SPACEMAP: VISUALIZING HIGH-DIMENSIONAL DATA BY SPACE EXPANSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dimensionality reduction (DR) and visualization of high-dimensional data is of theoretical and practical value in machine learning and related fields. In theory, there exists an intriguing, non-intuitive discrepancy between the geometry of high-dimensional space and low-dimensional space. Based on this discrepancy, we propose a novel DR and visualization method called Space-based Manifold Approximation and Projection (SpaceMAP). Our method establishes a quantitative space transformation to address the "crowding problem" in DR. With the proposed equivalent extended distance (EED) and function distortion (FD) theory, we are able to match the capacity of high-dimensional and low-dimensional space in a principled manner. To handle complex high-dimensional data with different manifold properties, SpaceMAP makes distinctions between the near field, middle field, and far field of data distribution in a data-specific, hierarchical manner. We evaluated SpaceMAP on a range of artificial and real datasets with different manifold properties, and demonstrated its excellent performance in comparison with classical and state-of-the-art DR methods. In addition, the concept of space expansion provides a generic framework for understanding nonlinear DR methods including t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).

1 INTRODUCTION

Real-world data, including images, videos, genetic expressions, natural languages, or financial statistics, usually have a high dimensionality. The intrinsic dimensionality of these data, however, is typically much lower than its ambient dimension, which property is recognized as an important underlying reason for modern machine learning to work (Levina & Bickel, 2004; Pope et al., 2021; Wright & Ma, 2021). To capture useful information from high-dimensional data, dimensionality reduction (DR) is of both theoretical and practical value.

DR is essential for data visualization. A space with a dimensionality higher than 3, however, is already beyond our accustomed way of observing data, and our intuition in 2-dimensional or 3-dimensional space may not apply. High-dimensional space is not a trivial extension of low-dimensional space; theoretical research on high-dimensional geometry and statistics revealed a number of intriguing, non-intuitive phenomena in high-dimensional space (Giraud, 2021). Imagine a hyper-sphere with a radius r in a d-dimensional Euclidean space whose central point is at the origin. Consider a "crust" of the d-dimensional hyper-sphere, which is between the surfaces of this hyper-sphere and a slightly smaller concentric hyper-sphere with radius $(1-\epsilon)r$, where ϵ is small (Figure 1 a). The ratio of the volume of the "crust" $C_d(r)$ to the hyper-sphere is $V_d(r)$ is $\frac{C_d(r)}{V_d(r)} = 1-(1-\epsilon)^d$. Take $\epsilon = 0.01$, it is easy to show that when d is small, the ratio is tiny (as our intuition goes), however this ratio grows exponentially fast to near 100% with the increase of dimensionality, as illustrated in figure 1 a-b. The volume of a high-dimensional hyper-sphere is therefore counter-intuitively concentrated on a crust (1 c). Such concentration explains the "crowding problem" of DR (van der Maaten & Hinton, 2008): a faithful preservation of distances in high-dimensional space would lead to crowded data points.

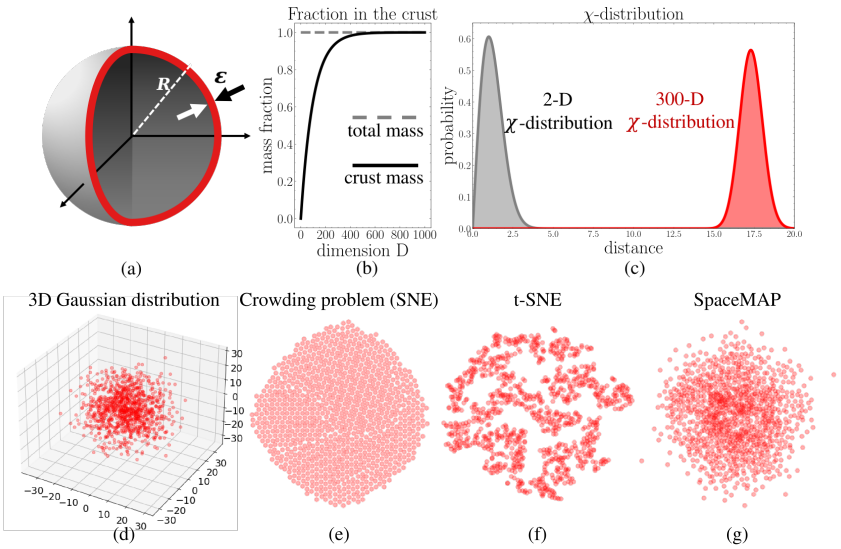

Figure 1: Upper panel: (a) The "crust" of a ball in 3-dimensional spaces with Euclidean distance metric. (b) The calculated fraction of volume between crust and sphere, with respect to the dimensionality. (c) Distributions of the distances of data points following Gaussian distribution from the origin ($\chi - distribution$) in 2-dimensional (gray) and 300-dimensional (red) spaces. Lower panel: (d) Sampling points following 3D Gaussian distribution. (e) DR result of the same data points by SNE, showing the "crowding problem" as all the points are tied together without reasonable distances and density. (f) DR by t-SNE, showing the false clusters generated by the t-distribution function. (g) DR by SpaceMAP, showing both the reasonable local fluctuation and the global density.

t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are among the most established DR methods today, both of which mitigate this crowding problem by modeling the low-dimensional similarity measure with a longer-tail distribution than high-dimensional similarity measure (e.g. t distribution vs. Gaussian in the case of t-SNE). Such a disparity in similarity measure amounts to a distortion of distance between two spaces, but in a highly *implicit* manner. Both methods perform well empirically, but the underlying distortion of distance measure in two spaces cannot be analytically expressed or validated. Furthermore, the implicity of distance transformation deters us from imposing priors in data if there is any (e.g. a Swiss Roll has an intrinsic dimensionality of 2).

Another important property of real-world high-dimensional data is that they often exhibit a hierarchical structure (sub-manifolds on large manifolds), governed by the underlying generative models. Such hierarchy demands different treatment of data points by DR at different relative position. The state-of-the-art DR methods, such as UMAP (McInnes et al., 2018) and Barnes-Hut t-SNE (van der Maaten, 2014), commonly consider on a selected neighborhood (as a hyper-parameter) while discarding the far field beyond it. Isomap (Balasubramanian et al., 2002), on the other hand, taking both near and far field into account by calculating distances on a connected graph. t-SNE and UMAP works well on data of disjoint manifolds such as MNIST, while Isomap works well on data of continuous manifold such as Swiss roll. However, it is generally difficult for one method to succeed in both disjoint and continuous data manifolds.

In this paper, we seek to develop a novel DR method that is based on the following two key ideas, and strive to address the two aforementioned issues:

- *Space Expansion:* Matching the "capacity" in high-dimensional and low-dimensional space, by explicit, quantitative transformation of distance measure;
- *Hierarchical Manifold Approximation:* Data-specific, hierarchical modeling of similarities in high-dimensional data, to accommodate both disjoint and continuous data manifolds.

2 RELATED WORK

Over the past decades, there have been active development of DR methods, which can be roughly categorized into global and local methods. The most well-known global method is principal component analysis (PCA) (Jolliffe, 2014), which finds a low number of directions that accounts for most variations. While PCA is linear, Laplacian eigenmaps (Belkin & Niyogi, 2001; 2003) and multidimensional scaling (MDS) (Buja et al., 2008) are nonlinear global methods. Local methods, including Isomap, Locally Linear Embedding (LLE) (Roweis, 2000), stochastic neighbor embedding (SNE) (Hinton & Roweis, 2003), and maximum variance unfolding (MVU) (Weinberger & Saul, 2006), focus on preserving local structure in data. While outperforming global methods on nonlinear datasets, local methods generally suffer from poor scalability, low speed, and stochasticity.

In 2008, t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008), a variant of SNE, was proposed and later became one of the most popular DR methods in research community. t-SNE alleviates the "crowding problem" by assuming t-distribution in low-dimensional space (in contrast to Gaussian distribution in the high-dimensional space). t-SNE is followed up by a number of prominent work to improve its scalability and speed, including among others Barnes-Hut t-SNE (BH t-SNE) (van der Maaten, 2014), fast Fourier-transform-accelerated interpolation-based t-SNE (FIt-SNE) (Linderman et al., 2019), and parametric t-SNE (ptSNE) (van der Maaten, 2009).

