



Article

Novel Loss Functions for Improved Data Visualization in *t*-SNESara Nassar ^{1,*} , Rachid Hedjam ² and Samir Brahim Belhaouari ¹ ¹ College of Science and Engineering, Hamad Bin Khalifa University, Education City, Gate 8, Doha P.O. Box 5825, Qatar; sbelhaouari@hbku.edu.qa² Department of Computer Science, Bishop's University, 2600 College Street, Sherbrooke, QC J1M 1Z7, Canada; rhedjam@ubishops.ca

* Correspondence: sana09516@hbku.edu.qa

Abstract

A popular method for projecting high-dimensional data onto a lower-dimensional space while preserving the integrity of its structure is t-distributed Stochastic Neighbor Embedding (*t*-SNE). This technique minimizes the Kullback–Leibler (*KL*) divergence to align the similarities between points in the original and reduced spaces. While *t*-SNE is highly effective, it prioritizes local neighborhood preservation, which results in limited separation between distant clusters and inadequate representation of global relationships. To improve these limitations, this work introduces two complementary approaches: (1) The Max-Flipped *KL* Divergence ($KL^{\max}$) modifies the original divergence by incorporating a contrastive term, KL' , which enhances the ranking of point similarities through maximum similarity constraints. (2) The *KL*-Wasserstein Loss ($\mathcal{L}_{KL-W}$) combines the *KL* divergence with the classic Wasserstein distance, allowing the embedding to benefit from the smooth and geometry-aware transport properties of Wasserstein metrics. Experimental results show that these methods lead to improved separation and better structural clarity in the low-dimensional space compared to standard *t*-SNE.

Keywords: *t*-SNE; visualization; Kullback–leibler divergence; Wasserstein distance; loss functions

1. Introduction

A core challenge in exploratory data analysis is producing two-dimensional views that are both locally faithful and globally interpretable. Popular methods include Principal Component Analysis (PCA) [1], Uniform Manifold Approximation and Projection (UMAP) [2], and t-distributed Stochastic Neighbor Embedding (*t*-SNE) [3]. Among these, *t*-SNE is widely adopted for preserving local neighborhoods, which makes clusters visually salient. However, optimizing the Kullback–Leibler (*KL*) divergence in the standard objective has known side effects: it tends to under-represent global relationships (inter-cluster layout, long-range structure), is sensitive to a few large P_{ij} values, and its asymmetry makes it more punitive for missing close neighbors than for misplacing distant ones. These behaviors can limit interpretability when cluster arrangement and global trends matter.

Understanding the tension between local and global fidelity is critical in many practical applications. For example, in single-cell biology, similar cell types may form distinct clusters that should still reflect developmental trajectories; meanwhile, in image retrieval and language modeling, globally meaningful structure supports semantic grouping and hierarchy discovery. Existing methods have sought to balance this trade-off: *f*-divergence variants of *t*-SNE re-weight neighborhood probabilities to adjust local sensitivity, while



Academic Editor: Massimo Ferri

Received: 5 October 2025

Revised: 19 November 2025

Accepted: 22 November 2025

Published: 18 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

global structure-aware algorithms such as UMAP introduce additional terms that encourage distant clusters to retain relative positioning. Despite these efforts, there remains no simple formulation that simultaneously enforces local ranking consistency and global geometric alignment within the original t -SNE optimization framework.

To address this gap, this work introduces two enhanced formulations that extend the t -SNE framework: (1) Max-Flipped KL Divergence ($KL^{\max}$) and (2) KL -Wasserstein Loss ($\mathcal{L}_{KL-W}$). Each method is designed to mitigate specific shortcomings of the original t -SNE formulation while improving the separation and interpretability of resulting visualizations.

First, Max-Flipped KL Divergence ($KL^{\max}$) augments the traditional KL divergence with an additional term, KL' , that focuses on discrepancies in similarity rankings. By explicitly penalizing violations in expected similarity orderings, this extension improves the model's sensitivity to subtle structural variations and enhances separation between distinct groups.

Second, KL -Wasserstein Loss ($\mathcal{L}_{KL-W}$) introduces a hybrid objective that combines KL divergence with the classic Wasserstein distance. This integration leverages the Wasserstein metric's ability to capture smooth, geometry-aware transport between distributions, thus reinforcing the separation between clusters while preserving meaningful local arrangements.

In contrast to prior t -SNE extensions based on alternative f -divergences such as Jensen-Shannon, reverse KL , and other symmetric formulations [4–8], as well as OT-regularized or transport-based embedding methods [9–14], the proposed $KL^{\max}$ and $\mathcal{L}_{KL-W}$ differ in two fundamental aspects. $KL^{\max}$ introduces a similarity-ranking inversion mechanism that directly penalizes deviations from the local maximum probability, a behavior not captured by symmetric or alternative f -divergence objectives [7,8]. $\mathcal{L}_{KL-W}$ incorporates a direct Wasserstein W_1 alignment term in the embedding space, unlike sliced-OT or entropy-regularized OT variants that operate in the data space or rely on smoothed surrogate distances [9,11,12]. This distinction allows our formulation to jointly address local ranking fidelity and global geometric consistency.

Recent work in representation learning has shown that preserving semantic relationships in complex data is essential for obtaining globally coherent embeddings. For example, robust feature representation methods for human parsing in complex scenes [15] emphasize maintaining consistent global structure, which is conceptually aligned with our objective of enhancing global structure consistency through the Wasserstein term. In addition, robustness to geometric transformations has been studied in other domains, such as light-field image watermarking. Techniques designed to withstand multidimensional geometric attacks [16] emphasize the importance of preserving structural consistency under complex transformations. Although the application domain differs, this broader theme is conceptually aligned with our motivation for incorporating the Wasserstein distance to strengthen the geometric structure preservation capability of the embedding.

Together, these considerations motivate the need for embedding objectives that more faithfully preserve both local and global relationships—a goal that the proposed $KL^{\max}$ and $\mathcal{L}_{KL-W}$ formulations are specifically designed to address.

2. t -Distributed Stochastic Neighbour Embedding (t -SNE)

t -SNE is a widely adopted method for visualizing high-dimensional data. It relies on KL divergence to align probability distributions in the original high-dimensional space with those in the low-dimensional embedding. The primary goal of t -SNE is to minimize the divergence between two distributions:

- (1) The high-dimensional probability distribution P_{ij} , which models the similarity between points x_i and x_j using a Gaussian kernel:

$$P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \quad (1)$$

where σ_i is a local bandwidth parameter. The final symmetric joint probability is given by:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n} \quad (2)$$

- (2) The low-dimensional probability distribution Q_{ij} , modeled with a Student's t -distribution to account for distant similarities:

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \quad (3)$$

The optimization objective in t -SNE is to minimize the KL divergence between P_{ij} and Q_{ij} :

$$KL(P\|Q) = \sum_i \sum_{j \neq i} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \quad (4)$$

This ensures that the low-dimensional representation preserves the pairwise similarities of the original space as accurately as possible. The gradient of the KL divergence with respect to the low-dimensional embeddings is:

$$\frac{\partial}{\partial y_i} KL(P\|Q) = 4 \sum_{j \neq i} (P_{ij} - Q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \quad (5)$$

This gradient is used to iteratively update the embeddings via gradient descent, ensuring the local structure is well preserved.

Although t -SNE is highly effective for clustering and visualizing high-dimensional data, its dependence on the KL divergence introduces several well-documented drawbacks. First, because the KL objective primarily focuses on minimizing local reconstruction errors, it tends to emphasize neighborhood-level relationships while neglecting large-scale structures such as hierarchies, inter-cluster layouts, and global trends [17–19]. As a result, clusters that are meaningfully related in the original space may appear arbitrarily separated or distorted in two-dimensional projections, reducing interpretability in tasks that depend on global context. Second, the asymmetric nature of KL divergence makes it more sensitive to underestimations in Q_{ij} than to overestimations, causing t -SNE to penalize missing close neighbors far more severely than misplacing distant ones. This imbalance strengthens local compactness but further weakens the preservation of global geometry. Addressing these limitations motivates our proposed extensions, which explicitly target both local ranking consistency and global geometric alignment.

