MILITARY TECHNICAL COURIER

Data Visualization with kNN t-SNE

Shifali Fatima[1,*], Mohammad Aamir Qayyoom[1].

[1] Department of Mathematics & Statistics, Faculty of Science, Integral University, Lucknow, India.

Original scientific paper, Review paper
mathematics

Keywords:
Riemannian manifold;
t-SNE;
geodesic distance;
dimensionality reduction;
spherical embedding;

Abstract

Introduction/Purpose: High-dimensional datasets often lie on nonlinear manifolds, where standard techniques like t-SNE, is unable to recover the basic geometric structure. We propose a geodesic-based extension of the t-SNE algorithm, which replaces Euclidean distances with pairwise geodesic distances computed via shortest paths on a k-nearest neighbor graph. Using random initialization and precomputed geodesic metrics, the method keeps up relationships in the embedding. Experiments with manifold-structured datasets show that the quality of the visualizations is better because they show structures that regular t-SNE distorts. **Method:** A modification to the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm wherein the standard use of pairwise Euclidean distances is replaced by approximate geodesic distances computed over a k-nearest neighbor graph. **Result:** The result shows better performance of kNN-based t-SNE, compared to standard t-SNE. **Conclusion:** A Riemannian variant of t-SNE that leverages localized neighborhood structures via k-nearest neighbors and adapts gradient updates to respect the geometry of underlying manifolds. By relying on sparse affinity computation and geodesic-aware optimization, the method retains both fine-grained local relationships and larger-scale cluster separation, while significantly reducing memory and computational overhead compared to classical t-SNE. The algorithm is especially suited to high-dimensional, non-Euclidean data and demonstrates consistent performance across a range of geometrically structured datasets. Empirical results validate that our approach produces more stable and interpretable visualizations while preserving the manifold-aware relational topology.

1 Introduction

Classical dimensionality reduction methods such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) have become basic tools for embedding high-dimensional datasets into low-dimensional spaces.However, these methods are essentially based on the idea that data lies on or near a Euclidean manifold, where linear approximations and Euclidean metrics is sufficient to represent local neighborhood structure. The assumption fails on the non-euclidean manifold.

This paper presents a geodesic-informed extension to t-SNE which replaces conventional Euclidean distances with geodesic distances using a neighborhood graph. By modifying the input metric to reflect intrinsic geometry, the proposed method more faithfully captures structural features arising from the manifold's curvature. This allows the resulting low-dimensional embeddings to reflect manifold-aware proximity relationships, yielding visualizations that are sensitive to both local and global curvature-induced distortions inherent in the data.

2 Background and Related Work

Reducing high-dimensional data into meaningful low-dimensional representations is a core challenge in data science and machine learning [1]. Early techniques like Stochastic Neighbor Embedding (SNE) [2] introduced the idea of preserving local relationships between points by comparing probability distributions. This approach was refined in t-SNE [3], which improves upon SNE by minimizing the Kullback-Leibler divergence between pairwise similarity distributions in the original and reduced spaces. Thanks to its ability to retain fine-grained neighborhood structures, t-SNE has become one of the most popular methods for exploring and visualizing high-dimensional datasets [4], [5].

Although t-SNE has proven effective in many practical settings, it assumes a Euclidean framework. The similarities it relies on—both in the original data space and in the lower-

* Corresponding author
E-mail address: shifali.k01@gmail.com; (Shifali Fatima)
Received: 25 October 2025;
Received in revised form: 20 March 2026;
Accepted: 25 May 2026;
Available online: 06 November 2026;
https://doi.org/10.5937/vojtehg75-62394

ISSN 2217-4753 (Online)
ISSN 0042-8469 (Print issue)
© 2026 The Authors. Published by Military Technical Courier (www.vtg.mod.gov.rs). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

dimensional projection—are computed using Euclidean distances. This becomes a limitation when working with data that naturally lies on curved or structured manifolds [6]. For instance, points distributed on spheres, hyperbolic spaces, or other Riemannian manifolds often defy the flat-space assumptions built into t-SNE's similarity measures and optimization process [7], [8], [9]. As a result, the resulting visualizations can distort the underlying geometry, making it difficult to faithfully represent the true structure of the data.