Recently, uniform manifold approximation and projection McInnes et al. (2018) (UMAP) was proposed, which shows competitive performance to t-SNE. UMAP is grounded in the mathematical theories of Riemannian geometry and algebraic topology. It computes local neighbor graph as a fuzzy simplicial set, which glues together the high-dimensional sparse data and differentiates local and global topology. UMAP further incorporates a number efficient algorithms such as approximated nearest neighbor and stochastic gradient descent (SGD) to improve speed and scalability. TriMAP (Amid & Warmuth, 2019) was recently proposed, which uses a triplet loss instead of pairwise. UMAP and TriMAP are reported to better preserve the global structure than t-SNE McInnes et al. (2018); Amid & Warmuth (2019).

We note that a different definition of "distance" in the low-dimensional space than in the high-dimensional space is core to these successful DR algorithms. In t-SNE, with a heavier tail, t-distribution implicitly expands the low dimensional space distance metrics, to allow more space for data points to be arranged. Likewise, the definition of distances of UMAP in low-dimensional space is much more heavy-tailed (inverse of polynomial) than in high-dimensional space (exponential), which also leads to implicit distortion of space.

3 METHOD

3.1 SPACE EXPANSION

High-dimensional space, irrespective of the distances defined, possesses exponentially higher capacity to express data than low-dimensional space. Here we define "capacity" as a measure of space volume to accommodate data points. Our intuition is therefore to *expand* the low-dimensional space to make up for this space capacity lost due to DR. To transfer the high-dimensional geometry into a low-dimensional one, we introduce the concept of *equivalent extended distance (EED)* in Section 3.1.1 and *function distortion (FD)* in Section 3.1.2.

3.1.1 EQUIVALENT EXTENDED DISTANCE (EED)

We first define space capacity to quantitatively calculate the amount of distortion needed to match low-dimensional space to high-dimensional space.

Definition 3.1 (Space Capacity)**.** *Let $R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$ be the distance between data point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the D-dimensional space. The space capacity $\mathcal{V}_D(R_{ij})$ from point i to point j is defined as the volume a D-dimensional ball with a radius R_{ij}.*

In Euclidean space, the capacity $\mathcal{V}_D(R_{ij})$ is simply the volume of a D-dimensional hyper-sphere $S_D(R_{ij})$:

$$\mathcal{V}_D(R_{ij}) = S_D(R_{ij}) = \frac{\pi^{D/2}}{\Gamma(D/2+1)} R_{ij}^D \tag{1}$$

As high-dimensional space has higher capacity than low-dimensional space, to preserve the capacity of the high-dimensional space in the low-dimensional space, the equivalent distance in the low-dimensional space naturally extends, hence Equivalent Extended Distance (EED):

Definition 3.2 (Equivalent Extended Distance (EED)). *Let $R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$ be the distance between data point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the D-dimensional space. The equivalent extended distance (EED) $\tilde{\mathcal{R}}_{ij,D\to d}$ is defined as the equivalent distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in d-dimensional space which can reach the same Space Capacity:*

$$\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij}) \tag{2}$$

Example 3.1 (2-dimensional EED in Euclidean space). *Consider a D-dimensional Euclidean space, the distance between two points i and j is $R_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. The 2-dimensional EED of R_{ij} is: $\tilde{\mathcal{R}}_{ij,D\to 2} = \sqrt{\frac{S_D(R_{ij})}{\pi}} = \alpha R_{ij}^{D/2}$ where $\alpha = \frac{\pi^{(D-2)/4}}{\sqrt{\Gamma(D/2+1)}}$ and Γ is the gamma function.*

Example 3.1 shows that when embedding D-dimensional Euclidean space into d-dimensional space, EED can be expressed in a generic form of $\tilde{\mathcal{R}}_{ij,D\to d} = \alpha R^{\beta}$ where $\beta = \frac{D}{d}$ and α is a constant determined by D and d.

In the following, we present prior theory of intrinsic dimension by the maximum likelihood estimation (MLE) method by Levina & Bickel (2004), and show that our proposed EED fits in this classical framework.

Theorem 3.1 (MLE of the intrinsic dimension (Levina & Bickel, 2004)). *Let $R_{ik} \in \mathbb{R}$ be the Euclidean distance between point i and its k-th nearest neighbor. Assume the distribution of the data points in the small sphere $S_i(R)$ is uniform. The maximum likelihood estimation (MLE) of the intrinsic dimension of the data around point i is:*

$$\hat{d}_i(R_{ik}) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{R_{ik}}{R_{ij}} \right)^{-1} \tag{3}$$

The MLE method provides a way to estimate the intrinsic dimension at point i in its k neighborhood based on the distribution of distances.

By applying EED transformation to a space with intrinsic dimension D, the following Lemma holds:

Lemma 3.2 (Transformation of the intrinsic dimension by applying EED). *$\forall R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$, if the MLE of the intrinsic dimension of the high-dimensional dataset around point i is $\hat{d}_i(R_{ij}) = D$, then the MLE of the intrinsic dimension after applying EED is $\hat{d}_i(\tilde{\mathcal{R}}_{ij,D\to d}) = d$.*

Proof. By replacing R_{ij} with $\tilde{\mathcal{R}}_{ij,D\to d} = \alpha R_{ij}^{D/d}$ into Equation (3), we have

$$\hat{d}_i(\tilde{\mathcal{R}}_{ik,D\to d}) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{\alpha R_{ik}^{D/d}}{\alpha R_{ij}^{D/d}} \right)^{-1} = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \frac{D}{d} \log \frac{R_{ik}}{R_{ij}} \right)^{-1} = \frac{d}{D} \hat{d}_i(R_{ik}) = d \tag{4}$$

□

Specifically, when $d = 2$ for visualization purposes, the MLE of the intrinsic dimension of the expanded space will be exactly 2. Essentially, the validity of this statement arises from the proposed EED transformation $\tilde{\mathcal{R}}_{ij,D\to d} = \alpha R^{\frac{D}{d}}$, where the exponential form of $\frac{D}{d}$ transforms the intrinsic dimension from D to d, based on the MLE formula (3) where the distances are taken logarithmic. In the following sections, the distance R_{ij} between point i and point j is described as $l(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which is suitable for several different distance metrics.

3.1.2 Function Distortion (FD)

Function distortion (FD) is the implementation to realize EED between high-dimensional and low-dimensional space similarity measures during optimization. Given that EED is well defined as

$\tilde{\mathcal{R}}_{ij,D\to d} = \alpha R^{\beta}$, FD between spaces can be explicitly computed. Figure 2 illustrates how FD transforms the similarity functions ($P_{j|i}$ and Q_{ij}) in high-dimensional and low-dimensional spaces. A much heavier tail can be observed in the low-dimensional space similarity measure, arising from EED.

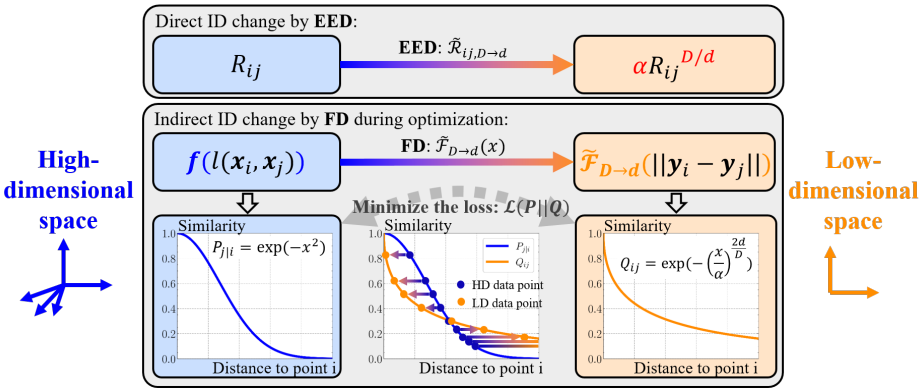

Figure 2: EED is the direct transformation of distance R_{ij}, e.g. distance between x_i and x_j, also expressed by $l(\boldsymbol{x}_i, \boldsymbol{x}_j)$, while FD integrates EED into the similarity measure, which is used for optimization. A much heavier tail can be observed in the low-dimensional similarity measure after FD (bottom-right). By matching the similarity in high-dimensional (blue) and low-dimensional (orange) spaces, we intuitively show how points are moved around by the optimization process (bottom-middle).