3. Wasserstein Distance

The p -Wasserstein distance ($p \in [1, \infty)$) quantifies the minimum “effort” required to transport probability mass from one distribution to another over a metric space (Ω, d) . It unifies continuous (density-based) and discrete (empirical) settings under the same optimal transport (OT) framework [20].

Given probability measures P and Q on (Ω, d) , the p -Wasserstein distance is

$$W_p(P, Q) = \left(\inf_{\gamma \in \Gamma(P, Q)} \int_{\Omega \times \Omega} d(x, y)^p d\gamma(x, y) \right)^{1/p}, \quad (6)$$

where $\Gamma(P, Q)$ is the set of couplings (transport plans) with the correct marginals:

$$\gamma(A \times \Omega) = P(A), \quad \gamma(\Omega \times B) = Q(B) \quad \text{for all Borel } A, B \subseteq \Omega. \quad (7)$$

When P is absolutely continuous with respect to the Lebesgue measure, the Kantorovich problem admits a map solution in many cases. The Monge formulation searches for a measurable transport map $f : \Omega \rightarrow \Omega$ pushing P to Q (denoted $f_{\#}P = Q$; i.e., $Q(A) = P(f^{-1}(A))$ for all Borel A):

$$W_p(P, Q) = \left(\inf_{f: f_{\#}P=Q} \int_{\Omega} d(x, f(x))^p dP(x) \right)^{1/p}. \quad (8)$$

For $p = 2$ on $\Omega \subseteq \mathbb{R}^d$, [21] shows that the optimal map exists and is the gradient of a convex potential (under mild conditions), linking (8) and (6).

In one dimension, the optimal map is monotone. Let F_P and F_Q be the CDFs of P and Q , and F_P^{-1}, F_Q^{-1} their quantile functions. Then

$$W_p(P, Q) = \left(\int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)|^p du \right)^{1/p}. \quad (9)$$

Equivalently, $f(x) = F_Q^{-1}(F_P(x))$ yields the Monge map and recovers (9) via (8).

For discrete/empirical measures, let $P = \sum_{i=1}^n a_i \delta_{x_i}$ and $Q = \sum_{j=1}^m b_j \delta_{y_j}$ with $a_i, b_j \geq 0$, $\sum_i a_i = \sum_j b_j = 1$. Writing $C_{ij} = d(x_i, y_j)^p$, the OT problem becomes a finite linear program:

$$W_p(P, Q)^p = \min_{\Pi \in \mathbb{R}_+^{n \times m}} \sum_{i=1}^n \sum_{j=1}^m \Pi_{ij} C_{ij} \quad \text{s.t.} \quad \sum_j \Pi_{ij} = a_i, \quad \sum_i \Pi_{ij} = b_j. \quad (10)$$

This discrete formulation is the Kantorovich linear program [22]. When $n = m$ and $a_i = b_j = \frac{1}{n}$, (10) searches over doubly-stochastic Π . In one dimension with uniform weights, sorting the supports gives a closed form:

$$W_p(P, Q)^p = \frac{1}{n} \sum_{i=1}^n |x_{(i)} - y_{(i)}|^p, \quad (11)$$

where $x_{(1)} \leq \dots \leq x_{(n)}$ and $y_{(1)} \leq \dots \leq y_{(n)}$.

For general (n, m) , (10) has nm variables and is typically solved by specialized LP/OT solvers. Exact network-simplex methods scale superlinearly (often near-cubic in n in practice), while entropic regularization leads to fast Sinkhorn matrix scaling with $\mathcal{O}(nm)$ work per iteration. In one dimension, computing W_p reduces to sorting (cost $\mathcal{O}(n \log n)$) and then linear time; if samples are pre-sorted, the evaluation is $\mathcal{O}(n)$.

Equations (6)–(10) provide a unified framework for both continuous measures (via maps or plans) and discrete empirical distributions (via finite couplings), while the one-dimensional case (9)–(11) admits closed-form solutions [23].

4. Proposed Methods

This section introduces two methods that directly target the well-known limitations of t -SNE. Standard t -SNE's asymmetric $KL(P \parallel Q)$ overemphasizes missed neighbors, tolerates false ones, and thus causes crowding and unreliable global geometry. We address this with $KL^{\max}$ to enforce bidirectional fidelity and reduce false neighbors, and with $\mathcal{L}_{\text{KL-W}} = KL + \lambda W$ to preserve global structure via a transport-aware term. Together, these objectives lower MSE, stabilize embeddings, and yield clearer, more interpretable separations with trustworthy inter-cluster distances for downstream analysis.

4.1. Max-Flipped KL Divergence

The proposed KL^{max} builds on the original loss by introducing an additional term, KL' , which focuses on differences in similarity rankings. By capturing deviations from the highest similarity values in both the high-dimensional and low-dimensional spaces, KL' ensures that not just the raw probabilities, but also their relative importance, are preserved. This enhancement enables KL^{max} to provide a more balanced embedding, effectively representing relationships between pairs of neighbors rather than focusing on a single dominant relation.

The KL^{max} divergence is defined as

$$KL^{max}(P\|Q) = KL(P\|Q) + \alpha KL'(P\|Q), \quad (12)$$

where $KL(P\|Q)$ is provided in Equation (4), and

$$KL'(P\|Q) = \sum_i \sum_{j \neq i} P'_{ij} \log \frac{P'_{ij}}{Q'_{ij}}. \quad (13)$$

Here,

$$P'_{ij} = \max(P_{i\cdot}) - P_{ij}, \quad (14)$$

$$Q'_{ij} = \max(Q_{i\cdot}) - Q_{ij}, \quad (15)$$

and both P'_{ij} and Q'_{ij} are subsequently normalized so that $\sum_{j \neq i} P'_{ij} = 1$ and $\sum_{j \neq i} Q'_{ij} = 1$ for each i .

This transformation effectively “flips” the focus of the divergence from comparing probabilities directly to measuring their deviations from the maximum probabilities within their respective distributions. Instead of emphasizing the values of P_{ij} or Q_{ij} themselves, KL^{max} prioritizes the distances of P_{ij} and Q_{ij} from their most probable counterparts, $\max(P_{i\cdot})$ and $\max(Q_{i\cdot})$.

The gradient of KL^{max} with respect to the low-dimensional embeddings y_i is the sum of the gradients of the two components:

$$\frac{\partial}{\partial y_i} KL^{max}(P\|Q) = \frac{\partial}{\partial y_i} KL(P\|Q) + \alpha \frac{\partial}{\partial y_i} KL'(P\|Q). \quad (16)$$

The gradient of $KL(P\|Q)$ is given in Equation (5). Using the Student- t kernel $q_{ij} \propto (1 + \|y_i - y_j\|^2)^{-1}$ and the standard t -SNE force decomposition, the contribution of $KL'(P\|Q)$ takes the same force form with (P', Q') :

$$\frac{\partial}{\partial y_i} KL'(P\|Q) = 4 \sum_{j \neq i} (P'_{ij} - Q'_{ij}) \frac{(y_i - y_j)}{1 + \|y_i - y_j\|^2}. \quad (17)$$

Combining (16) with (17) and the standard KL term yields the final expression:

$$\frac{\partial}{\partial y_i} KL^{max}(P\|Q) = 4 \sum_{j \neq i} \left((P_{ij} - Q_{ij}) + \alpha (P'_{ij} - Q'_{ij}) \right) \frac{(y_i - y_j)}{1 + \|y_i - y_j\|^2}. \quad (18)$$

To avoid undefined terms when $Q'_{ij} = 0$ with $P'_{ij} > 0$, we (i) apply the same ε -floor to both P' and Q' , and (ii) compute KL' on the union of active indices per row (equivalently, after flooring-and-renormalization all entries are strictly positive). With these two steps, KL' is well-defined.

