Fit-SNE 嵌入；（b）在原始 Fashion MNIST 数据集的前 50 个主成分上获得的 NeuroDAVIS 嵌入，（d）t-SNE 嵌入，（f）UMAP 嵌入，（h）Fit-SNE 嵌入。

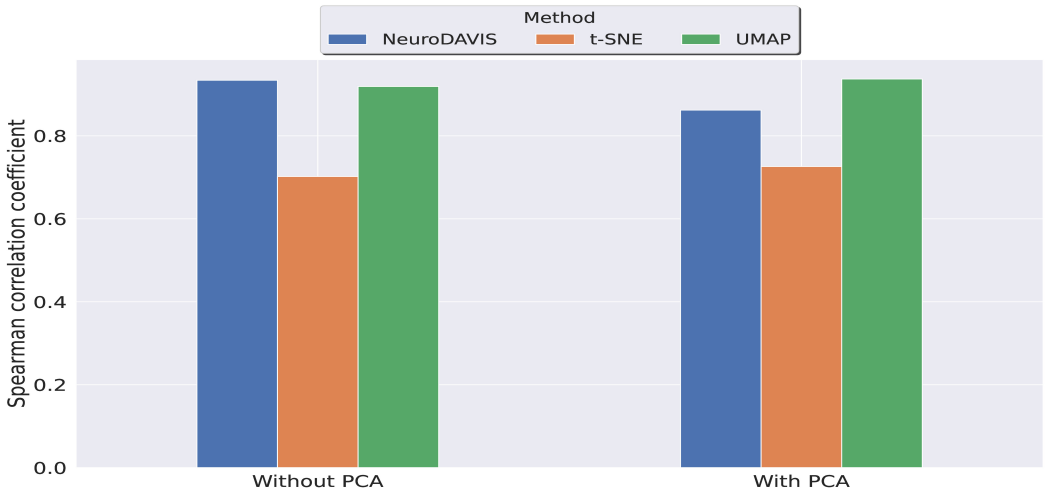


图 11：使用 NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE（有无基于 PCA 的预处理）生成的 Fashion MNIST 数据集嵌入中，低维空间簇质心之间的成对距离与高维空间簇质心之间的成对距离的斯皮尔曼秩相关系数。

使用曼-惠特尼U检验得到的 p 值分别为 0.0007（NeuroDAVIS-t-SNE）、0.0001（NeuroDAVIS-UMAP）和 0.002（NeuroDAVIS-Fit-SNE）。

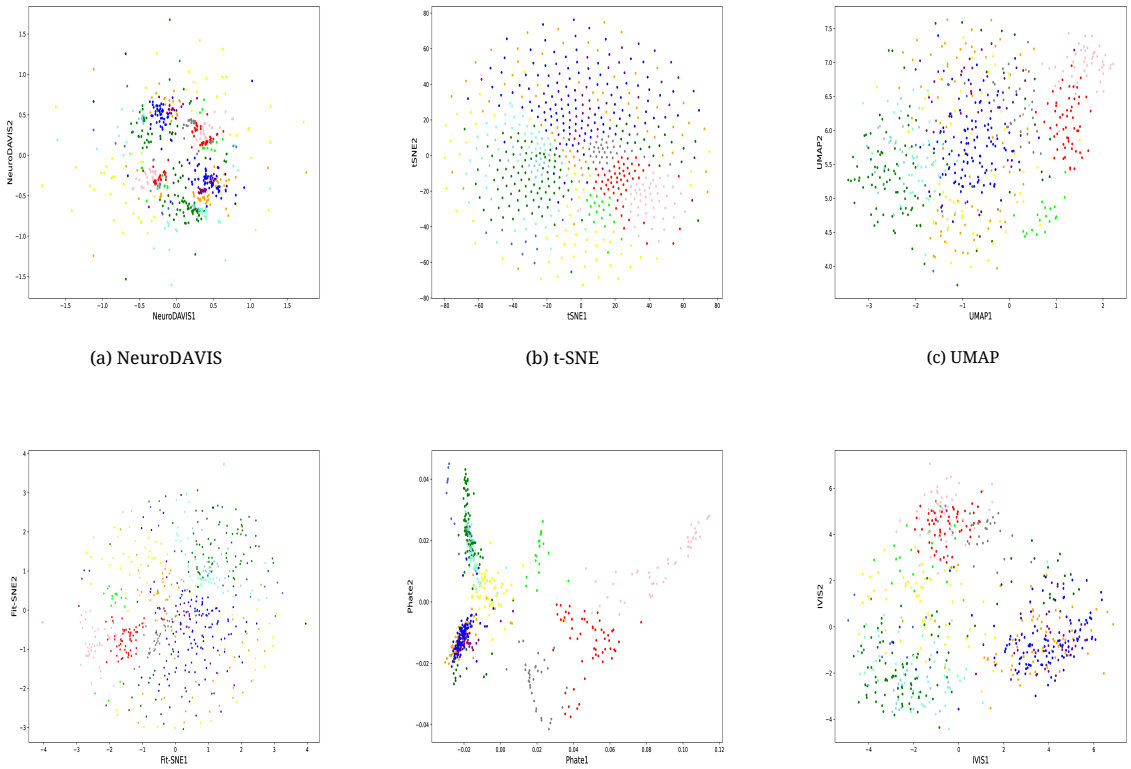
在 Coil20 图像数据集的 NeuroDAVIS 嵌入上，k近邻和随机森林分类器取得的分类准确率和 F1 分数优于 UMAP。不过，正如补充材料中的图 S3 所示，t-SNE 和 Fit-SNE 在该数据集上取得了更好的分类结果。