Recent studies have tried to improve t-SNE by blending it with distance measures that respect the shape of curved spaces (manifolds) [10]. Some promising results have come from combining these geometry-aware ideas with stochastic methods [11], [12]. However, there's still no version of t-SNE that works smoothly on all types of Riemannian manifolds. The challenge lies in how differently each manifold behaves—the way distances curve and stretch (geodesics and curvature) can vary a lot [13]. Because of this, it's very hard to design one single algorithm that works well across all curved spaces while still being both accurate in geometry and efficient to compute.

In this work, we take a simple but effective approach to adapt t-SNE for curved data. Instead of using standard Euclidean distances, we estimate distances using a nearest-neighbor graph and shortest paths—similar to what Isomap does [14], [15]. This lets us better capture the true shape of the data in high dimensions, without changing how t-SNE works in low dimensions. Unlike other methods, we don't rely on complex tools like Riemannian gradient descent or retractions.

2.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a statistical method that helps visualize high dimensional data by reducing it to fewer dimensions, typically two or three. It aims to keep similar data points close together in the reduced space while keeping points far apart in the original data remain distant.

2.1.1 High-Dimensional Similarities

In t-SNE, the measure of similarity between two data points u_i and u_j in the space of high dimensionality is modeled using a conditional probability. The probability $\mathcal{P}_{ij}$, representing the likelihood of point j being a neighbor of point indexed i, is computed employing a Gaussian distribution with u_i as its mean:

$$\mathcal{P}_{j|i} = \frac{\exp\left(-\frac{\|u_i - u_j\|^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{\|u_i - u_k\|^2}{2\sigma_i^2}\right)},$$

where $\|u_i - u_j\|$ represents the Euclidean distance between u_i and u_j, and σ_i controls the variance of the Gaussian kernel for each point i. The value $\mathcal{P}_{j|i}$ is set to zero. To obtain a symmetric joint probability, the conditional probabilities are combined as:

$$\mathcal{P}_{ij} = \frac{\mathcal{P}_{j|i} + \mathcal{P}_{i|j}}{2n},$$

where n is the total number of data points. This symmetrization ensures that $\mathcal{P}_{ij} = \mathcal{P}_{ji}$.

2.1.2 Low-Dimensional Similarities

In the construction of the low-dimensional embedding, the affinities between projected points v_i and v_j are encoded via a probability measure $\mathcal{Q}_{ij}$ derived from a normalized heavy-tailed kernel. Concretely, Student's t-distribution with one degree of freedom is used, whose polynomial decay counters the crowding problem by allowing to represent pairwise dissimilarities. The induced probabilities are defined as

$$\mathcal{Q}_{ij} = \frac{\left(1 + \|v_i - v_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|v_k - v_l\|^2\right)^{-1}},$$

where the numerator prioritizes local proximity while the denominator ensures normalization over all unordered pairs. This particular form penalizes small pairwise separations among globally distant data points, thereby preserving the intrinsic disjointness of remote clusters in the embedding space. Such behavior is crucial for actual topological unfolding, particularly when the manifold structure exhibits nontrivial curvature or varying local densities.

2.1.3 The Cost Function

The objective of tSNE is to make the low-dimensional distribution Q resemble the high-dimensional distribution P as closely as possible. To this end, the algorithm minimizes the Kullback-Leibler divergence:

$$C = KL(\mathcal{P}\|\mathcal{Q}) = \sum_i \sum_j \mathcal{P}_{ij} \log \frac{\mathcal{P}_{ij}}{\mathcal{Q}_{ij}},$$

where $\mathcal{P}_{ij}$ and $\mathcal{Q}_{ij}$ are the similarity probabilities in the high- and low-dimensional feature spaces, respectively.

2.1.4 Gradient Descent Optimization

Minimization of the Kullback–Leibler divergence between the high-dimensional and embedded distributions is performed via gradient descent. The gradient of the objective function C with respect to the embedding coordinate $v_i \in \mathbb{R}^d$ is given by:

$$\frac{\partial C}{\partial v_i} = 4\sum_j \left(\mathcal{P}_{ij} - \mathcal{Q}_{ij}\right) \frac{v_i - v_j}{1 + \|v_i - v_j\|^2}.$$

To accelerate convergence and stabilize updates, a momentum-based scheme is adopted:

$$\Delta v_i^{(t)} = \eta \frac{\partial C}{\partial v_i^{(t)}} + \alpha \Delta v_i^{(t-1)}, \quad v_i^{(t+1)} = v_i^{(t)} - \Delta v_i^{(t)},$$

where $\eta > 0$ denotes the learning rate and $\alpha \in [0, 1)$ the momentum coefficient.