4.2. KL–Wasserstein Loss

The proposed KL–Wasserstein objective augments the standard t -SNE loss with a transport-based regularization term. Formally, the combined loss is

$$\mathcal{L}_{KL-W} = KL(P\|Q) + \beta W_p(P, Q), \quad (19)$$

where $KL(P\|Q)$ is the conventional similarity-based divergence (Equation (4)). The Wasserstein component $W_p(P, Q)$ captures global geometric discrepancies between the original data $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ and the embedded points $Y = \{y_i\}_{i=1}^n \subset \mathbb{R}^2$. In our formulation, we evaluate the first Wasserstein distance W_1 between the empirical measures $\mu_X = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\mu_Y = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$, computed using the `wasserstein_distance_nd` routine from `scipy.stats`. This computation is applied on mini-batches and incurs only moderate cost.

Because `wasserstein_distance_nd` provides only the scalar distance, we estimate its gradient with respect to the embedding Y using a stochastic finite-difference approximation inspired by SPSA. Small random perturbations are applied to Y , and the resulting changes in W_1 are used to approximate $\partial W_1 / \partial Y$.

For reference, the analytic subgradient associated with an explicit optimal assignment σ is

$$\frac{\partial}{\partial y_i} W_1(\mu_X, \mu_Y) = \frac{1}{n} \frac{y_i - x_{\sigma(i)}}{\|y_i - x_{\sigma(i)}\|_2}, \quad (20)$$

or in a smoothed form,

$$\frac{y_i - x_{\sigma(i)}}{\sqrt{\|y_i - x_{\sigma(i)}\|_2^2 + \delta^2}}, \quad \delta \in [10^{-6}, 10^{-3}], \quad (21)$$

which avoids division by zero while converging to the true subgradient as $\delta \rightarrow 0$. In practice, because recomputing σ at each iteration is computationally expensive, our implementation relies on the finite-difference estimator instead of the analytic form.

The KL and Wasserstein terms contribute complementary geometric information to the embedding. The divergence $KL(P\|Q)$ induces short-range attractive forces that preserve local neighborhood relations, becoming increasingly influential as clusters form. In contrast, the Wasserstein component provides global structural guidance by encouraging each embedded point y_i to align with its high-dimensional counterpart $x_{\sigma(i)}$.

To illustrate their interaction, consider the gradient of the combined objective,

$$\nabla_{y_i} \mathcal{L}_{KL-W} = \nabla_{y_i} KL(P \| Q) + \beta \nabla_{y_i} W_1(\mu_X, \mu_Y). \quad (22)$$

The KL gradient takes the well-known t -SNE form,

$$\nabla_{y_i} KL(P \| Q) = 4 \sum_{j \neq i} (P_{ij} - Q_{ij}) \frac{(y_i - y_j)}{1 + \|y_i - y_j\|^2}, \quad (23)$$

which strengthens local attraction whenever $P_{ij} > Q_{ij}$.

In contrast, away from assignment changes, a valid subgradient of W_1 is

$$\nabla_{y_i} W_1(\mu_X, \mu_Y) = \frac{1}{n} \frac{y_i - x_{\sigma(i)}}{\|y_i - x_{\sigma(i)}\|_2}, \quad (24)$$

which pulls each embedded point toward its high-dimensional origin. This long-range signal counteracts crowding and improves the global layout.

Although the two forces may point in different directions at times, their effects are complementary. The Wasserstein component typically dominates in the early iterations, promoting a globally coherent structure, while the KL term refines local neighborhoods as optimization progresses. Consequently, the combined objective achieves both improved global organization and enhanced local fidelity compared with standard t -SNE.

5. Experiments

This section evaluates the performance of the proposed methods, $KL^{\max}$ and $\mathcal{L}_{KL-W}$, across a variety of datasets. We organize the experiments as follows: (i) we first evaluate the impact of the weighting parameters α and β on the quality of the embeddings; (ii) we then examine the evolution of the loss function across optimization iterations; (iii) next, we compare the clustering and structure preservation performance of $KL^{\max}$ and $\mathcal{L}_{KL-W}$ against the baseline using quantitative metrics; and (iv) finally, we provide qualitative assessments through 2D visualizations of the resulting embeddings. All experiments were conducted on a MacBook Pro (13-inch, M1, 2020) running macOS Sequoia 15.1 with an Apple M1 chip and 8 GB of RAM. These hardware specifications are provided to contextualize the computational performance of the proposed methods.

5.1. Datasets

We conducted experiments on a diverse collection of benchmark datasets, including Pendigits [24], MNIST [25], Fashion-MNIST [26], COIL-20 [27], and Olivetti Faces [28]. These datasets vary in terms of structure, modality, and complexity, allowing us to evaluate the effectiveness of our proposed methods under different distributional settings.

- MNIST contains 70,000 handwritten digit images (28×28 pixels, grayscale), evenly distributed across 10 classes. It is widely used for evaluating clustering quality in embedding spaces.
- Pendigits comprises 1797 instances of pen-based digit trajectories (8×8 grayscale images) representing digits 0 to 9, each assigned to one of 10 classes.
- Fashion-MNIST includes 70,000 grayscale images (28×28 pixels) of clothing items from 10 fashion categories. Compared to MNIST, this dataset presents more subtle visual variations between classes.
- Olivetti Faces consists of 400 grayscale facial images (64×64 pixels) of 40 individuals, with 10 different facial expressions or lighting conditions per subject. This introduces intra-class variation and is suitable for testing embedding robustness.
- COIL-20 contains 1440 grayscale images of 20 objects taken from 72 different viewpoints (5-degree increments), introducing continuous rotational changes. Successful embeddings must preserve the geometric progression of these views.

In contrast, synthetic datasets such as Swiss Roll and Three Gaussians provide controlled environments for evaluating structure preservation. Real-world datasets present more complex distributions. MNIST, Pendigits, and Fashion-MNIST, composed of grayscale images with well-separated categories, serve as benchmarks for cluster fidelity. Olivetti Faces challenges embeddings with variation in lighting and expression, while COIL-20 emphasizes the need for preserving smooth geometric transformations.

5.2. Evaluation Metrics

We assess the performance of the K -means algorithm across various embedding spaces using both external and internal validation metrics. The external metrics include the F1-score [29,30] and Normalized Mutual Information (NMI) [31], both of which yield scores from 0 to 1, where higher values indicate better alignment with ground truth labels. For

internal validation, we use the Davies–Bouldin Index (DBI) [32], which quantifies cluster compactness and separation. Lower DBI values indicate better clustering performance.

5.3. Parameter Sensitivity

To analyze the influence of parameter weighting in our proposed loss functions, we conducted a sensitivity analysis of α (for $KL^{\max}$) and β (for $\mathcal{L}_{KL-W}$) on two datasets: Pendigits and COIL-20. We evaluated both quantitative results using NMI, F1-score, and DBI, as well as qualitative 2D visualizations of the embeddings.

While the full sensitivity sweep of α (for $KL^{\max}$) and β (for $\mathcal{L}_{KL-W}$) was performed on Pendigits and COIL-20, these two datasets were selected because they represent complementary geometric regimes: discrete, well separated clusters and smooth manifold structure. These regimes are known to generalize the behavior of neighbor-embedding objectives on high-dimensional image datasets such as MNIST and Fashion-MNIST. In practice, we found that applying the same parameter ranges ($\alpha \approx 0.5\text{--}0.7$ and $\beta \approx 250\text{--}500$) to the remaining datasets produced stable embeddings without degradation, suggesting that these values are reasonably robust across datasets.

5.3.1. Effect on NMI and F1-Score

Figures 1 and 2 illustrate how NMI and F1-score change under varying α and β values. For the Pendigits dataset, NMI initially increases slightly as α moves from 0.2 to 0.5, then drops sharply beyond $\alpha = 0.7$. This suggests that a moderate influence from $KL^{\max}$ enhances class separability, while excessive emphasis leads to embedding degradation.