为了展示 NeuroDAVIS 在大型数据集上的有效性，我们还在另一个名为 Fashion MNIST 的图像数据集上评估了 NeuroDAVIS，该数据集包含 6 万张训练图像，每张图像含有 28×28 个像素。图 10a、10c、10e 和 10g 分别展示了 NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE 生成的 Fashion MNIST 数据集嵌入。为了从高维数据集中获得高质量嵌入，研究人员通常会将 PCA 作为预处理步骤 [23]。因此，我们进一步进行了调查，将 NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE 应用于 Fashion MNIST 数据的前 50 个主成分，并将结果与仅使用 NeuroDAVIS、t-SNE、UMAP 和 Fit-SNE 得到的结果进行比较。图 10b、10d、10f 和 10h 展示了在应用 NeuroDAVIS、t-SNE、UMAP 或 Fit-SNE 之前使用 PCA 作为预处理步骤的效果。

通常很难通过视觉评估簇的质量。因此，为了量化簇的质量，我们计算了高维空间中簇质心之间的成对距离与低维嵌入中簇质心之间的成对距离的斯皮尔曼秩相关系数。相关系数越高，表示簇间距离的保留效果越好。如图 11 所示，我们发现 NeuroDAVIS 得到的相关系数更高，为 0.93；相比之下，t-SNE 的相关系数为 $= 0.70$，UMAP 为 $= 0.91$，Fit-SNE 为 $= 0.89$。有趣的是，使用 PCA 进行预处理后，NeuroDAVIS 的相关系数反而降低了，为 $= 0.86$；而 t-SNE（相关系数为 $= 0.72$）和 UMAP（相关系数为 $= 0.93$）的结果有所改善。Fit-SNE 则不受将 PCA 用作预处理步骤的影响（相关系数为 $= 0.89$）。这些结果如图 11 所示。由此我们推断，使用 PCA 作为预处理工具时，t-SNE 或 UMAP 的表现更好。另一方面，NeuroDAVIS 无需依赖 PCA 即可生成高质量嵌入。在应用 NeuroDAVIS 之前进行基于 PCA 的预处理可能会造成信息损失，从而损害嵌入质量。

4.2.4 生物数据集

最后，我们探究了 NeuroDAVIS 在一些生物数据集上的有效性。为此，我们使用了两个单细胞 RNA 测序数据集。Usoskin数据集 [30] 包含 622 个细胞在 25334 个基因上的基因表达值，而 Jurkat数据集 [31] 包含 3388 个细胞在 32738 个基因上的基因表达值。这两个数据集均已按照教程建议的标准流程，使用 Scanpy [34] 进行了预处理