While this gradient promotes the colocation of similar points and repulsion of dissimilar ones, the cost function's pairwise structure yields quadratic complexity, and its performance is susceptible to the crowding phenomenon, in which distant points undesirably collapse in the low-dimensional representation.

Managing neighborhood fidelity in local areas, t-SNE brings in a concept of perplexity, which implicitly defines the number of effective neighbors using the Shannon entropy of the prospective distribution $\mathcal{P}_{j|i}$:

$$H(\mathcal{P}_{j|i}) = -\sum_j \mathcal{P}_{j|i} \log_2 \mathcal{P}_{j|i}, \quad \text{Perplexity}(\mathcal{P}_{j|i}) = 2^{H(\mathcal{P}_{j|i})}.$$

The bandwidth σ_i is tuned for each data point via a binary search on the entropy to match a user-specified perplexity, yielding:

$$p_{j|i} = \frac{\exp\left(-\frac{\|u_i - u_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|u_i - u_k\|^2}{2\sigma_i^2}\right)}.$$

However, such formulations presume an ambient Euclidean structure and isotropic locality—assumptions that fail in the presence of nontrivial curvature. When the data resides on a Riemannian manifold $(\mathcal{M}, g)$, the Euclidean metric no longer faithfully captures intrinsic distances, and local neighborhoods often exhibit anisotropic diffusion aligned with geodesic flows.

To deal with this, we replace Euclidean kernels with geodesic-informed similarity functions, and adapt bandwidths to respect directional curvatures. In particular, we consider anisotropic kernels aligned with intrinsic coordinates—e.g., exhibiting differential scaling along geodesics in the y-direction—thereby ensuring that local affinity structures remain compatible with the manifold geometry.

3 kNN based t-SNE

We propose a modification to the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm wherein the standard use of pairwise Euclidean distances is replaced by approximate geodesic distances computed over a k-nearest neighbor graph. Given a dataset $\mathcal{U} = \{u_i\}_{i=1}^n \subset \mathbb{R}^D$ sampled from a manifold $\mathcal{M} \subset \mathbb{R}^D$, we construct an undirected graph G by connecting each point to its k nearest neighbors in Euclidean norm, assigning edge weights $w_{ij} = \|u_i - u_j\|$. Approximate geodesic distances d_{ij} are then obtained via shortest paths in G. These distances replace $\|u_i - u_j\|$ in the high-dimensional conditional probability computations of t-SNE. The remaining steps, including the symmetrization of affinities and optimization of the Kullback–Leibler divergence in the embedding space, are preserved. This adjustment allows the algorithm to better capture the intrinsic geometry of non-Euclidean datasets while retaining the core structure of the original method.

3.1 Conceptual Overview

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_n\} \subset \mathcal{M}$ be a set of n samples embedded in a high-dimensional Riemannian manifold $\mathcal{M}$. The aim is to attain a low-dimensional embedded space $\mathcal{V} = \{v_1, v_2, \ldots, v_n\} \subset \mathbb{R}^2$ that preserves the pairwise structure of $\mathcal{U}$ as reflected through intrinsic distances.

The original t-SNE constructs similarity probabilities p_{ij} in the high-dimensional space via conditional Gaussians:

$$p_{j|i} = \frac{\exp\left(-\|u_i - u_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|u_i - u_k\|^2/2\sigma_i^2\right)}, \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

However, this formulation assumes linearity in the ambient space, disregarding curvature or topological obstructions.