3.2 Hierarchical Manifold Approximation

Another fundamental assumption of SpaceMAP is that real world data often lie on a complex manifold, of a hierarchical structure (Figure 3 a). In SpaceMAP, we divided the high-dimensional space with respect to each data point i into near field $\mathbb{S}_{i,near}$, middle field $\mathbb{S}_{i,middle}$, and far field $\mathbb{S}_{i,far}$. Such a definition of neighborhood is point-specific, implying that we can treatment different points differently, depending on their relative location on the manifold. An analogue can be made to UMAP, which defines a different distance metric for each data point depending on its nearest neighbor. Similarly, for a generic formulation, the determination of the fields is based on K-nearest neighbors in SpaceMAP:

$$\mathbb{S}_{i,near} = \{\boldsymbol{x}_j \,|\, l(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq l(\boldsymbol{x}_i, \boldsymbol{x}_{k_1})\}, \quad k_1 = n_{near} \tag{5}$$

$$\mathbb{S}_{i,middle} = \{\boldsymbol{x}_j \,|\, l(\boldsymbol{x}_i, \boldsymbol{x}_{k_1}) < l(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq l(\boldsymbol{x}_i, \boldsymbol{x}_{k_2})\}, \quad k_2 = n_{near} + n_{middle} \tag{6}$$

$$\mathbb{S}_{i,far} = \{\boldsymbol{x}_j \,|\, l(\boldsymbol{x}_i, \boldsymbol{x}_j) > l(\boldsymbol{x}_i, \boldsymbol{x}_{k_2})\} \tag{7}$$

where n_{near} is the number of the nearest neighbors in near field, and n_{middle} is the number of nearest neighbors in the middle field. Like UMAP, we can use a fixed number to define the neighborhood by empirically setting $n_{near} = 20$, $n_{middle} = 50$, or setting the number based on our prior knowledge on data. Additional experiments in Appendix (Figure 10) shows that the results are not particularly sensitive to these two parameters.

To describe hierarchical manifold $\mathcal{M}$, two intrinsic dimensionalities, namely, $d_{local}(i)$ of the near field of point i, and d_{global} of the middle and far field of point i, can be calculated based on the MLE method with n_{near} and n_{middle} set. (Alternatively, d_{local} or d_{global} can also be set manually with prior knowledge of data, e.g. the d_{local} and d_{global} of a Swill Roll dataset are both 2. However in this work, for simplicity, we used n_{near} and n_{middle} to calculate the two dimensionalities.)

Base on d_{global} and d_{local}, the conditional similarities of the data points $P_{j|i}$ are defined as:

$$P_{j|i} = \begin{cases} \exp\left(-\left(\frac{l(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\alpha_i}\right)^{\frac{2d_{local,i}}{d}}\right), & x_j \in \mathbb{S}_{near} \\ \exp\left(-\frac{\left(l(\boldsymbol{x}_i,\boldsymbol{x}_j) - l(\boldsymbol{x}_i,\boldsymbol{x}_{k_1}) + \sqrt{-\ln \eta}\right)^2}{\sigma_i}\right), & x_j \in \mathbb{S}_{middle} \\ 0, & x_j \in \mathbb{S}_{far} \end{cases} \tag{8}$$

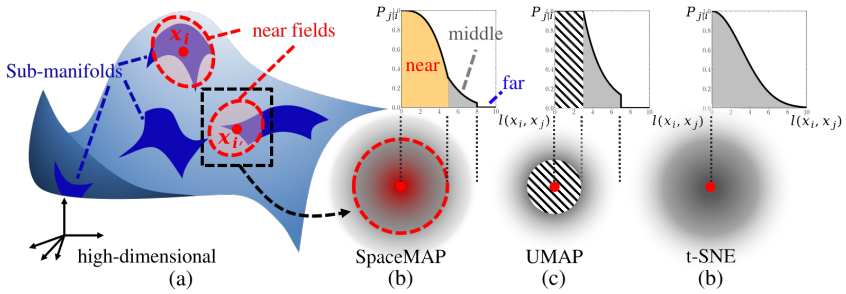

Figure 3: (a) Illustration of a hierarchical manifold: sub-manifolds (dark blue), on which the data points are distributed, are within the global manifold with different geometric properties, reflecting the necessity of local-specific and hierarchical similarity functions. (b) The definition of near/middle/far fields and the shape of the corresponding similarity functions (high-dimensional similarity kernel) of SpaceMAP. UMAP (c) glues the nearest neighbor of each point with similarity 1. (d) The Gaussian similarity kernel of t-SNE.

where α_i, d and d_{local} denotes FD in the near field, $\eta \in (0, 1)$ is a constant representing $P_{k_1|i}$ at the border of near field and middle field. In Appendix (Figure 11) we tested a range of η values and showed the choice of its value is not critical to the final results. σ_i is the normalization factor of the middle field $\mathbb{S}_{middle}$ defined in the similar way as in UMAP (McInnes et al., 2018):

$$\sum_{j=k_1+1}^{k_2} P_{j|i} = \eta \log(n_{middle}) \tag{9}$$

The operation in equation amounts to hierarchical space expansion in the near field and middle field. The definition is in a large part for the ease of computation, as by modeling P instead of Q, we keep the expression of Q as simple as possible for better differentiality during optimization.

The resulting hierarchical similarity function is illustrated in Figure 3, where the computation of $P_{j|i}$ in near field, middle field, and far field are shown for SpaceMAP (b), UMAP (c), and t-SNE (c), respectively. It can be seen from the figure that t-SNE makes no distinction between fields, with a uniform Gaussian function, while UMAP differentiates three fields, but with a uniform profile for near field (1 nearest neighbor). In comparison, SpaceMAP has a subtle perception of both near and middle fields.

To symmetrise the conditional similarity $P_{j|i}$ and $P_{i|j}$ into a symmetric metric P_{ij}, we use the average between the two: $P_{ij} = \frac{P_{j|i}+P_{i|j}}{2}$. We next define the the pairwise similarities in the low-dimensional space:

$$Q_{ij} = \exp\left(-(\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|}{\alpha})^{\frac{2d}{d_{global}}}\right) \tag{10}$$

Where α, d and d_{global} denotes FD in the middle field. Figure 2.a shows the exact effect of our method to model $P_{j|i}$ and Q_{ij}, where the high-dimensional neighbors of point i (blue points on $P_{j|i}$ function) are distorted to match the pattern of geometry in low-dimensional space during optimization, which is illustrated in the next section.

3.3 Optimization Pipeline

The loss function of SpaceMAP is defined in a similar way as in UMAP:

$$\mathcal{C} = \mathcal{L}(P_{ij}\|Q_{ij}) = \sum_i \sum_j \left[P_{ij} \log \frac{P_{ij}}{Q_{ij}} + (1 - P_{ij}) \log \frac{1 - P_{ij}}{1 - Q_{ij}}\right] \tag{11}$$

where the first term is the KL divergence between two probability distributions, and the second term generates the repulsive force between point pairs, contributing to the global structure preservation. For stochastic gradient descent, the derivative of the loss with respect to the coordinates of each point in low-dimensional space $\frac{\partial \mathcal{C}}{\partial \boldsymbol{y}_i}$ is calculated and divided into two parts during iteration, which

can be seen as the attractive force $F_{ij,attractive}$ and the repulsive force $F_{ij,repulsive}$ between point pairs:

$$F_{ij,attractive} = -\left[\frac{2d}{d_{global}\alpha^{2d/d_{global}}}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^{(\frac{2d}{d_{global}}-2)}P_{ij}\right](\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{12}$$

$$F_{ij,repulsive} = \left[\frac{2d}{d_{global}\alpha^{2d/d_{global}}}\frac{Q_{ij}(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|)}{1 - Q_{ij}(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|)}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^{(\frac{2d}{d_{global}}-2)}(1 - P_{ij})\right](\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{13}$$

In the implementation, the neighbors to generate the forces can be chosen as the full batch or stochastically (therefore stochastic gradient descent, SGD) to achieve scalability of SpaceMAP to large datasets. The overall algorithm of SpaceMAP is describe in Appendix A.1 Algorithm 1 followed by the detailed functions.