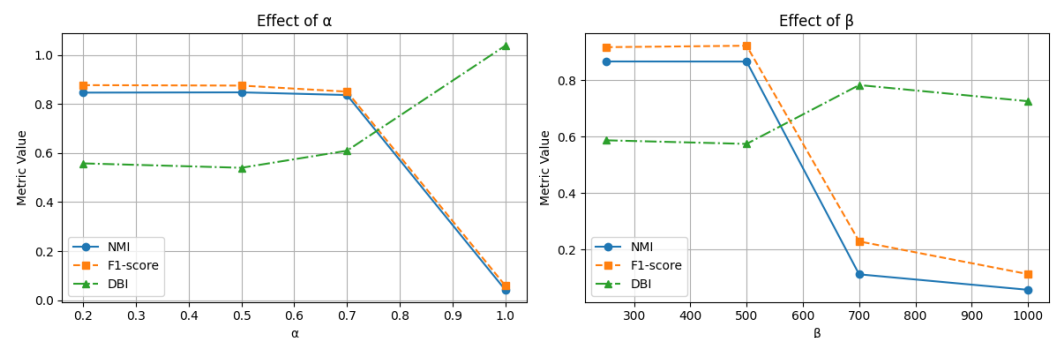


Figure 1. Effect of alpha and beta on performance based on Pendigits dataset.

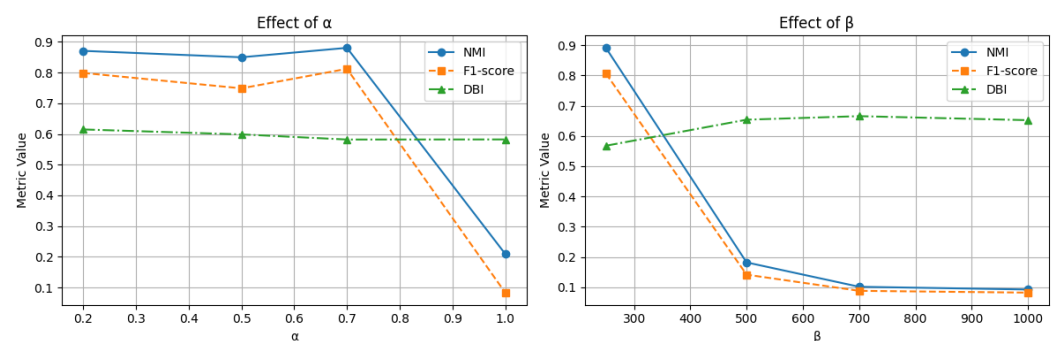


Figure 2. Effect of alpha and beta on performance based on COIL-20 dataset.

A similar trend is observed with β : NMI and F1 peak at $\beta = 250\text{--}500$, before degrading at $\beta = 1000$. These results confirm that both loss functions require careful parameter tuning to balance structural preservation and noise amplification. In the COIL-20 dataset, optimal NMI and F1 are reached at $\alpha = 0.7$ and $\beta = 250$. However, performance drops at higher values, indicating over-regularization may suppress the dataset’s inherent geometric continuity.

5.3.2. Effect on DBI and Interpretation

Interestingly, the DBI values for COIL-20 remain nearly constant across parameter changes, even when NMI and F1 degrade significantly (see Figure 2). This suggests that while the semantic alignment with class labels is affected, the geometric structure of the clusters remains stable. This is a direct result of the DBI's nature. It evaluates clustering quality based on intra-cluster compactness and inter-cluster separation, without relying on ground truth labels.

In contrast, for Pendigits (Figure 1), DBI trends align closely with NMI and F1. Since Pendigits consists of discrete digit classes, degradation in the embedding affects both structural and label-based clustering quality.

These differences highlight that DBI is more suitable for assessing structure-aware embeddings, particularly in datasets like COIL-20, where the class boundaries follow a smooth, continuous transformation (e.g., object rotation). The model may preserve structural integrity (low DBI) even as class labels become less distinguishable (low NMI/F1), especially under high α or β values.

5.3.3. Effect on Embedding Structure

These trends are reflected in the visualizations in Figures 3 and 4. In Pendigits (Figure 3), increasing α initially improves cluster tightness, but at $\alpha = 0.7$ and above, the layout collapses or degenerates into scattered blobs. Similar degradation is observed for large β values.

For COIL-20 (Figure 4), low α and β values produce visually clean and well-separated loops corresponding to object rotations. At higher values, the embeddings become tangled, yet the overall shape and spacing between clusters are partially preserved, explaining why DBI remains low despite declining NMI and F1.

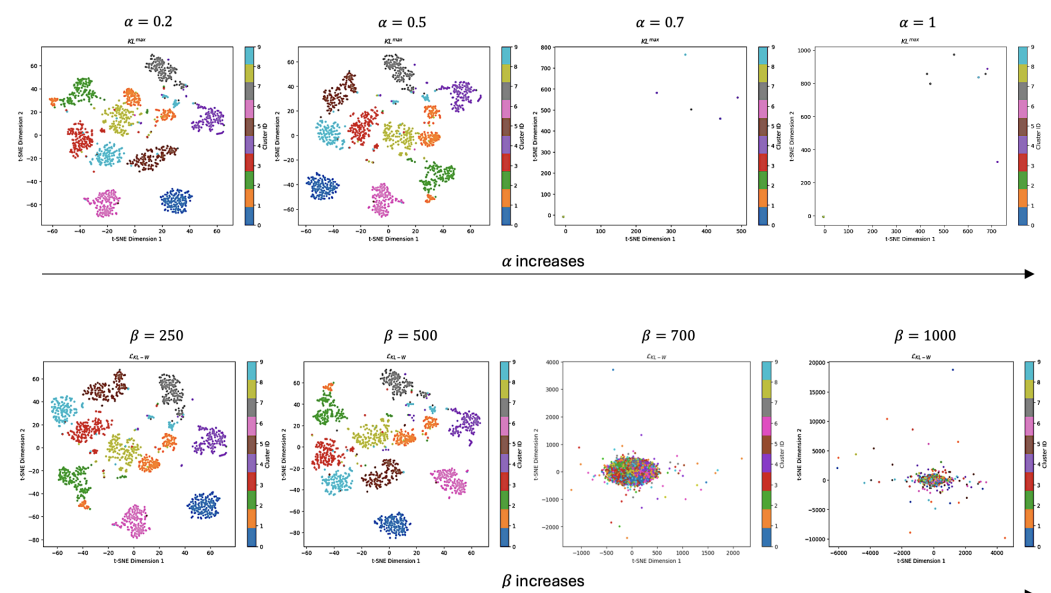


Figure 3. The change effect of α and β parameters on Pendigits dataset in terms of visualization on KL^{max} (Row 1) and $\mathcal{L}_{KL-W}$ (Row 2).

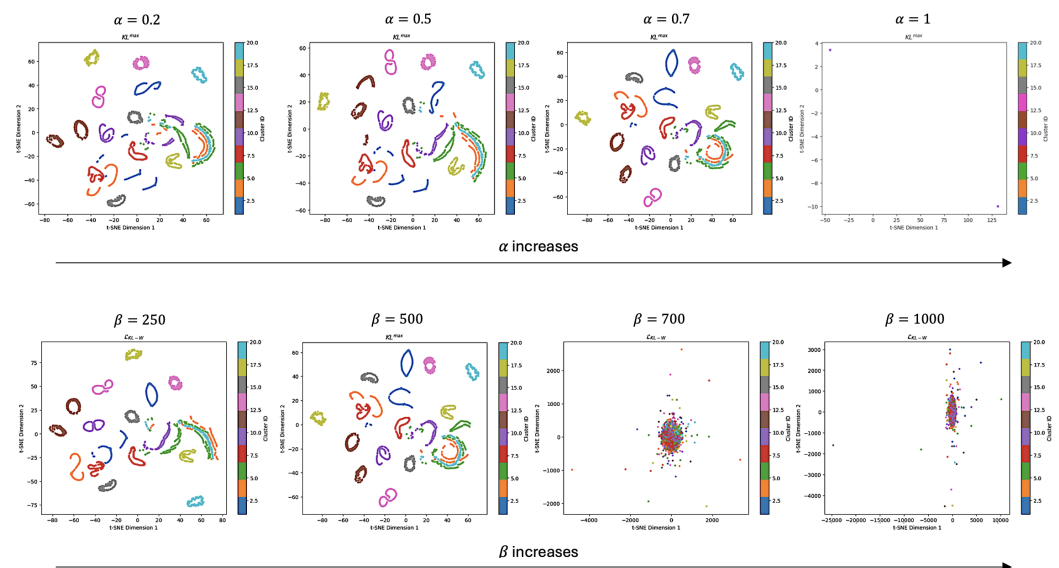


Figure 4. The change effect of α and β parameters on COIL-20 dataset in terms of visualization $KL^{\max}$ (Row 1) and $\mathcal{L}_{KL-W}$ (Row 2).

