ROBUSTVIS: HIGH-DIMENSIONAL DATA VISUALIZATION FROM SPARSE SIMILARITY GRAPHS WITH UNCERTAIN NON-EDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neighborhood-graph-based high-dimensional data visualization methods depend on pairwise relations that are inevitably imperfect. Under conventional cross-entropy objectives, a single false-positive edge or false-negative non-edge can have unbounded loss and gradient influence near the boundary of the low-dimensional affinity. We introduce RobustVis, a bounded-influence optimization framework for high-dimensional data visualization. Its asymmetric positive- and negative-pair losses place robustness parameters inside the logarithms, uniformly bounding pairwise losses and derivatives while retaining strong attraction for trusted edges. We establish perturbation bounds for both the objective and the embedding gradient under graph corruption, and derive an importance-weighted negative-pair estimator for scalable optimization. Moreover, we develop an extension for RobustVis that can effectively account for uncertainty in positive edges. Experiments on synthetic manifold data and diverse real datasets demonstrate the effectiveness of the proposed methods in comparison to strong baselines.

1 INTRODUCTION

High-dimensional data visualization maps complex observations to two or three dimensions for exploratory analysis, scientific discovery, and downstream modeling (Jolliffe & Cadima, 2016; Tenenbaum et al., 2000; Hinton & Salakhutdinov, 2006; van der Maaten et al., 2009; Moon et al., 2019; Schneider et al., 2023). Modern methods such as t-SNE (Van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018) are particularly successful at producing visually informative embeddings and have become standard tools across many domains. Their success has motivated numerous extensions that improve scalability, global organization, parametric mapping, or the balance between local and global structure (Tang et al., 2016; Amid & Warmuth, 2019; Wang et al., 2021; Sainburg et al., 2021; Jeon et al., 2025; Dagès et al., 2025; Kury et al., 2026).

Many nonlinear visualization methods rely on a pairwise affinity matrix or a neighborhood graph. In practice, however, neighborhood graphs are finite, data-dependent approximations to latent geometric or semantic relations. Noise, irrelevant features, heterogeneous densities, distance concentration, and metric misspecification can introduce false-positive edges, while finite-K truncation and mutual-neighbor filtering can omit pairs that should have non-negligible affinity. Although not every missing edge is a false negative, both the observed edge set and its much larger complement inevitably contain uncertain relations. Such inaccuracies can propagate through the embedding objective and contribute to misleading local or global geometry in the resulting visualization. Prior studies report failures to recover manifold structure (Im et al., 2018), sensitivity to false negatives (Yang et al., 2021), and unreliable geometric or cluster-related visual cues (Chari & Pachter, 2023; Bergam et al., 2025). Although these effects are not caused solely by graph errors, they motivate a complementary question: *how should an embedding objective behave when the supplied neighborhood relations are imperfect?*

A natural way to learn an embedding from a neighborhood graph is to treat each pairwise relation as a Bernoulli target and match it with the corresponding low-dimensional affinity. This yields a binary attraction–repulsion objective closely related to the fuzzy-set cross-entropy in the nominal formulation of UMAP (McInnes et al., 2018). Practical UMAP, however, uses asymmetric nega-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tive sampling and is more accurately characterized by a reweighted effective objective (Damrich & Hamprecht, 2021). The objective of UMAP is retained or extended by methods such as Parametric UMAP (Sainburg et al., 2021), densMAP (Narayan et al., 2021), UMATO (Jeon et al., 2025), and StarMAP (Watanabe et al., 2025). We emphasize that binary graph-matching objectives are not universal in dimensionality reduction. For example, t-SNE minimizes a KL divergence between normalized pairwise distributions (Van der Maaten & Hinton, 2008), PaCMAP uses bounded rational potentials over neighbor, mid-near, and further pairs (Wang et al., 2021), and classical manifold-learning methods employ spectral, reconstruction, or stress-based criteria. Our analysis therefore focuses specifically on the robustness of binary logarithmic graph-matching objectives under imperfect neighborhood relations.

We propose *RobustVis*, which limits the influence of uncertain relations while retaining strong attraction for trusted edges. Its asymmetric design reflects the imbalance between relatively few observed edges and abundant non-edges that may include latent neighbors. Our contributions are:

- We identify the unbounded influence of erroneous relations in binary graph matching and propose an asymmetric objective with uniformly bounded pairwise losses and affinity derivatives.
- We establish objective- and gradient-perturbation bounds under false-positive edges and false-negative non-edges, and derive importance-weighted negative sampling that preserves the objective and its gradient in expectation.
- We develop a confidence-weighted extension that retains the binary graph support while adjusting attraction according to positive-edge reliability.

Experiments on 3 synthetic and 13 real datasets demonstrate the effectiveness of RobustVis and its variants against 10 competitors.