4 Experiments and Results

4.1 Datasets and Evaluation Metrics

We applied the proposed SpaceMAP to a variety of datasets and computed quantitative metrics to evaluate the performance of DR. The method is benchmarked by other classical or state-of-the-art methods including PCA, Laplacian Eigenmaps, t-SNE, and UMAP.

We tested on a large range of high-dimensional datasets, including the standard MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), Swiss Roll (Van der Maaten et al., 2007) (Here we artificially created a hole on the Swiss Roll to challenge the DR methods if they can preserve the local property on a continuous manifold), COIL-20 (Nene et al., 1996), RNA-seq (Tasic et al., 2018), and an experimental cardiac MRI dataset from three different vendors. Examples of cardiac MRI are given in Fig. 14. The details of the chosen datasets are in Appendix A.6. Here the original data is simply vectorized and taken as input of DR. We also tested all methods on the GoogleNews Word2Vec 3 Million dataset (Mikolov et al., 2013). For quantitative evaluation, we computed the 20-fold cross-validated k-nearest neighbor classifier accuracy (KNN accuracy), trustworthiness, continuity, Shepard goodness, and normalized stress to evaluate both the local and global structure preservation of the datasets (Espadoto et al., 2021; Nonato & Aupetit, 2019). KNN accuracy measures the local structure preservation along different neighborhood scale k. Trustworthiness and continuity evaluate the local pattern of the embedding by calculating the true neighbor rate and the missing neighbor rate. The Shepard goodness and the normalized stress are two measurements of the goodness in global structure preservation. The detailed introduction of these evaluation metrics is in Appendix A.7.

4.2 Qualitative and Quantitative Results

We present the visualization results of SpaceMAP in Figure 4, along with PCA, Laplacian Eigenmap, t-SNE, and UMAP. It can be seen that SpaceMAP nicely preserved the structures of the datasets on both local and global scale, for real-world datasets with disjoint manifolds (MNIST, Fashion-MNIST), as well as for synthetic data of continuous manifold (row 3: Swiss roll with a hole). The space extension from dimensionality 3 to 2 nicely preserves the continuous manifold, while for t-SNE and UMAP the degree of extension is more difficult to control, resulting in artificial clusters. Figure 7 shows SpaceMAP DR results of GoogleNews Word2Vec 3 million data, in comparison with UMAP. More distinctive clusters are observed in the SpaceMAP result, while a zoomed-in view clearly shows the natural clustering of geographical semantics. Again the improvement can be attributed to the strategy to explicitly expand the low-dimensional space, as in UMAP (or t-SNE) the extent of expansion is difficult to calculate or control.

We further compare the quantitative measure of DR performance of all DR methods using the metrics described before. Figure 5 and Table 1 shows the superior, or non-inferior performance of SpaceMAP compared to the classical and state-of-the-art methods in most datasets. As the optimization pipeline resembles that of UMAP, the computation time of SpaceMAP is comparable to UMAP.

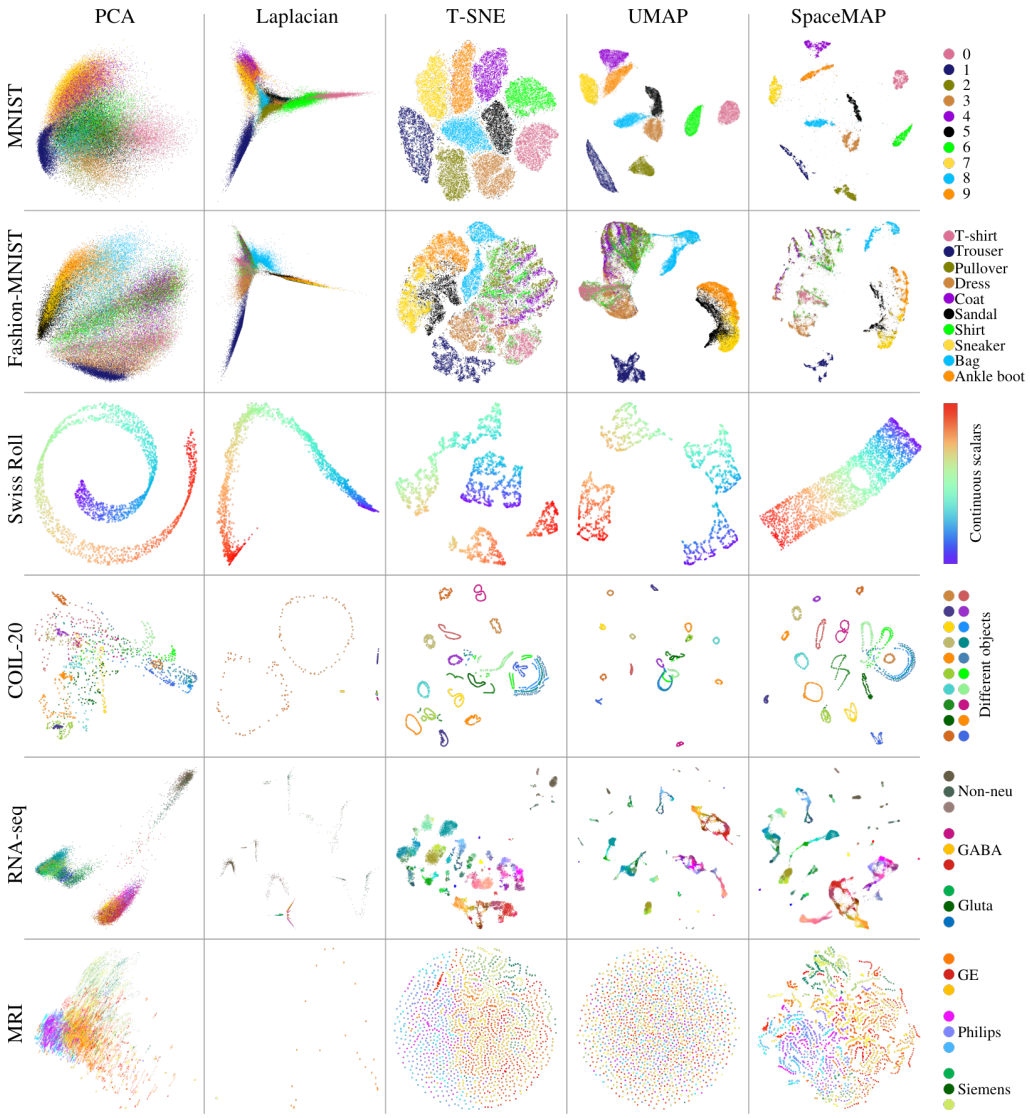

Figure 4: Results 2D visualization by SpaceMAP compared to PCA, Laplacian eigenmaps, t-SNE and UMAP on different datasets. The detailed implementations of the datasets and the hyper-parameters are introduced in Appendix A.6.

5 Discussion and Conclusion

We propose a new DR method called SpaceMAP, which, in principle, can map data of any dimensionality onto a 2-dimensional space for visualization with the calculated *space expansion*. Different from established methods such as t-SNE or UMAP that perform implicit transformation of distances, we analytically derived a quantitative EED transformation of distances between high-dimensional space and low-dimensional spaces. We further show that the EED transformation fits in the classical framework of MLE of intrinsic dimension, effectively altering the intrinsic dimension thereby realizing low-dimensional mapping.

We argue that all successful DR methods, including among others t-SNE and UMAP, make use of the rationale of *space expansion* to enable data visualization in a space of drastically reduced dimension. However, previous methods did such transformations in a highly implicit manner, with the rules concealed in the self-defined similarity measure (e.g. t-distribution in t-SNE or inverse of

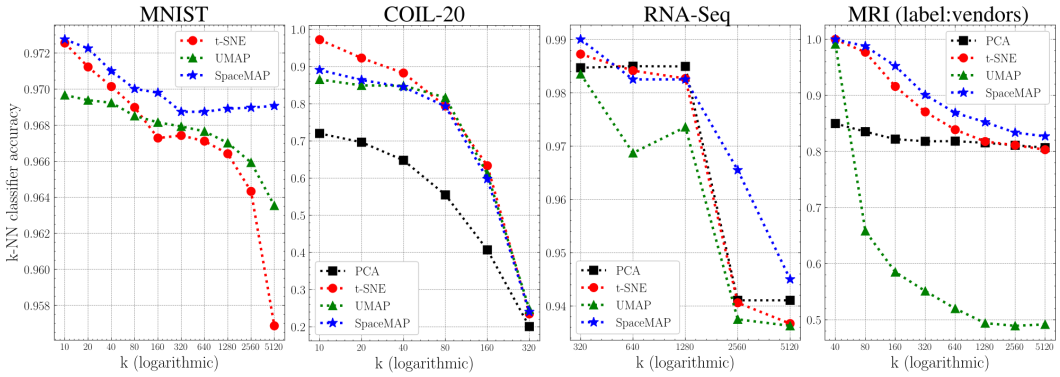

Figure 5: 20-fold cross-validated k-NN classifier accuracy as a function of neighborhood size (k) on different datasets. From left to right: MNIST, COIL-20, RNA-seq, and multi-vendor MRI. In general, SpaceMAP outperforms UMAP in different neighborhood sizes (k), while t-SNE is competitive to SpaceMAP when the neighborhood size (k) is small.