5.3.4. Summary

Moderate parameter values ($\alpha = 0.5\text{--}0.7$, $\beta = 250\text{--}500$) consistently yield the best trade-off between local fidelity and global structure, in terms of both external (NMI, F1) and internal (DBI) metrics. The near-constant DBI observed in COIL-20 underscores that structural cluster quality can remain high even when class-label agreement is weakened, an important insight for applications involving manifold or rotational structures. These findings reinforce the value of DBI as a robust structure, preservation metric and highlight the importance of understanding dataset-specific geometry when tuning model parameters.

5.4. Convergence Analysis

Despite the slightly higher loss values observed in the Kernel Density Estimation (KDE) plots in Figure 5, these increases are expected due to the inclusion of additional structural terms in the proposed losses. It is important to emphasize that higher numerical loss does not imply degraded embedding quality. Instead, the added terms in $KL^{\max}$ and $\mathcal{L}_{KL-W}$ help guide the optimization toward embeddings that are more structurally meaningful. This is validated by the consistently lower DBI scores and stronger clustering metrics (e.g., NMI and F1-score) achieved by our methods across all datasets.

The KDE plots, based on 50 independent runs per method and dataset, reveal that both $KL^{\max}$ and $\mathcal{L}_{KL-W}$ exhibit tighter and more symmetric loss distributions compared to standard KL divergence. This reflects greater optimization stability and reduced sensitivity to random initialization, suggesting that the proposed methods converge more consistently to robust minima.

To statistically verify these observations, we employed two non-parametric tests—the Kolmogorov–Smirnov (KS) test and the Mann–Whitney U test—because they make minimal distributional assumptions, are robust to outliers and skew, and provide complementary sensitivity: KS detects any distributional change (location, scale, or shape), while Mann–Whitney focuses on differences in central tendency dominance between groups [33,34]. The results, reported in Table 1, show that for all dataset comparisons, the p -values are below the significance threshold of 0.05. This confirms that the differences in loss distributions between KL and our proposed losses are statistically significant.

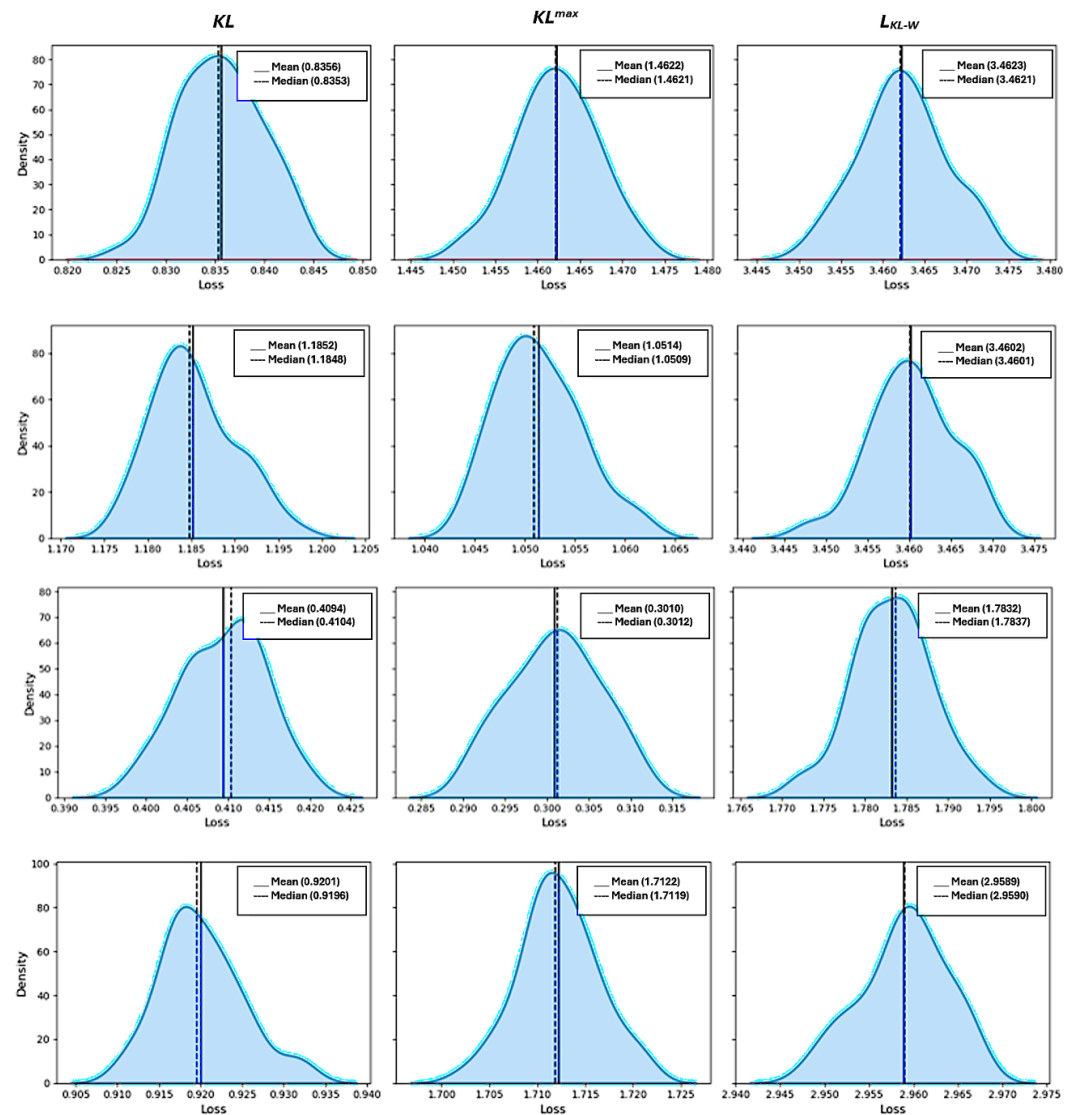


Figure 5. Kernel Density Estimation (KDE) plots illustrating the distribution of loss values based on 50 independent runs for each method. From left to right: KL , $KL^{\max}$, and $\mathcal{L}_{KL-W}$. From top to bottom: Pendigits, 5 classes from 5000 digits of MNIST, Coil-20, and FMNIST datasets.

These findings further strengthen our conclusion that $KL^{\max}$ and $\mathcal{L}_{KL-W}$ not only improve the quality of low-dimensional embeddings in terms of structure preservation and cluster separation, but also provide more stable and reproducible optimization behavior.

Table 1. Statistical comparison of loss distributions across KL , $KL^{\max}$, and $\mathcal{L}_{KL-W}$ using KS and Mann–Whitney U tests (p -values). The reported p -values appear as 0.0000 because they were rounded to four decimal places in the output; all are smaller than 0.0001 (<0.0001).

Dataset	Test	KL vs. $KL^{\max}$	KL vs. $\mathcal{L}_{KL-W}$
Pendigits	KS-Test	0.0000	0.0000
	Mann-Whitney U	0.0000	0.0000
MNIST	KS-Test	0.0000	0.0000
	Mann-Whitney U	0.0000	0.0000
Coil-20	KS-Test	0.0000	0.0000
	Mann-Whitney U	0.0000	0.0000
FMNIST	KS-Test	0.0000	0.0000
	Mann-Whitney U	0.0000	0.0000

5.5. Computational Complexity Analysis

To evaluate the practical feasibility of the proposed loss functions, we measured the total execution time required to optimize the embeddings on four benchmark datasets. Table 2 summarizes the runtime (in seconds) of standard t -SNE with KL divergence and the modified versions using $KL^{\max}$ and $\mathcal{L}_{KL-W}$.