To incorporate manifold structure, we replace the Euclidean norm $\|u_i - u_j\|$ with a geodesic approximation δ_{ij} computed as the shortest path length over a k-NN graph. The formation of the k-NN graph requires finding the k nearest neighbors for each data point u_i based on their pairwise Euclidean distances. Formally, the set of neighbors $\mathcal{N}(u_i)$ of a point u_i is given by:

$$\mathcal{N}(u_i) = \{u_j \in \mathcal{X} : \|u_i - u_j\| \leq \text{distance to the } k\text{-th nearest neighbor of } u_i\}.$$

Using the graph $\mathcal{G}$ formed from these k nearest neighbors, we compute the geodesic distance matrix $D = [\delta_{ij}]$ where:

$$\delta_{ij} = d_{\mathcal{G}}(u_i, u_j), \quad \text{with } \mathcal{G} \text{ being the k-NN graph of } \mathcal{X}.$$

This geodesic distance matrix is used to redefine the similarities between points, changing the Euclidean distance with the geodesic distance.

The similarities in the high-dimensional space are now given by:

$$p_{j|i} = \frac{\exp\left(-\delta_{ij}^2/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\delta_{ik}^2/2\sigma_i^2\right)}, \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

The low-dimensional embedding $\mathcal{V}$ is constructed by optimizing the KL divergence between the high- and low-

dimensional similarities:

$$C = \mathrm{KL}(\mathcal{P}\|\mathcal{Q}) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

where q_{ij} is defined using a Student$t$-t kernel:

$$q_{ij} = \frac{\left(1 + \|v_i - v_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|v_k - v_l\|^2\right)^{-1}}.$$

The use of geodesic distances aligns with manifold learning principles: if $\mathcal{M}$ is a smooth Riemannian manifold, then geodesic distances provide better estimators of intrinsic similarities than Euclidean metrics. This is particularly significant when the manifold has non-zero curvature or contains topological bottlenecks.

Standard t-SNE fails to respect such curvature, leading to visual artifacts like cluster tearing or poor local-global balance. By leveraging δ_{ij}, the Riemannian t-SNE embeds points according to their true connectivity along $\mathcal{M}$, preserving both shor$t$- and long-range structure more faithfully.

3.2 Pseudocode

The geodesic-based Riemannian t-SNE thus is a principled modification of the classical algorithm that directly addresses its limitations in non-Euclidean domains. By constructing embeddings over geodesic metrics derived from local connectivity graphs, the proposed method better respects the topological and metric structure of curved data manifolds, providing a robust tool for high-dimensional data visualization and analysis.

4 Experiment

To evaluate kNN based t-SNE,it is experimented and we compare it with standard t-SNE. We conducted experiments on five different datasets spanning multiple application domains. However, due to space constraints, we present results from only three of these datasets in this paper.

In Section 4.1, the data sets used in our experiments are defined. The setup of the experiments is in Section 4.2. In Section 4.3, we highlight the results of our experiments.

4.1 Datasets

The data sets used in experiments are: (1) the MNIST data set, (2) COIL 100 dataset and (3) the Olivetti dataset. We evaluate our approach on three diverse image datasets. The The MNIST dataset comprises 60,000 grayscale images, each depicting a handwritten digit, each of size 28×28 pixels, from which we randomly sample 6,000 for analysis. The Olivetti Faces dataset includes 400 grayscale facial images of 40 individuals (10 per person), sized 92×112 pixels, featuring variations in expression, pose, and accessories. The COIL-100 dataset comprises 7,200 images of 100 objects captured at 5° rotational increments, with each 32×32 image reflecting smooth viewpoint changes, making it suitable for modeling continuous transformations.

4.2 Experimental Setup

For our experiments, we begin by normalizing the data and applying PCA to reduce the original dimensionality to 50, preserving most of the variance while improving computational efficiency. We then construct a k-NN graph (set $k = 15$) in this reduced space and compute geodesic distances using shortest paths over the resulting graph. These geodesic distances are used in place of Euclidean distances within a modified t-SNE framework to better respect the underlying manifold structure. For comparison, we also generate embeddings using standard t-SNE on the same PCA-reduced data. In both cases, the final low-dimensional embeddings are visualized in two dimensions, with class labels used solely for coloring the points to aid in visual interpretation—no label information is used in the computation itself.

4.3 Results

Figures 1, 2, and 3 show the result of the t-SNE algorithm. The result shows better performance of kNN-based t-SNE, compared to standard t-SNE.