Table 1: Quantitative measure of DR performance by SpaceMAP and other reference methods: M_t, M_c, M_s and M_σ indicate the trustworthiness, continuity, Shepard goodness and normalized stress, respectively. M_t and M_c are local metrics (left side of the table) while M_s and M_σ are global metrics (right side of the table) for the evaluation of the DR performance. Detailed introduction of the metrics is in Appendix A.7.

Local Metrics (M_t and M_c)

Experiments		PCA	Laplacian	t-SNE	UMAP	SpaceMAP
MNIST	M_t	0.74	0.81	**0.98**	0.96	0.97
	M_c	0.94	0.93	0.97	0.97	**0.98**
FMNIST	M_t	0.91	0.89	0.98	0.98	**0.99**
	M_c	0.98	0.87	0.98	**0.99**	**0.99**
COIL-20	M_t	0.86	0.92	0.99	0.99	**1.00**
	M_c	0.93	0.79	0.99	0.99	**1.00**
RNA-Seq	M_t	0.89	0.85	**1.00**	0.99	**1.00**
	M_c	0.98	0.94	0.99	**1.00**	**1.00**
MRI	M_t	0.90	0.59	0.99	**1.00**	**1.00**
	M_c	**1.00**	0.51	0.99	**1.00**	**1.00**

Global Metrics (M_s and M_σ)

Experiments		PCA	Laplacian	t-SNE	UMAP	SpaceMAP
MNIST	M_s	**0.50**	0.23	0.35	0.32	0.35
	$1-M_\sigma$	0.42	0.33	0.54	0.53	**0.55**
FMNIST	M_s	**0.88**	0.41	0.64	0.58	0.67
	$1-M_\sigma$	0.65	0.36	0.66	0.62	**0.67**
COIL-20	M_s	**0.89**	0.72	0.80	0.56	0.61
	$1-M_\sigma$	0.55	0.19	**0.62**	0.58	0.60
RNA-Seq	M_s	**0.80**	0.11	0.61	0.22	0.63
	$1-M_\sigma$	**0.60**	0.15	0.59	0.50	0.52
MRI	M_s	**0.59**	0.28	0.22	0.03	0.37
	$1-M_\sigma$	0.54	0.00	0.53	0.50	**0.58**

polynomial in UMAP, with different parameters). Despite their empirical success, we posit that an analytical form of distance transformation is desirable to deal with situations where we would like to take more control of the DR results or impose prior knowledge for DR.

SpaceMAP further differentiates different range of neighborhoods to model the hierarchical structure existent in many real-world datasets. SpaceMAP is generic and has a limited number of hyperparameters, with the most important ones related to the selection of number of nearest neighbors in the near field and middle field. We observed that the final results are not particularly sensitive to the selection of these parameters (Appendix Figure 10).

In conclusion, we have proposed a new DR method, SpaceMAP, which is based on a principled way to explicitly transform distances in high- and low-dimensional spaces, and models the hierarchical structure of data based on the intrinsic dimension of local and global manifolds. Our experiments on a diverse range of datasets demonstrated its excellent performance in comparison with other state-of-the-art DR methods.

REFERENCES

Ehsan Amid and Manfred K Warmuth. Trimap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.

Mukund Balasubramanian, Eric L Schwartz, Joshua B Tenenbaum, Vin de Silva, and John C Langford. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Nips*, volume 14, pp. 585–591, 2001.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, jun 2008. doi: 10.1198/106186008x318440.

Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, mar 2021. doi: 10.1109/tvcg.2019.2944182.

Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

Ian Jolliffe. Principal component analysis, sep 2014.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pp. 777–784, Cambridge, MA, USA, 2004. MIT Press.

George C. Linderman, Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245, feb 2019. doi: 10.1038/s41592-018-0308-4.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. February 2018.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). 1996.

Luis Gustavo Nonato and Michael Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, aug 2019. doi: 10.1109/tvcg.2018.2846735.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

S. T. Roweis. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, dec 2000. doi: 10.1126/science.290.5500.2323.

Bosiljka Tasic, Zizhen Yao, Lucas T. Graybuck, Kimberly A. Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N. Economo, Sarada Viswanathan, Osnat Penn, Trygve Bakken, Vilas Menon, Jeremy Miller, Olivia Fong, Karla E. Hirokawa, Kanan Lathia, Christine Rimorin, Michael Tieu, Rachael Larsen, Tamara Casper, Eliza Barkan, Matthew Kroll, Sheana Parry, Nadiya V. Shapovalova, Daniel Hirschstein, Julie Pendergraft, Heather A. Sullivan, Tae Kyung Kim, Aaron Szafer, Nick Dee, Peter Groblewski, Ian Wickersham, Ali Cetin, Julie A. Harris, Boaz P. Levi, Susan M. Sunkin, Linda Madisen, Tanya L. Daigle, Loren Looger, Amy Bernard, John Phillips, Ed Lein, Michael Hawrylycz, Karel Svoboda, Allan R. Jones, Christof Koch, and Hongkui Zeng. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, oct 2018. doi: 10.1038/s41586-018-0654-5.

Laurens van der Maaten. Learning a parametric embedding by preserving local structure. In David van Dyk and Max Welling (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 384–391, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/maaten09a.html.

Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014. URL http://jmlr.org/papers/v15/vandermaaten14a.html.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Laurens Van der Maaten, Eric O Postma, and Hendrik J van den Herik. Matlab toolbox for dimensionality reduction. *MICC, Maastricht University*, 2007.

Kilian Q Weinberger and Lawrence K Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, volume 6, pp. 1683–1686, 2006.

John Wright and Yi Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A Appendix

A.1 SpaceMAP Pseudo Code

Algorithm 1 SpaceMAP algorithm

function SpaceMAP($X, n_{near}, n_{middle}, near\_field\_range, \eta, n\_epochs, learning\_rate$)
  $knn\_dist \leftarrow \text{NearestNeighbor}(X, n_{near}, n_{middle})$
  $d_{local}, d_{middle} \leftarrow \text{MLEIntrinsicDimension}(knn\_dist, n_{near}, n_{middle})$
  for $\boldsymbol{x}_i \in X$ **do** ▷
    $\alpha'[i] \leftarrow \text{FDScalingFactor}(knn\_dist[i, n_{near}], \eta, d_{local}[i], 2)$
  end for
  $d_{global} \leftarrow \frac{1}{N}\sum_i^N d_{middle}[i]$
  $\alpha \leftarrow \text{FDScalingFactor}(near\_field\_range, d_{global})$
  $P_{ij} \leftarrow \text{HierarchicalManifoldP}(knn\_dist, n_{near}, n_{middle}, d_{local}, d_{global}, \alpha')$
  $Y \leftarrow \text{InitializeEmbedding}(X)$
  $Y \leftarrow \text{OptimizeEmbedding}(Y, P_{ij}, d_{local}, d_{global}, \alpha, learning\_rate, n\_epochs)$
  return Y
end function