As expected, both $KL^{\max}$ and $\mathcal{L}_{KL-W}$ introduce additional computational overhead compared to the baseline. This increase stems from the inclusion of extra components in the objective function, ranking deviation terms in $KL^{\max}$ and $\mathcal{L}_{KL-W}$. However, this added cost is relatively modest.

Importantly, the improvements in embedding quality reflected in superior clustering scores, lower DBI values, and increased visual interpretability justify the modest increase in runtime. The structure aware regularization terms provide better guidance during optimization, resulting in embeddings that are both more informative and more stable, without incurring prohibitive computational costs.

Table 2. Execution time (in seconds) of each method on the benchmark datasets.

Dataset	KL	$KL^{\max}$	$\mathcal{L}_{KL-W}$
Pendigits	55.14	79.93	83.43
MNIST	1008.94	1202.22	1111.88
Fashion-MNIST	2069.63	2212.24	2087.94
COIL-20	400.52	630.45	570.55

5.6. Comprehensive Evaluation of Clustering, Structure, and Reconstruction

This section evaluates the proposed divergence measures, $KL^{\max}$ and $\mathcal{L}_{KL-W}$, against the standard KL divergence in terms of clustering quality, structural preservation, and reconstruction accuracy. We apply the methods across six benchmark datasets with varying characteristics, including synthetic and real-world data.

5.6.1. Clustering Metrics

As shown in Table 3, both proposed losses outperform the standard KL divergence in terms of F1-score and NMI across most datasets. On simpler datasets such as 3-Gaussians, all methods perform comparably. However, on more challenging datasets like Pendigits, MNIST, and COIL-20, $KL^{\max}$ and $\mathcal{L}_{KL-W}$ achieve clear improvements. For instance, in Pendigits, $\mathcal{L}_{KL-W}$ improves the NMI from 0.833 to 0.872 and the F1-score from 0.852 to 0.908. These enhancements demonstrate better alignment between the embedded space and the underlying class structure.

Table 3. Comparison of Metrics for KL , $KL^{\max}$, and $\mathcal{L}_{KL-W}$.

Dataset	Method	NMI	F1-Score	DBI
3-Gaussians	KL	0.929	0.983	0.458
	$KL^{\max}$	0.929	0.983	0.347
	$\mathcal{L}_{KL-W}$	0.920	0.975	0.353
Swiss Roll	KL	0.294	0.835	0.744
	$KL^{\max}$	0.300	0.886	0.698
	$\mathcal{L}_{KL-W}$	0.303	0.877	0.602
Pendigits	KL	0.833	0.852	0.597
	$KL^{\max}$	0.860	0.881	0.540
	$\mathcal{L}_{KL-W}$	0.872	0.908	0.563

Table 3. Cont.

Dataset	Method	NMI	F1-Score	DBI
5-MNIST	KL	0.802	0.919	0.692
	$KL^{\max}$	0.862	0.963	0.629
	$\mathcal{L}_{KL-W}$	0.842	0.930	0.638
Coil-20	KL	0.851	0.763	0.637
	$KL^{\max}$	0.865	0.794	0.615
	$\mathcal{L}_{KL-W}$	0.888	0.808	0.600
Fashion-MNIST	KL	0.826	0.805	0.564
	$KL^{\max}$	0.831	0.880	0.672
	$\mathcal{L}_{KL-W}$	0.810	0.835	0.619

5.6.2. Internal Clustering Quality

The DBI is used to assess cluster compactness and separation. As shown in Table 3, both $KL^{\max}$ and $\mathcal{L}_{KL-W}$ generally yield lower DBI scores than the baseline KL , indicating tighter and more distinct clusters. For instance, on the COIL-20 dataset, the DBI decreases from 0.637 (KL) to 0.600 ($\mathcal{L}_{KL-W}$), reflecting improved structural grouping. However, the slightly lower DBI of the baseline on Fashion-MNIST can be attributed to that dataset's simple, compact cluster structure, its local neighborhoods are already well defined, favoring methods that prioritize local cohesion. In contrast, our proposed methods emphasize global geometric alignment, so a small trade-off in intra-cluster compactness is expected.

5.6.3. Visual Inspection of Embeddings

Figure 6 provides qualitative comparisons of the 2D embeddings produced by each method. Embeddings generated using $\mathcal{L}_{KL-W}$ show more distinct and well-separated clusters, especially in datasets with complex manifolds like Swiss Roll and COIL-20. The consistency and continuity in structure indicate that $\mathcal{L}_{KL-W}$ effectively captures long-range relationships in the data. Meanwhile, $KL^{\max}$ produces embeddings with sharper cluster boundaries and reduced within-cluster variance, showing improvements over standard KL divergence in terms of local structure.

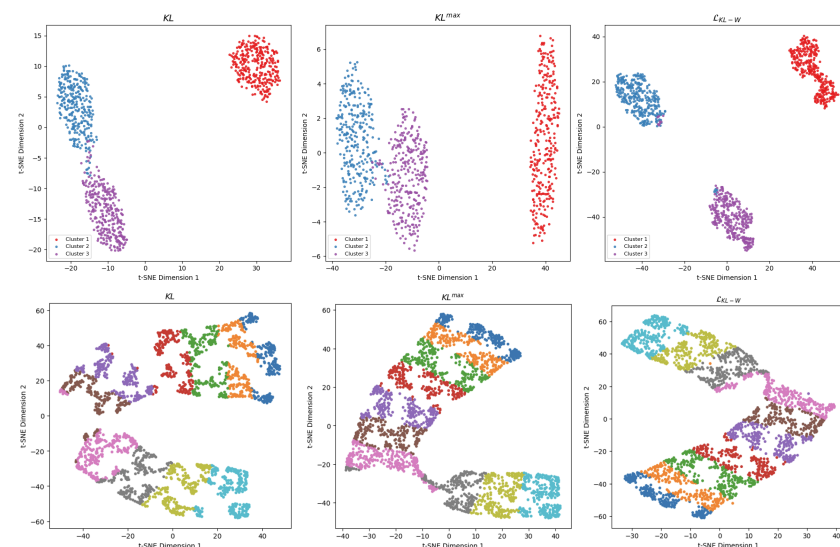


Figure 6. Cont.