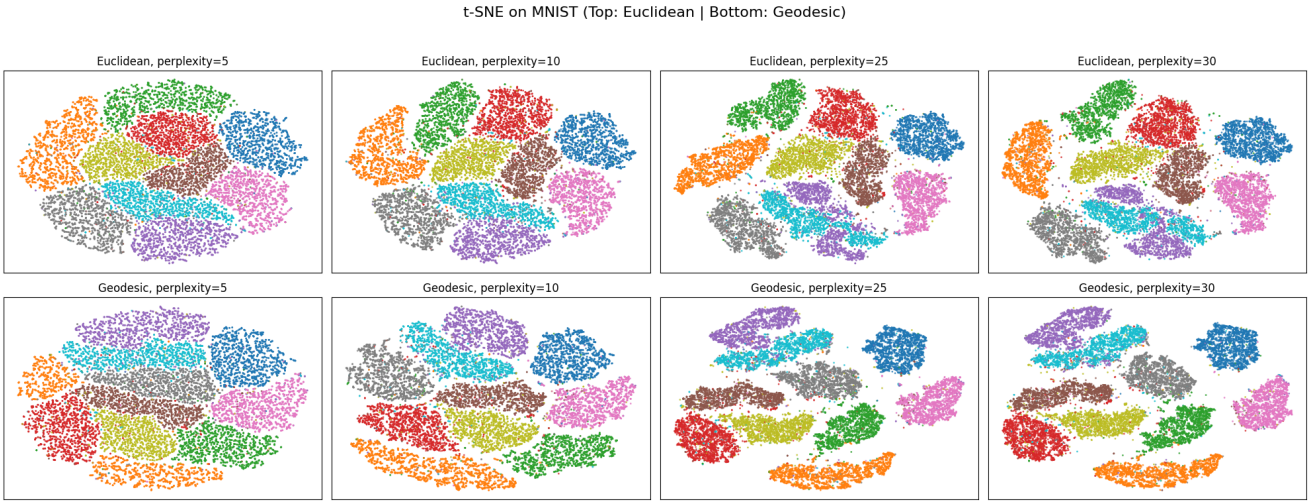


Figure 1. Visualisation of MNIST.

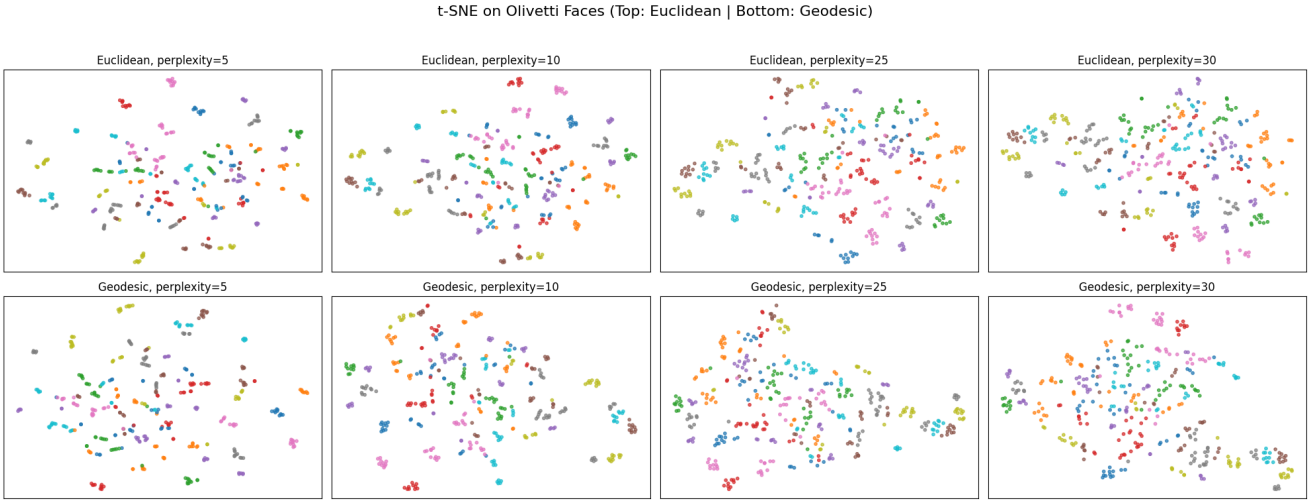


Figure 2. Visualisation of Olivetti Faces.