Algorithm 2 Maximum likelihood estimation (MLE) of the intrinsic dimension

function MLEINTRINSICDIMENSION($knn\_dist, n_{near}, n_{middle}$)
  $k_1 \leftarrow n_{near}$
  $k_2 \leftarrow n_{near} + n_{middle}$
  for $i \leftarrow 1, ...,$n-data **do**
    $R_{i,k_1} \leftarrow knn\_dist[i, k_1]$
    $R_{i,k_2} \leftarrow knn\_dist[i, k_2]$
    $\hat{d}_{near}[i] \leftarrow (\frac{1}{k_1-1}\sum_{j=1}^{k_1-1}\log\frac{R_{i,k_1}}{R_{i,j}})^{-1}$
    $\hat{d}_{middle}[i] \leftarrow (\frac{1}{k_2-1}\sum_{j=1}^{k_2-1}\log\frac{R_{i,k_2}}{R_{i,j}})^{-1}$
  end for
  return $\hat{d}_{near}, \hat{d}_{middle}$
end function

Algorithm 3 Calculation of the FD scaling factor

function FDSCALINGFACTOR(r, η, D, d)
  $\beta \leftarrow \frac{D}{d}$
  $\alpha \leftarrow \frac{r}{(-\ln\eta)^{\beta/2}}$ ▷ α is chosen to satisfy $\tilde{\mathcal{F}}_{D\to d}(r) = \eta$ or $\tilde{f}_{d\to D}(r) = \eta$
  return α
end function

Algorithm 4 Hierarchical Manifold Approximation in high-dimensional space (P_{ij} calculation)

function HIERARCHICALMANIFOLDP($knn\_dist, n_{near}, n_{middle}, d_{local}, d_{global}, \alpha'$)
  for $i \leftarrow 1, ...,$n-data **do**
    for $r_{ij} \in knn\_dist[i]$ **do** ▷ knn_dist is arranged in ascending order
      if $j \le n_{near}$ **then** ▷ $P_{j|i}$ in near field
        $P_{j|i} \leftarrow \exp\left(-(\frac{r_{ij}}{\alpha'[i]})^{d_{local}}\right)$
      else if $j \le n_{near} + n_{middle}$ **then** ▷ $P_{j|i}$ in middle field
        $r \leftarrow r_{ij} - knn\_dist[i, n_{near}]$
        Binary search for σ_i such that $\sum_{j=n_{near}+1}^{n_{middle}} e^{(-\frac{(r-\sqrt{-\ln\gamma})^2}{\sigma_i})} = \eta\log_2(n_{middle})$
        $P_{j|i} \leftarrow \exp\left(-\frac{r_{ij}^2}{\sigma_i}\right)$
      else
        $P_{j|i} \leftarrow 0$
      end if
    end for
  end for
  for all $P_{j|i}$ **do**
    $P_{ij} = \frac{P_{j|i}+P_{i|j}}{2}$ ▷ Symmetrization of the similarities
  end for
  return P_{ij}
end function

A.2 MORE SPACEMAP: HIGH-DIMENSIONAL DATA VISUALIZATION

A.3 VISUALIZATION OF LARGE DATASETS

A.4 OTHER EXPERIMENTS

A.5 MNIST VISUALIZATION

A.6 IMPLEMENTATION

A.7 INTRODUCTION OF THE METRICS

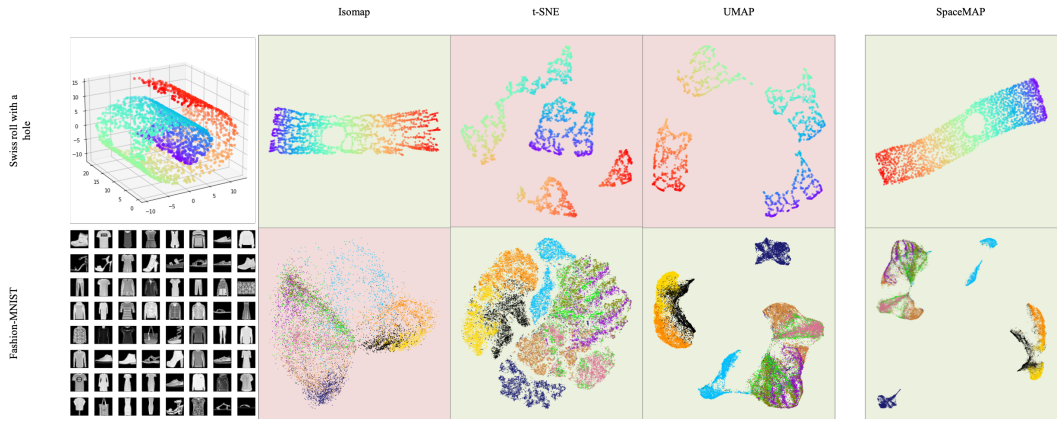

Figure 6: Upper panel: Swill Roll with a hole. Lower panel: Fashion MNIST. SpaceMAP can handle both continuous and disjoint manifolds, while Isomap, t-SNE, and UMAP are more sensitive to the manifold properties. In particular, both t-SNE and UMAP resulted in artificial clusters, broken at the location of the hole.

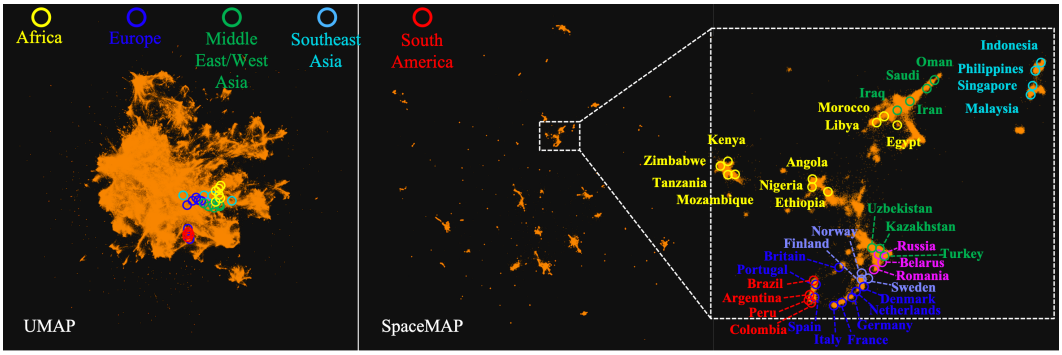

Figure 7: Visualization of the GoogleNews Word2Vec 3 million dataset by UMAP (left) and SpaceMAP (middle), where the SpaceMAP result is zoomed in for better visualization of the word semantics (right).

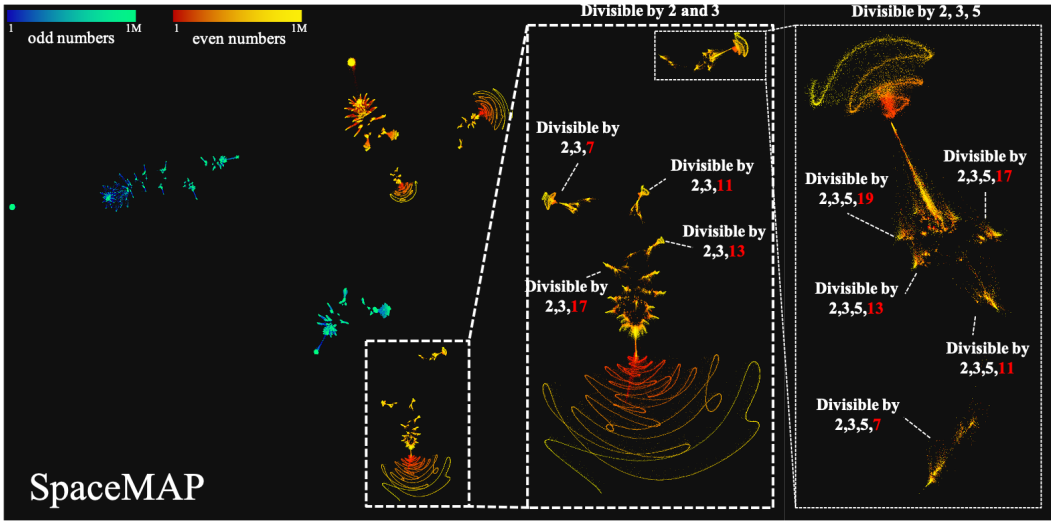

Figure 8: Embedding integers from 0 to 1,000,000, as represented by the sparse binary vectors of prime number divisibility, as described in UMAP (McInnes et al., 2018). Sub-structures of data can be identified, which corresponds to divisibility of multiple prime numbers (zoomed-in windows).