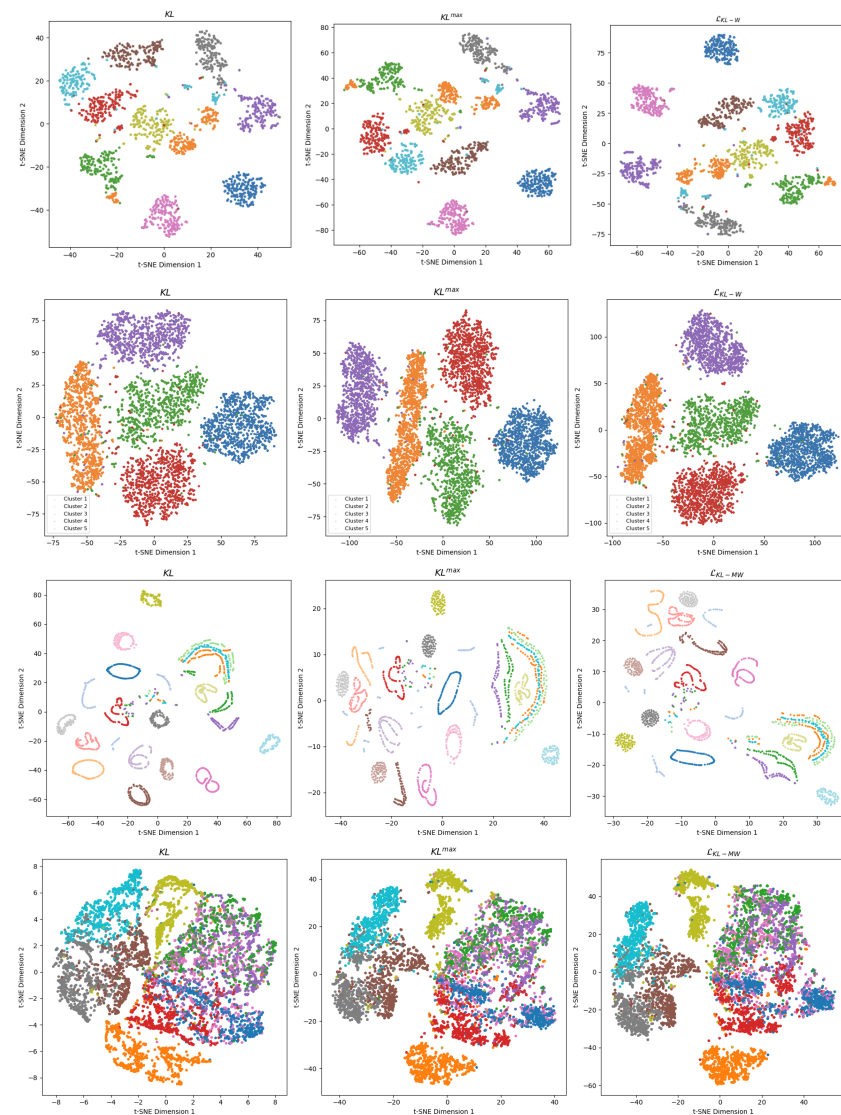


Figure 6. Visual comparison of embeddings using, from left to right, KL , $KL^{\max}$, and $\mathcal{L}_{KL-W}$; and from top to bottom, 3-Gaussian clusters, Swiss Roll, Pendigits, 5 classes of 5000 MNIST digits, Coil-20, and FMNIST datasets. In all panels, point colors indicate the ground-truth class or cluster label in the corresponding dataset (e.g., Gaussian component, roll arm, digit identity, or object category).

5.6.4. Out-of-Sample and Reconstruction Quality

To address the out-of-sample limitation inherent in t -SNE and its variants, a Multi-Layer Perceptron (MLP) regressor was employed to approximate the mapping from high-dimensional input data to their corresponding low-dimensional embeddings. Separate models were trained using embeddings generated by t -SNE with KL , $KL^{\max}$, and $\mathcal{L}_{KL-W}$ loss functions. Once trained, the MLP enables fast and efficient projection of new, unseen data points into the embedding space without rerunning the full optimization procedure, making the method more scalable and practical for dynamic or streaming scenarios.

The quality of these learned mappings was evaluated using MSE between the predicted embeddings and the original embeddings produced by each method. As shown in Table 4, both $KL^{\max}$ and $\mathcal{L}_{KL-W}$ achieve lower MSE than the standard KL -based t -SNE, with $KL^{\max}$ achieving the best reconstruction performance (MSE = 0.0014). This indicates stronger generalization capabilities and more faithful preservation of the underlying data structure, because the standard asymmetric $KL(P \parallel Q)$ primarily penalizes missed neighbors and can tolerate false neighbors (leading to crowding and global distortions),

whereas $KL^{\max}$ penalizes discrepancies in both directions, reducing spurious attractions and better aligning local–global relationships; additionally, $\mathcal{L}_{KL-W} = KL + \lambda W$ augments the objective with a Wasserstein term that encourages globally coherent layouts, further lowering reconstruction error.

To further assess the preservation of structural and semantic features, a reconstruction experiment was conducted using the Olivetti Faces dataset. Figure 7 shows that the embeddings derived from $KL^{\max}$ and $\mathcal{L}_{KL-W}$ allow for the recovery of faces with clearer contours, sharper expressions, and reduced artifacts. In contrast, reconstructions from KL -based embeddings suffer from noticeable blurring and loss of detail. These qualitative and quantitative results confirm the superiority of the proposed loss functions in preserving essential information for both embedding quality and out-of-sample extension.



Figure 7. Reconstruction of Olivetti Faces using KL , $KL^{\max}$ and $\mathcal{L}_{KL-W}$, with MSE values shown for each method. Images from the AT/T Laboratories Cambridge Database of Faces; used with attribution [28].

Table 4. Average MSE of Reconstructed Faces in the Olivetti Dataset.

Method	KL	$KL^{\max}$	$\mathcal{L}_{KL-W}$
MSE	0.0027	0.0014	0.0019

6. Advantages, Limitations, and Applications

6.1. Advantages

The proposed loss functions, $KL^{\max}$ and $\mathcal{L}_{KL-W}$, offer several advantages over the KL divergence in the context of dimensionality reduction and data visualization:

- **Improved Similarity Ranking:** The introduction of $KL^{\max}$ enhances the model's ability to preserve similarity rankings, making the embeddings more consistent with the underlying class structure.
- **Better Reconstruction Quality:** Both loss functions reduce reconstruction error (as reflected in lower MSE values), and the visual results demonstrate sharper and more accurate data recovery.
- **Robustness Across Datasets:** Empirical results show consistent performance gains on a variety of datasets, including structured (e.g., Swiss Roll) and real-world datasets (e.g., MNIST, Olivetti Faces).

6.2. Limitations

Despite the demonstrated benefits, the proposed approaches also have certain limitations:

- **Increased Computational Complexity:** Integrating Wasserstein distance or ranking-based terms into the loss functions introduces additional computational overhead relative to standard t -SNE. Specifically, $KL^{\max}$ adds $O(N^2)$ pairwise comparisons with an additional ranking step that marginally increases runtime (about 10%), while $\mathcal{L}_{KL-W}$ incorporates a batch-wise Wasserstein distance term with complexity $O(mb \log b)$, where m is the number of random directions and b is the batch size.
- **Hyperparameter Sensitivity:** Parameters such as the exaggeration factor or weighting coefficients require tuning, which may limit generalizability without cross-validation.

6.3. Applications

The proposed loss functions and enhancements to t -SNE are well-suited for many applications:

- **Bioinformatics and Genomics:** For interpreting gene expression data or cell populations where subtle structures (e.g., differentiation trajectories) matter.
- **Facial Recognition and Image Retrieval:** Improved clustering and reconstruction quality make these methods effective for feature extraction and identity preservation.
- **Out-of-Sample Generalization:** One of the key limitations of traditional t -SNE lies in its inability to naturally handle out-of-sample data. Since the algorithm lacks an explicit mapping function, incorporating new data points requires re-running the full embedding process, which is computationally expensive and unsuitable for dynamic or streaming environments.

7. Discussion

This work set out to address two well-known limitations of t -SNE's KL -based objective, its emphasis on local neighborhoods at the expense of global structure and its asymmetric sensitivity, by introducing $KL^{\max}$ and the $\mathcal{L}_{KL-W}$ loss. Across synthetic and real-world datasets, our experiments show that both objectives improve cluster separability and structural clarity relative to standard KL , with $\mathcal{L}_{KL-W}$ most consistently enhancing long-range organization and $KL^{\max}$ sharpening local neighborhoods. These trends are reflected quantitatively (higher NMI/F1 and lower or comparable DBI) and qualitatively in the 2D embeddings. Notably, modest increases in aggregate loss values for the proposed objectives do not translate to degraded embeddings; instead, they arise from additional

structure-aware terms that guide the optimizer toward configurations that better respect the data geometry.

Our findings align with and extend prior observations that alternative divergences or geometry-aware criteria can remedy KL 's biases in neighbor embedding. By explicitly penalizing rank deviations (KL^{max}) and by incorporating transport geometry (L_{KL-W}), the proposed losses reconcile local fidelity with broader organization. In particular, L_{KL-W} leverages Wasserstein's smooth, geometry-aware transport to better capture manifold continuity (e.g., COIL-20 rotations), while KL^{max} reduces within-cluster variance by enforcing consistency among the highest-similarity relationships. The resulting embeddings exhibit clearer global layouts without sacrificing the hallmark local detail that makes t -SNE appealing.