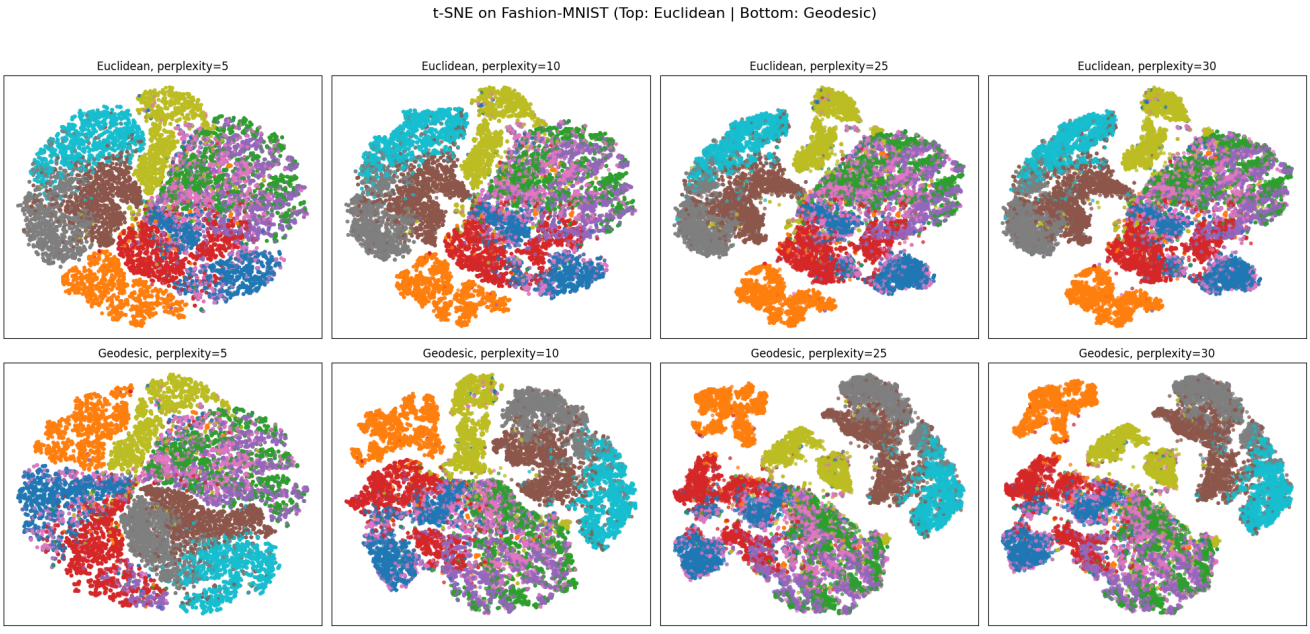


Figure 3. Visualisation of Fashion MNIST.

To assess the quality of the embeddings, we report three standard evaluation metrics: *Trustworthiness*, *KL Divergence*, and *Silhouette Score*. The table below compares these metrics for the proposed kNN-based t-SNE variant with the classical t-SNE across multiple datasets.

Table 1. Comparison of t-SNE and kNN-based t-SNE across datasets.

Dataset	Method	Trustworthiness	Silhouette Score	KL Divergence
MNIST	Standard t-SNE	0.9926	0.3495	1.7643
	kNN-based t-SNE	0.9933	0.3426	2.1211
Fashion-MNIST	Standard t-SNE	0.9930	0.1611	1.5964
	kNN-based t-SNE	0.9928	0.1664	1.8638
Olivetti Faces	Standard t-SNE	0.9771	0.0483	0.5917
	kNN-based t-SNE	0.9799	0.0870	0.9086

On MNIST and Fashion-MNIST, our method achieved comparable or slightly better trustworthiness and silhouette scores, indicating that it preserves local structures effectively. Notably, on the Olivetti Faces dataset — a more challenging case due to limited samples per class and high intra-class variance — the kNN-based t-SNE improved the silhouette score substantially (from 0.0483 to 0.0870), suggesting enhanced class separation. While the KL divergence increased across all datasets, this is a known consequence of using geodesic distances, which deviate from the assumptions underlying standard t-SNE. Nonetheless, the preservation of local structure and improved separability in challenging scenarios support the potential benefits of incorporating manifold-aware distances in t-SNE.