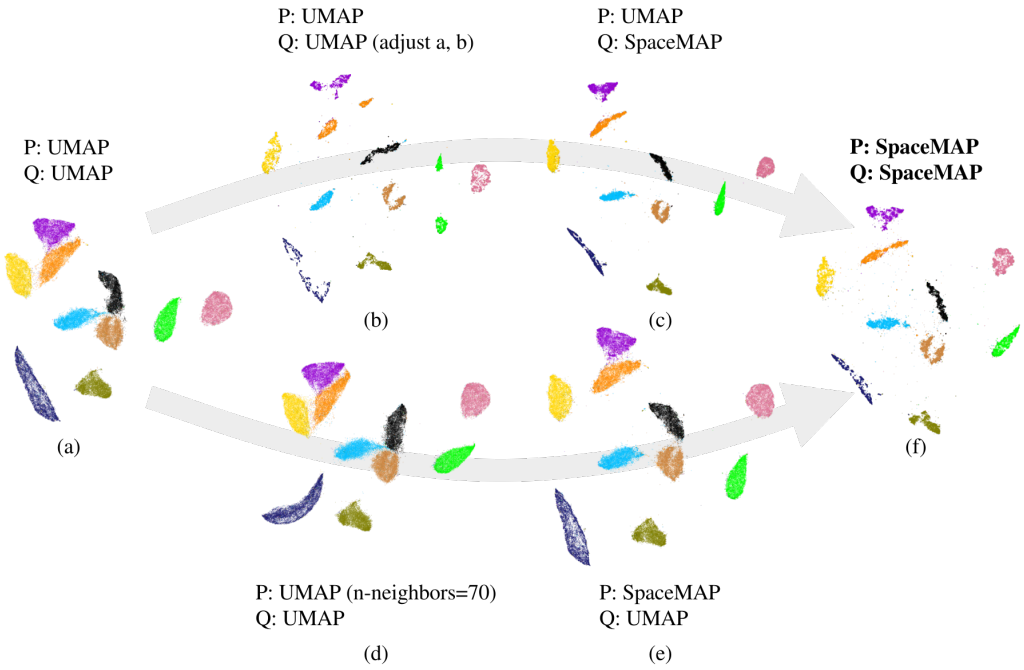

Figure 9: Ablation study to compare the definition of similarity functions in SpaceMAP and UMAP, on the MNIST dataset. Following the gray arrow, we adapted the P and Q from UMAP to SpaceMAP. It can be observed that our proposed multi-scale P (differentiating near field and middle field) had a significant influence on the final visualization. Compared to UMAP, SpaceMAP resulted in better separation of the clusters, and less uniform intra-class distribution. We argue that the uniform distribution within classes, as promoted by the UMAP rationale, can be artificial as data often have sub-structures within a class. See Figure 13 for a zoomed-in view of MNIST handwritten digits.

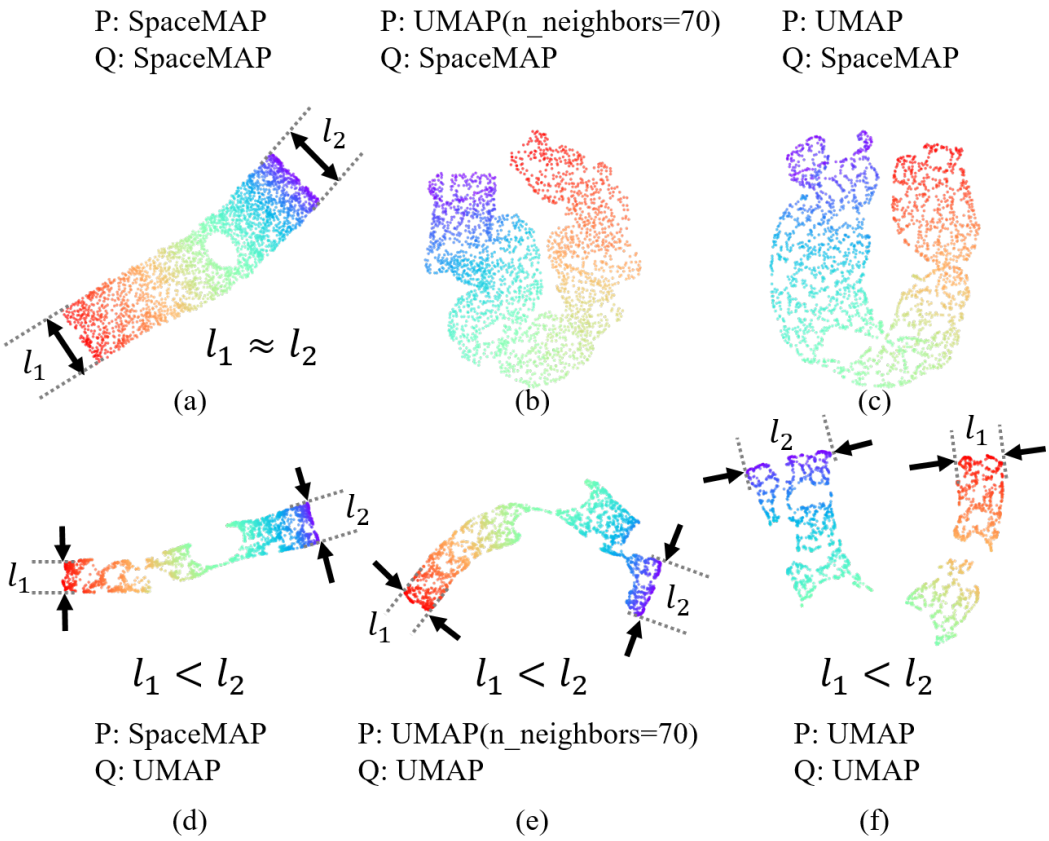

Figure 10: Ablation study to compare the definition of similarity functions in SpaceMAP and UMAP, on the Swiss-Roll-with-a-Hole dataset. Different combinations of P and Q were tested. Here we observed that SpaceMAP potentially better preserves the geometry of the manifold than UMAP. As the Swiss Roll has a denser distribution of points on the purple end than on the red end, UMAP resulted in a wider band on the purple end, promoted by the uniform approximation.