A practical contribution of this study is guidance on hyperparameter regimes. Sensitivity analyses indicate that moderate values, $\alpha \approx 0.5$ – 0.7 for KL^{max} and $\beta \approx 250$ – 500 for L_{KL-W} , strike a robust balance between local ranking fidelity and global structure, with performance typically degrading if either term dominates the objective. Interestingly, COIL-20's DBI remains relatively stable even when NMI/F1 decline at high α or β , suggesting that structure-only metrics can under- or over-estimate semantic separability when manifold continuity drives class organization. This divergence between internal and external indices highlights the importance of reporting both metric families and of visually inspecting embeddings when the data have strong geometric factors.

From an optimization perspective, both KL^{max} and L_{KL-W} demonstrate tighter, more symmetric loss distributions over repeated runs and statistically significant differences relative to the KL baseline (KS and Mann–Whitney U tests, $p < 0.05$ across datasets), indicating improved robustness to initialization. While the added terms introduce a modest runtime overhead, the computational cost remains practical on commodity hardware and is offset by gains in stability and interpretability. Moreover, out-of-sample mapping via a simple MLP shows lower reconstruction MSE, particularly for KL^{max} , suggesting that the learned embeddings expose structure that is easier to approximate with smooth predictors, a property valuable for streaming or incremental use cases.

Limitations remain; for example, the methods introduce additional hyperparameters and, for L_{KL-W} , require approximations to make transport terms tractable at scale. Although our experiments cover diverse datasets, broader evaluations—e.g., high-throughput single-cell data or large-vocabulary text embeddings—could further validate generality. In addition, while we observed that higher composite losses need not imply worse visualization quality, a deeper theoretical characterization of the proposed objectives' minima and their relationship to cluster topology would be valuable.

For future work, several avenues follow naturally: (i) combine KL^{max} and L_{KL-W} into a single objective to jointly enforce ranking and transport geometry, potentially with adaptive weighting during training; (ii) explore scalable transport surrogates (e.g., sliced or linearized optimal transport) tailored to neighbor-embedding kernels to further reduce overhead; (iii) integrate label information in a semi-supervised variant to explicitly balance semantic and geometric structure; (iv) develop principled, data-dependent schedules for α and β based on online diagnostics (e.g., changes in DBI vs. NMI), reducing manual tuning; and (v) extend the out-of-sample mapping with uncertainty estimation to flag projections likely to deviate from the manifold learned by the embedding.

Author Contributions: Conceptualization, R.H. and S.B.B.; methodology, R.H. and S.B.B.; software, S.N.; validation, S.N., R.H. and S.B.B.; formal analysis, S.N.; investigation, S.N.; resources, S.N.; data curation, S.N.; writing original draft preparation, S.N.; reviewing and editing, R.H. and S.B.B.; supervision, R.H. and S.B.B.; project administration, R.H. and S.B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Qatar National Research Fund (QNRF), grant number GSRA9-L-1-0512-22011.

Data Availability Statement: All datasets used in this study are publicly available. The MNIST dataset is accessible at <https://www.kaggle.com/datasets/hojjatk/mnist-dataset>. The Pendigits dataset can be obtained from the UCI Machine Learning Repository at <https://archive.ics.uci.edu/dataset/81/pen+based+recognition+of+handwritten+digits>. The Fashion-MNIST dataset is available at <https://github.com/zalandoresearch/fashion-mnist>. The Olivetti Faces (ORL) dataset is available at <https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>. The COIL-20 dataset can be accessed at <https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
2. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
3. van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
4. Wan, N.; Li, D.; Hovakimyan, N. F-divergence variational inference. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17370–17379.
5. Arora, S.; Hu, W.; Kothari, P.K. An analysis of the t-SNE algorithm for data visualization. In *Proceedings of the Conference on Learning Theory, Stockholm, Sweden, 6–9 July 2018*; PMLR: Cambridge, MA, USA, 2018; pp. 1455–1462.
6. Im, D.J.; Verma, N.; Branson, K. Stochastic neighbor embedding under f-divergences. *arXiv* **2018**, arXiv:1811.01247. [CrossRef]
7. Yao, W.; Yang, W.; Wang, Z.; Lin, Y.; Liu, Y. Revisiting weak-to-strong generalization in theory and practice: Reverse KL vs. Forward KL. *arXiv* **2025**, arXiv:2502.11107.
8. Yang, Z.; Chen, Y.; Corander, J. t-SNE is not optimized to reveal clusters in data. *arXiv* **2021**, arXiv:2110.02573. [CrossRef]
9. Naderializadeh, N.; Li, R.; Xiao, D.; Shrivastava, A.; Soatto, R. Set representation learning with generalized sliced-Wasserstein embeddings. *arXiv* **2021**, arXiv:2103.03892. [CrossRef]
10. Kolouri, S.; Park, S.R.; Thorpe, M.; Slepcev, D.; Rohde, G.K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* **2017**, *34*, 43–59. [CrossRef]
11. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2292–2300.
12. Genevay, A.; Cuturi, M.; Peyré, G.; Bach, F. Stochastic optimization for large-scale optimal transport. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3440–3448.
13. Seguy, V.; Damodaran, B.B.; Flamary, R.; Courty, N.; Rolet, A.; Blondel, M. Large-scale optimal transport and mapping estimation. *arXiv* **2017**, arXiv:1711.02283.
14. Gallouët, T.; Ghezzi, R.; Vialard, F.X. Regularity theory and geometry of unbalanced optimal transport. *J. Funct. Anal.* **2025**, *289*, 111042. [CrossRef]
15. Liu, Y.; Wang, C.; Lu, M.; Yang, J.; Gui, J.; Zhang, S. From simple to complex scenes: Learning robust feature representations for accurate human parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5449–5462. [CrossRef] [PubMed]
16. Gao, Z.; Wang, Y.; Chen, Y.; Feng, X.; Ji, H. Light-field image multiple reversible robust watermarking against geometric attacks. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7457–7471. [CrossRef]
17. Wang, Y.; Wan, R.; Wang, Q.; Chan, K.C.C.; Chen, H. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *J. Mach. Learn. Res.* **2021**, *22*, 1–45.
18. Zhou, Y.; Sharpee, T.O. Using global t-SNE to preserve intercluster data structure. *Neural Comput.* **2022**, *34*, 1637–1651. [CrossRef]
19. Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10*, 5416. [CrossRef]
20. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 338.
21. Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* **1991**, *44*, 375–417. [CrossRef]

22. Kantorovich, L.V. On the translocation of masses. *Dokl. Akad. Nauk. USSR (NS)* **1942**, *37*, 199–201. [\[CrossRef\]](#)
23. Peyré, G.; Cuturi, M. Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [\[CrossRef\]](#)
24. Alpaydin, E.; Alimoglu, F. Pen-based recognition of handwritten digits data set. *Mach. Learn. Repos.* **1998**, *4*, 2.
25. LeCun, Y. The MNIST Database of Handwritten Digits. Available online: <http://yann.lecun.com/exdb/mnist/> (accessed on 18 November 2025).
26. Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv* **2017**, arXiv:1708.07747. [\[CrossRef\]](#)
27. Nene, A.; Nayar, K.; Murase, H. *Columbia Object Image Library (COIL-20)*; Tech. Rep.; Columbia University: New York, NY, USA, 1996; pp. 223–303.
28. AT&T Laboratories Cambridge. *The Database of Faces (Olivetti/ORL Faces)*; AT&T Laboratories Cambridge: Cambridge, UK, 1994.
29. Van Rijsbergen, C. Information retrieval: Theory and practice. In Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems, Suita, Japan, 4–7 September 1979; Volume 79; pp. 1–14.
30. Powers, D.M.W. What the F-measure doesn't measure: Features, flaws, fallacies and fixes. *arXiv* **2015**, arXiv:1503.06410.
31. Strehl, A.; Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
32. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *PAMI-1*, 224–227. [\[CrossRef\]](#)
33. Massey, F.J., Jr. The Kolmogorov–Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78. [\[CrossRef\]](#)
34. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.