5 Weakness

While the proposed kNN-based t-SNE formulation offers clear benefits in preserving neighborhood structures, several mathematical limitations remain. Since the algorithm builds upon discrete nearest-neighbour graphs, its effectiveness hinges on the accuracy of the local connectivity structure, which can be difficult to estimate when the manifold is irregular, highly curved, or unevenly sampled. In such cases, the graph construction may fail to reflect the underlying geometric relationships faithfully, leading to distortions in the embedding. Moreover, the method inherits the non-convex optimization landscape of classical t-SNE, making the final result dependent on initialization and hyperparameter choices, including the neighborhood size k. Although these parameters provide flexibility, they also introduce variability across runs and datasets. From a manifold learning perspective, the algorithm primarily captures local geometry, but its ability to represent broader geometric or topological features—such as continuity across disconnected components or global curvature—is more limited. Extensions incorporating continuous kernel-based similarities or graph Laplacians adapted to manifold structure may help address these concerns in future work.

6 Conclusion

This work introduces a Riemannian variant of t-SNE that leverages localized neighborhood structures via k-nearest neighbors and adapts gradient updates to respect the geometry of underlying manifolds. By relying on sparse affinity computation and geodesic-aware optimization, the method retains both fine-grained local relationships and larger-scale cluster separation, while significantly reducing memory and computational overhead compared to classical t-SNE. The algorithm is especially suited to high-dimensional, non-Euclidean data and demonstrates consistent performance across a range of geometrically structured datasets. Empirical results validate that our approach produces more stable and interpretable visualizations while preserving the manifold-aware relational topology. Future directions include exploring more expressive metrics on manifolds, extending the approach to parametric or out-of-sample settings, and scaling to very large datasets using approximate neighbors and hierarchical embeddings.

This work lays a geometric foundation for manifold-aware visualization, yet it opens more questions than it resolves — a hallmark of a fertile mathematical framework. Several deep and natural directions for future inquiry remain. First, a rigorous theoretical characterization of the algorithm's behavior under varying curvature regimes is essential to understand its convergence, stability, and sensitivity to underlying geometry. Second, extending the current framework to encompass more general spaces — such as manifolds with boundary, stratified structures, or singular geometries — would significantly broaden its applicability to real-world data that deviates from ideal smooth settings. Third, most current evaluations of embedding quality rely on Euclidean benchmarks; developing intrinsic metrics and principled evaluation criteria tailored for non-Euclidean targets remains an open and nontrivial challenge. Moreover, the interplay between sampling density, neighborhood construction, and geodesic approximation demands further exploration, especially in cases where manifold learning is compounded by data sparsity or noise.

Ultimately, this work should be viewed not as a culmination but as a blueprint — an early step toward a more principled integration of geometry and data representation. Advancing this line of research will require both sharper mathematical tools and empirical validation across diverse domains, with the long-term aim of developing algorithms that are not only powerful in practice but also interpretable and geometrically faithful by design.

References

[1] Ma Y., Fu Y., 2012. Manifold learning theory and applications, vol. 434, CRC press, Boca Raton, available at: https://doi.org/10.1201/b11431.

[2] Hinton G. E., Roweis S., 2002. Stochastic neighbor embedding, Advances in neural information processing systems, 15. Available at: https://shorturl.at/rFaxF.

[3] Van der Maaten L., Hinton G., 2008. Visualizing data using t-sne., Journal of machine learning research, 9(11). Available at: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[4] Wattenberg M., Viégas F., Johnson I., 2016. How to use t-sne effectively, Distill, 1(10), e2. Available at: http://doi.org/10.23915/distill.00002.

[5] Arora S., Hu W., Kothari P. K., 2018. An analysis of the t-sne algorithm for data visualization, in: Conference on learning theory, PMLR, 1455–1462. Avaiable at: https://proceedings.mlr.press/v75/arora18a.html.

[6] Li Y., 2018. Curvature-aware manifold learning, Pattern Recognition, 83, 273–286. Available at: https://doi.org/10.1016/j.patcog.2018.06.007.

[7] Le T. M., Lauw H. W., 2014. Semantic visualization for spherical representation, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1007–1016. Available at: 10.1145/2623330.