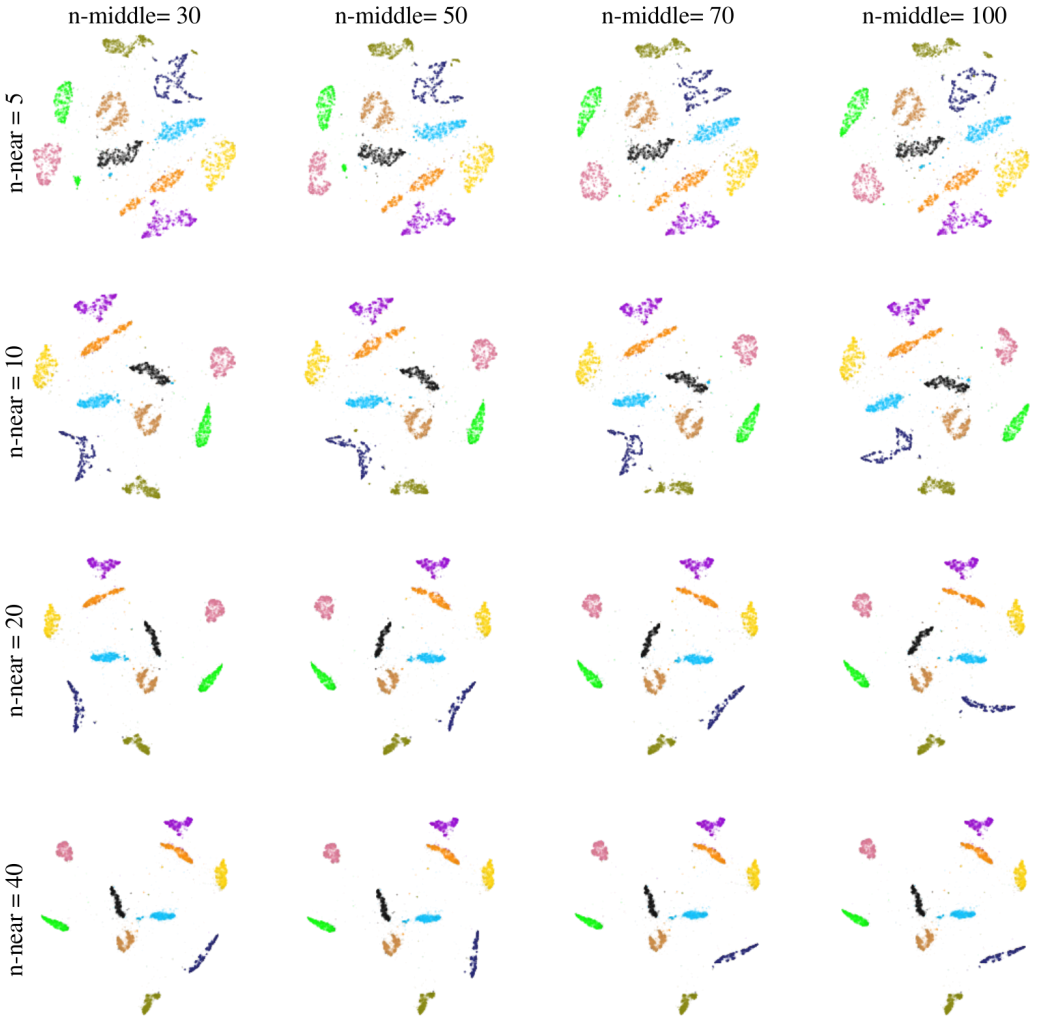

Figure 11: Hyper-parameter selection in SpaceMAP: the influence of n_{near} and n_{middle}.

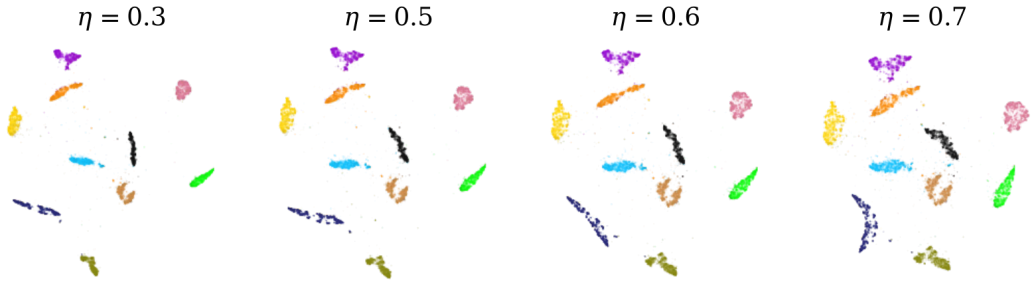

Figure 12: Hyper-parameter selection in SpaceMAP: the influence of η.

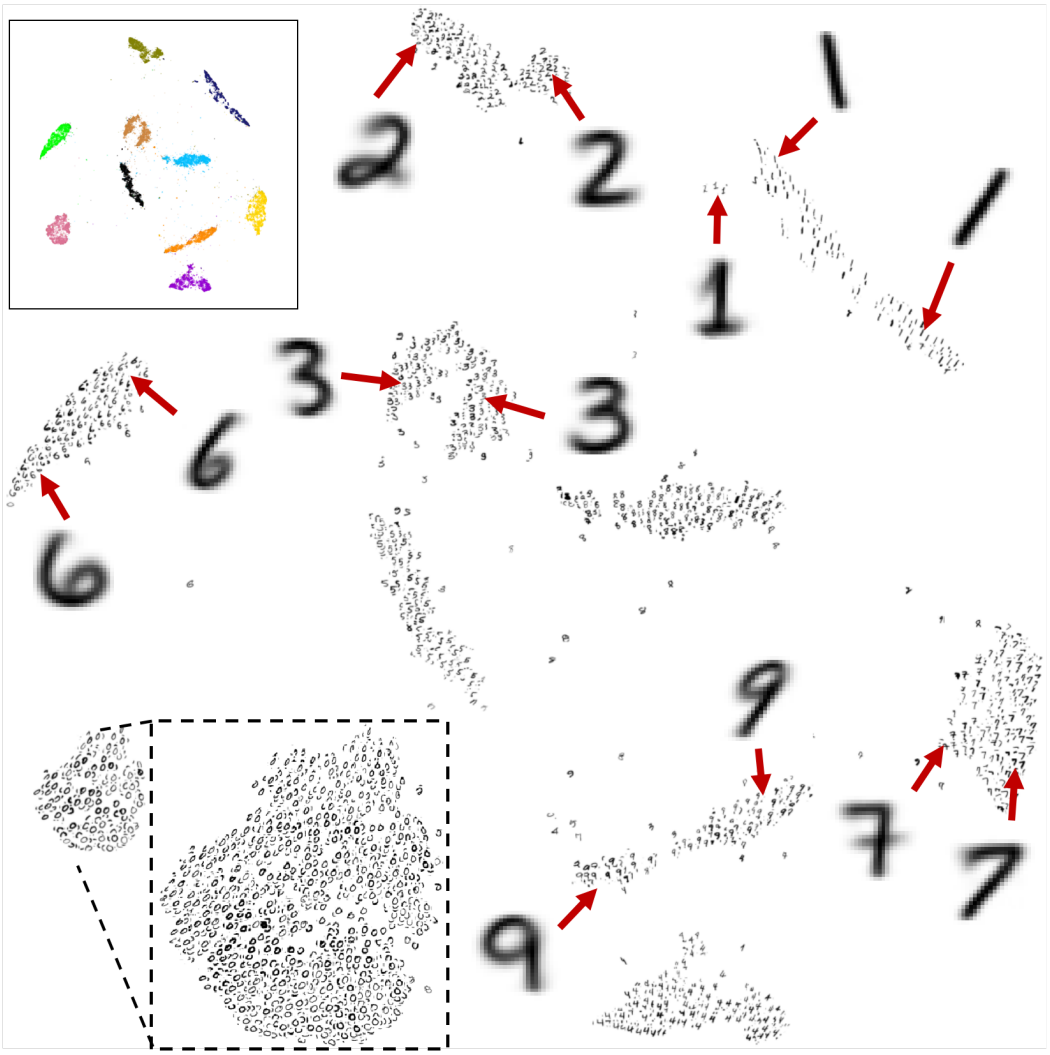

Figure 13: Visualizing MNIST images on SpaceMAP: the distribution of data points may not be uniform by nature. For example, a split-up of digit 2, 3 can be observed depending on the style of writing.

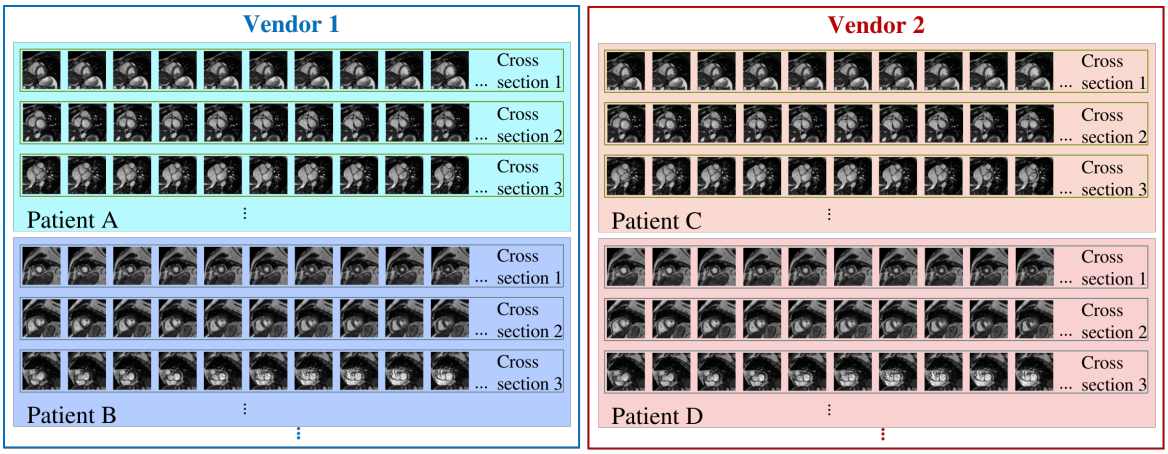

Figure 14: The original images of the cardiac cine MRI datasets as described in Section 4.1 and visualized in the last row of Figure. 4. The images depict human heart during one cardiac cycle at different cross section levels (short-axis). The worm-like structure in t-SNE and SpaceMAP visualization represents the spatial and temporal continuity in the image data.