(a) NeuroDAVIS (b) t-SNE (c) UMAP

(d) Fit-SNE (e) PHATE (f) IVIS

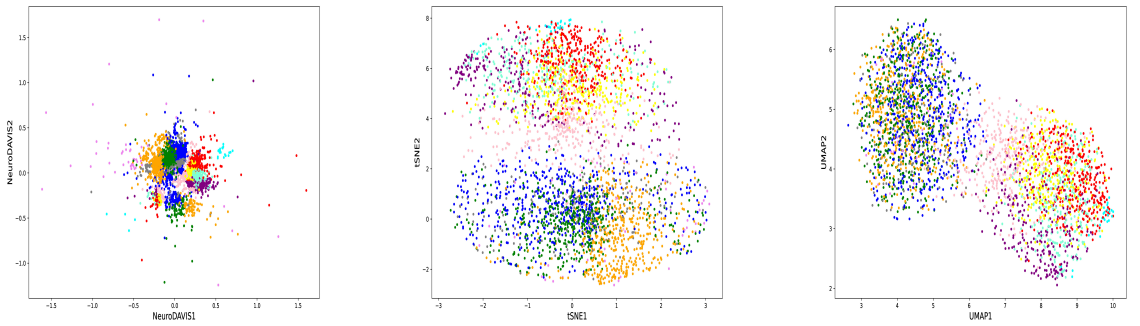
图 12：分别由 NeuroDAVIS、t-SNE、UMAP、Fit-SNE、PHATE 和 IVIS 生成的 Usoskin 数据二维嵌入。

`https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html`。除了 t-SNE、UMAP 和 Fit-SNE 之外，我们还将 PHATE 和 IVIS 纳入这些数据集的性能比较；除了 t-SNE 和 UMAP 外，它们也是目前适用于单细胞 RNA 测序数据集的最先进算法。值得一提的是，这两个数据集中的样本数量都远低于维度数量，这进一步增加了学习难度。

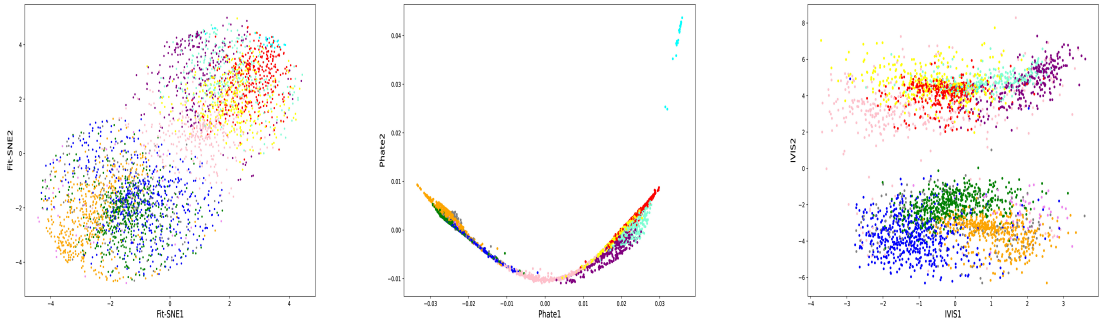
如图 12 和图 13 所示，NeuroDAVIS、t-SNE、UMAP、Fit-SNE、PHATE 和 IVIS 分别为 Usoskin 和 Jurkat 数据集生成的二维嵌入表明，t-SNE 和 UMAP 都未能清晰地呈现数据中的簇。对于 Usoskin 数据集，NeuroDAVIS 能够较好地呈现其中一些簇；而对于 Jurkat 数据集，与其他方法相比，NeuroDAVIS 投影中的簇更加密集且分离得更好。总体而言，NeuroDAVIS 投影优于 t-SNE、UMAP 和 Fit-SNE 投影，仅与 PHATE 或 IVIS 投影相当。

图 14 报告了原始数据中的成对距离与由 NeuroDAVIS、t-SNE、UMAP、Fit-SNE、PHATE 和 IVIS 生成的嵌入中的成对距离之间的相关系数值。我们再次观察到，对于 Jurkat 数据集，NeuroDAVIS 的相关性（中位数相关系数 = 0.41）优于所有其他方法；而对于 Usoskin 数据集，NeuroDAVIS 的中位数相关系数为 0.23，优于除 t-SNE（中位数相关系数 = 0.27）和 Fit-SNE（中位数相关系数 = 0.29）之外的所有其他方法。对于 Jurkat 数据集，使用曼-惠特尼U检验得到的相应 p 值为 0.0001（NeuroDAVIS-t-SNE、NeuroDAVIS-UMAP、NeuroDAVIS-Fit-SNE、NeuroDAVIS-PHATE 和 NeuroDAVIS-IVIS）；对于 Usoskin 数据集，使用曼-惠特尼U检验得到的 p 值为 0.677（NeuroDAVIS-t- SNE）、0.472（NeuroDAVIS-Fit-SNE）和 0.0001（NeuroDAVIS-UMAP、NeuroDAVIS-PHATE 和 NeuroDAVIS-IVIS）。