[8] Wang M., Wang D., 2016. Vmf-sne: embedding for spherical data, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2344–2348. Available at: https://doi.org/10.1109/ICASSP.2016.7472096.

[9] Guo Y., Guo H., Yu S. X., 2022. Co-sne: Dimensionality reduction and visualization for hyperbolic data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 21–30. Available at: https://shorturl.at/fy79c.

[10] de Silva V., Tenenbaum J. B., 2003. Unsupervised learning of curved manifolds, in: Nonlinear estimation and classification, Springer, 453–465. Available at: https://doi.org/10.1007/978-0-387-21579-2_31.

[11] Bergsson A., Hauberg S., 2022. Visualizing riemannian data with rie-sne, arXiv preprint arXiv:2203.09253. Available at: https://doi.org/10.48550/arXiv.2203.09253.

[12] Archambault D.and Nabney I., Peltonen J., 2024. Visualizing riemannian data with rie-sne, arXiv preprint arXiv:2203.09253. Available at: https://doi.org/10.48550/arXiv.2203.09253.

[13] Ghojogh B., Crowley M., Karray F., Ghodsi A., 2018. Visualizing riemannian data with rie-sne, arXiv preprint. Available at: https://doi.org/10.48550/arXiv.2110.09620.

[14] Tenenbaum J. B., Silva V. d., Langford J. C., 2000. A global geometric framework for nonlinear dimensionality reduction, science, 290(5500), 2319–2323. Available at: https://doi.org/10.1126/science.290.5500.2319.

[15] Cayton L., 2008. Algorithms for manifold learning, eScholarship, University of California. Available at: http://cseweb.ucsd.edu/~lcayton/resexam.pdf.

Military Technical Courier, Vol. 75, Issue 1 (2027)

Vizuelizacija podataka pomoću kNN t-SNE

Shifali Fatima[1,*], Mohammad Aamir Qayyoom[1].

[1] Department of Mathematics & Statistics, Faculty of Science, Integral University, Lucknow, India.

Abstarkt

Visokodimnzionalni skupovi podataka često leže na nelinearnim mnogostrukostima, gde standardne tehnike poput t-SNE nisu u stawu da regenerišu osnovnu geometrijsku strukturu.Predlažemo geodetsko proširenje t-SNE algoritma, koje zamenjuje euklidske rastojanja parnim geodetskim rastojanjima izračunatim preko najkraćih puteva na grafu k najbližih suseda.Koristeći slučajnu inisijallizaciju i unapred izračunate geodetske metrike, metod održava odnose u ugrađivanju. Eksperimenti sa skupovima podataka sa višestruko strukturiranom strukturom pokazuju da je kvalitet vizuelizacija bolji jer prikazuje strukture koje regularna t-SNE iskrivljuje. **Metod:** Modifikacija t-distribuiranog stohastičkog uređivanja suseda (t-SNE) algoritma gde se standardna upotreba parnih euklidskih rastojanja zamenjuje približnim geodetskim rastojanjima izračunatim preko k-grafa najbližih suseda. **Rezultat:** Rezultat pokazuje bolje performanse t-SNE zasnovanog na kNN-u, u poređenju sa standardnim t-SNE. **Zaključak:** Rimanova varijanta t-SNE koja koristi lokalizovane strukture susedstva putem k-najbližih suseda i prilagođava ažuriranja gradijenta kako bi se poštovala geometrija osnovnih mnogostrukosti. Oslanjajući se na izračunavanje retke afinitetnosti i geodetsku optimizaciju, metod zadržava i fino zrnaste lokalne odnose i razdvajanje klastera većih razmera, uz značajno smanjenje memorije i računarskoh opterećenja u poređenju sa klasičnim t-SNE. Algoritam je posebno pogodan za visokodimenzione, neeuklidske podatke i pokazuje konzistentne perormanse u nizu geometrijski struktuiranih skupova podataka. Empirijski rezultati potvrđuju da naš pristup proizvodi stabilnije i interoperatibilnije vizuelizacije, uz očuvanje racionalne topologije koja je svesna mnogostrukosti.

Originalni naučni rad, Pregledni rad
matematika

Ključne reči:
Rimanova varijanta;
t-SNE;
geodetska rastojanja;
dimensionality reduction;
spherical embedding;