如补充材料中的图 S4 所示，我们进一步观察到，就调整兰德指数和 Fowlkes–Mallows 指数得分而言，在 Usoskin 数据上，k 均值和凝聚层次聚类方法应用于 NeuroDAVIS 嵌入时的聚类性能优于 t-SNE 和 IVIS 嵌入，但不及 PHATE 和 Fit-SNE 嵌入，同时与 UMAP 嵌入



(a) NeuroDAVIS (b) t-SNE (c) UMAP



(d) Fit-SNE (e) PHATE (f) IVIS

图 13：NeuroDAVIS、t-SNE、UMAP、Fit-SNE、PHATE 和 IVIS 分别生成的 Jurkat 数据二维嵌入。

的表现非常接近。对于 Jurkat数据集，NeuroDAVIS 嵌入的聚类性能在调整兰德指数和 Fowlkes–Mallows 指数得分方面均为最佳，只有 IVIS 嵌入可以与之媲美。

5 讨论与结论

在本研究中，我们开发了一种名为 NeuroDAVIS 的新型无监督深度学习模型，用于数据可视化。NeuroDAVIS 能够从数据中生成低维嵌入，并提取有助于后续分析流程的相关特征。我们在多种数据集上验证了 NeuroDAVIS 的有效性，包括二维合成数据以及高维数值型、文本、图像和生物数据。该模型展现出出色的可视化能力，并在下游分析（聚类/分类）中表现优异，因此具备与 t-SNE、UMAP 和 Fit-SNE 等最先进方法竞争的实力。结果也在统计学上显著支持 NeuroDAVIS。NeuroDAVIS 的优势在于能够很好地保留数据中的簇形状、大小和结构。数学分析表明，NeuroDAVIS 生成的低维嵌入还能够保留高维空间中观测值之间的局部关系。此外，该模型能够提取数据中固有的非线性特征。它也不预先假定数据分布，因此更加实用。该模型可称为一种通用的降维与可视化系统，不受嵌入维度限制。它无需任何预处理，即可独立生成可解释的可视化结果。

NeuroDAVIS 作为一种降维与可视化方法，也有其不足之处。它的输入是一个单位矩阵，其阶数等于输入数据中的观测值数量。因此，与其他现有方法相比，它会占用更多内存。此外，NeuroDAVIS 的超参数对细微改动十分敏感，而网络调优需要业余人员投入大量时间和精力。

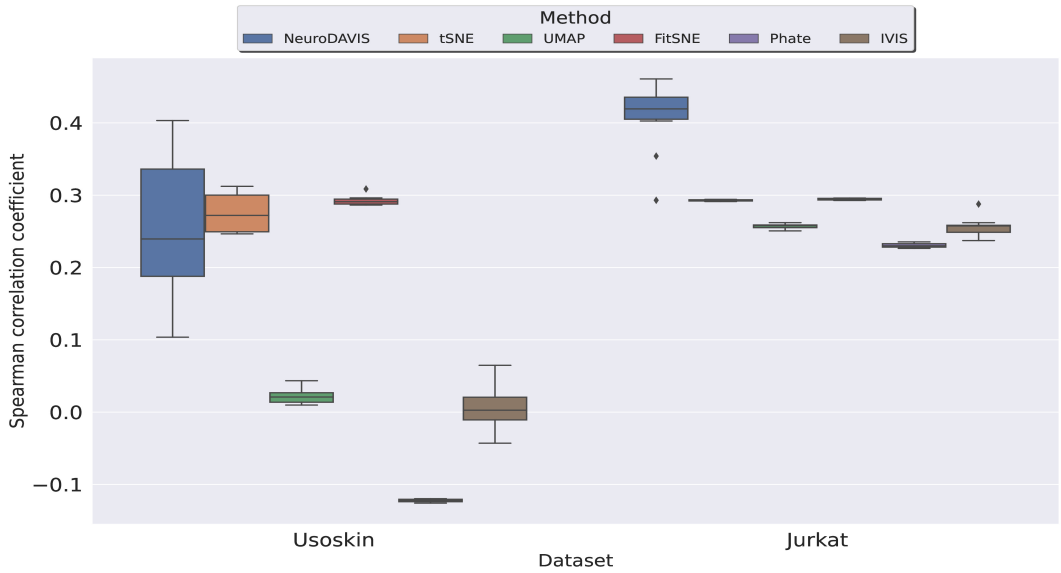


图14：Usoskin和Jurkat数据集的原始分布中的成对距离与NeuroDAVIS、t-SNE、UMAP、Fit-SNE、PHATE和IVIS生成的嵌入中的成对距离之间的斯皮尔曼秩相关。对于Usoskin数据集，得到的中位数相关系数分别为0.24（NeuroDAVIS）、0.27（t-SNE）、0.02（UMAP）、0.29（Fit-SNE）、−0.12（PHATE）和0.002（IVIS）。对于Jurkat数据集，得到的中位数相关系数分别为0.42（NeuroDAVIS）、0.29（t-SNE）、0.26（UMAP）、0.29（Fit-SNE）、0.23（PHATE）和0.25（IVIS）。使用曼-惠特尼U检验得到的Usoskin数据集对应的p值为0.677（NeuroDAVIS-t-SNE）、0.472（NeuroDAVIS-Fit-SNE）、0.0001（NeuroDAVIS-UMAP、NeuroDAVIS-PHATE和NeuroDAVIS-IVIS）；Jurkat数据集对应的p值为0.0001（NeuroDAVIS-t-SNE、NeuroDAVIS-UMAP、NeuroDAVIS-Fit-SNE、NeuroDAVIS-PHATE和NeuroDAVIS-IVIS）。

尽管如此，NeuroDAVIS 仍是可视化高维数据集的一种有竞争力的方法。它生成的潜在嵌入具有较强的鲁棒性，优于 t-SNE、UMAP 和 Fit-SNE 等部分现有基准方法，同时也解决了局部与全局形状保持这一难题。因此，在这项工作中，我们为长期存在的、如何可视化属于不同类别／领域／数据模态的数据集这一问题，提出了一种统一的解决方案。未来可以为现有架构添加特征选择模块，以扩展 NeuroDAVIS 的功能。它还可以进一步扩展，用于可视化高维多模态数据集，通过生成能够捕捉数据重要特征的低维嵌入来实现。

数据与代码可用性

NeuroDAVIS 使用 Python 3 实现。用于复现结果的代码可在 `https://github.com/shallowlearner93/NeuroDAVIS` 获取。本研究使用的预处理数据集可从 `https://doi.org/10.5281/zenodo.7315674` 下载。

补充信息

补充图已收录于补充文件中。

作者贡献

方法论构思：CM、DBS、RKD。数据整理、数据分析、正式分析、可视化、调查、实现、验证、原始稿件撰写：CM、DBS。验证、审阅、编辑、总体监督：DBS、RKD。

致谢

本研究得到 DST-NSF 资助，该资助由 IDEAS-TIH、加尔各答印度统计研究所提供给 RKD。

声明

DBS 担任副　　研究数据科学家，任职于 Tatras Data Services Pvt. Ltd. 他未因这项工作获得任何资金。

参考文献

[1] Hotelling, H.：将一组统计变量分解为主成分。《教育心理学杂志》**24**(6), 417 (1933)[2] Jutten, C., Herault, J.：盲源分离，第一部分：一种基于神经拟态架构的自适应算法。《信号处理》**24**(1), 1–10 (1991)[3] Kruskal, J.B.：通过优化对非度量假设的拟合优度进行多维尺度分析。《心理测量学》**29**(1), 1–27 (1964)[4] Tenenbaum, J.B., Silva, V.d., Langford, J.C.：一种用于非线性降维的全局几何框架。《科学》**290**(5500), 2319–2323 (2000)[5] Lee, D.D., Seung, H.S.：通过非负矩阵分解学习物体的组成部分。《自然》**401**(6755), 788–791 (1999)[6] Stewart, G.W.：奇异值分解的早期历史。《SIAM评论》**35**(4), 551–566 (1993)[7] Van der Maaten, L., Hinton, G.：使用 t-SNE 可视化数据。《机器学习研究杂志》**9**(11) (2008)[8] McInnes, L., Healy, J., Melville, J.：UMAP：用于降维的统一流形近似与投影。arXiv 预印本 arXiv:1802.03426 (2018)[9] Izenman, A.J.：线性判别分析。载于：《现代多元统计技术》，第237–280页。施普林格，??? (2013)[10] Ding, J., Condon, A., Shah, S.P.：利用深度生成模型对单细胞转录组数据进行可解释的降维。《自然-通讯》**9**(1), 1–13 (2018)[11] Wattenberg, M., Viégas, F., Johnson, I.：如何有效使用 t-SNE。Distill (2016). https://doi.org/10.23915/distill.00002[12] Schubert, E., Gertz, M.：用于可视化和离群点检测的内在 t-随机邻域嵌入。载于：相似性搜索与应用国际会议，第188–203页 (2017)。施普林格[13] Amid, E., Warmuth, M.K.：一种利用三元组实现更高全局准确性的降维方法。arXiv 预印本 arXiv:1803.00854 (2018)[14] Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.：用于改善单细胞 RNA 测序数据可视化的基于快速插值的 t-SNE。《自然·方法》**16**(3), 243–245 (2019)[15] Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.v.d., Hirn, M.J., Coifman, R.R., 等：可视化高维生物数据中的结构与转变。《自然·生物技术》**37**(12), 1482–1492 (2019)[16] Hinton, G.E., Salakhutdinov, R.R.：利用神经网络降低数据维度。《科学》**313**(5786), 504–507 (2006) https://www.science.org/doi/pdf/10.1126/science.1127647. https://doi.org/10.1126/science.1127647[17] Lin, C., Jain, S., Kim, H., Bar-Joseph, Z.：使用神经网络降低单细胞 RNA 测序数据的维度。《核酸研究》**45**(17), 156–156 (2017)[18] Wang, W., Huang, Y., Wang, Y., Wang, L.：广义自编码器：一种用于降维的神经网络框架。载于：2014年IEEE计算机视觉与模式识别会议研讨会，第496–503页 (2014). https://doi.org/10.1109/CVPRW.2014.79[19] Kohonen, T.：自组织映射。《IEEE论文集》**78**(9), 1464–1480 (1990)[20] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.：掩码自编码器是可扩展的视觉学习器。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第16000–16009页 (2022)

[21] Talwar, D., Mongia, A., Sengupta, D., Majumdar, A.：Autoimpute：基于自编码器的单细胞 RNA 测序数据插补。《科学报告》**8**(1), 1–11 (2018)[22] Xie, R., Wen, J., Quitadamo, A., Cheng, J., Shi, X.：用于基因表达预测的深度自编码器模型。《BMC基因组学》**18**(9), 39–49 (2017)[23] Szubert, B., Cole, J.E., Monaco, C., Drozdov, I.：高维单细胞数据集的保结构可视化。《科学报告》**9**(1), 1–10 (2019)[24] Kingma, D.P., Ba, J.：Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980 (2014)[25] Street, W.N., Wolberg, W.H., Mangasarian, O.L.：用于乳腺肿瘤诊断的细胞核特征提取。载于：生物医学图像处理与生物医学可视化，第1905卷，第861–870页 (1993)。SPIE[26] Aeberhard, S., Coomans, D., de Vel, O.：RDA 的分类性能。北昆士兰詹姆斯库克大学计算机科学系与数学和统计学系，技术报告，92–01 (1992)[27] Almeida, T.A., Hidalgo, J.M.G., Yamakami, A.：对短信垃圾信息过滤研究的贡献：新数据集与结果。载于：第11届ACM文档工程研讨会论文集，第259–262页 (2011)[28] Nene, S.A., Nayar, S.K., Murase, H.：哥伦比亚大学图像库（COIL-20）(1996)[29] Xiao, H., Rasul, K., Vollgraf, R.：Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。CoRR**abs/1708.07747** (2017) arXiv:1708.07747[30] Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P.V., 等：通过大规模单细胞 RNA 测序对感觉神经元类型进行无偏分类。《自然-神经科学》**18**(1), 145–153 (2015)[31] Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., 等：大规模并行数字转录谱分析单细胞。《自然-通讯》**8**(1), 1–12 (2017)[32] Seal, D.B., Das, V., De, R.K.：CASSL：一种使用半监督学习进行单细胞转录组数据细胞类型注释的方法。《应用智能》，1–19 (2022)[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.：Scikit-learn：用于机器学习的 Python。《机器学习研究杂志》**12**, 2825–2830 (2011)[34] Wolf, F.A., Angerer, P., Theis, F.J.：Scanpy：大规模单细胞基因表达数据分析。《基因组生物学》**19**(1), 1–5 (2018)