2 Related Work

High-dimensional data visualization. Classical dimensionality-reduction methods preserve different aspects of the input geometry. PCA seeks a linear subspace of maximum variance (Pearson, 1901; Hotelling, 1933; Jolliffe & Cadima, 2016), while multidimensional scaling preserves pairwise dissimilarities (Torgerson, 1952; Shepard, 1962). Nonlinear manifold methods, including Isomap (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000), Laplacian eigenmaps (Belkin & Niyogi, 2003), Hessian LLE (Donoho & Grimes, 2003), and LTSA (Zhang & Zha, 2004), exploit local geometric structure, but can be sensitive to noise, graph disconnection, or inappropriate neighborhood scales (Fan et al., 2018). Modern visualization methods use probabilistic affinities, graphs, or sampled pairwise constraints to balance local and global structure. Representative examples include t-SNE (Van der Maaten & Hinton, 2008), LargeVis (Tang et al., 2016), UMAP (McInnes et al., 2018), TriMap (Amid & Warmuth, 2019), PaCMAP (Wang et al., 2021), PHATE (Moon et al., 2019), UMATO (Jeon et al., 2025), LocalMAP (Wang et al., 2025), Finsler MDS (Dagès et al., 2025), and DREAMS (Kury et al., 2026). Our RobustVis differs from these methods in its primary focus: rather than introducing a new geometric constraint or neighborhood topology, it studies how the embedding objective should respond when the supplied neighborhood relations are imperfect.

Neighborhood graphs and pairwise objectives. Neighborhood graphs are central to many manifold-learning and visualization methods. Isomap, LLE, and Laplacian eigenmaps construct local neighborhoods using KNN-type graphs and preserve geodesic, reconstruction, or spectral structure, respectively (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003). Although t-SNE does not explicitly construct a sparse graph, its locally scaled high-dimensional affinities strongly emphasize nearby samples (Van der Maaten & Hinton, 2008). UMAP instead constructs a weighted fuzzy neighborhood graph (McInnes et al., 2018), while mutual-KNN constructions retain only reciprocally supported relations (Dalmia & Sia, 2021).

LargeVis and the nominal formulation of UMAP optimize logarithmic attraction and repulsion terms for positive and negative pairs (Tang et al., 2016; McInnes et al., 2018). Practical UMAP uses asymmetric negative sampling, which induces a reweighted effective objective rather than the nominal full cross-entropy (Damrich & Hamprecht, 2021). Nevertheless, these objectives retain logarithmic singularities: an erroneous positive or negative relation can produce an unbounded loss or affinity derivative near the boundary. This issue is not universal across pairwise visualization methods; for example, PaCMAP uses bounded rational potentials over neighbor, mid-near, and further pairs

(Wang et al., 2021). Our RobustVis specifically robustifies weighted Bernoulli graph matching by uniformly bounding the positive- and negative-pair losses and derivatives.

3 RobustVis: Bounded-Influence Neighborhood Embedding

3.1 Problem Formulation and Affinity Graph

Given $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$, we seek an embedding $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]^\top \in \mathbb{R}^{n\times l}$, typically $l \in \{2, 3\}$ with $l \ll d$, that preserves relevant geometric or semantic relations. We represent the input relations by a symmetric sparse affinity matrix $\mathbf{S} \in [0, 1]^{n\times n}$ with $S_{ii} = 0$, where $S_{ij} > 0$ indicates an observed neighbor relation and its strength, whereas $S_{ij} = 0$ denotes an unobserved relation. The matrix $\mathbf{S}$ can be constructed using different neighborhood procedures. For example, let $\mathcal{N}_K(i)$ contain the K nearest neighbors of $\mathbf{x}_i$ under a chosen input-space distance. Symmetric KNN graphs connect a pair when either endpoint is in the other's neighborhood, whereas mutual-KNN graphs require both. Edge weights may be binary, distance-based, rank-based, or fuzzy affinities such as those used by UMAP. The resulting graph $G = (\mathcal{V}, \mathcal{E})$ partitions all unordered pairs into observed positive and negative sets:

$$\Omega^+ = \{(i, j) : i < j,\ S_{ij} > 0\}, \qquad \Omega^- = \{(i, j) : i < j,\ S_{ij} = 0\}, \tag{1}$$

where dependence on $\mathbf{S}$ is omitted for brevity and $N = n(n-1)/2 = |\Omega^+| + |\Omega^-|$.

These observed relations need not reflect the underlying structure. To characterize graph corruption, let $\mathbf{S}^\star$ denote the unknown ideal affinity matrix. A pair is a false-positive edge if $S_{ij} > 0$ but $S^\star_{ij}$ is negligible, or a false-negative non-edge if $S_{ij} = 0$ but $S^\star_{ij}$ is non-negligible. Such errors may arise from noisy or irrelevant features, metric misspecification, heterogeneous sampling densities, finite-K truncation, or mutual-neighbor filtering. In KNN graphs, increasing K may recover missing latent neighbors but also connect samples that are close in the observed space yet distant in the latent geometry. $\mathbf{S}^\star$ serves only as a conceptual reference and is not available during embedding.

To represent pairwise relations in the embedding, we define $\mathbf{A}$ using a scaled Cauchy kernel:

$$A_{ij} = \left(1 + \frac{\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}{\sigma^2}\right)^{-(l+1)/2}, \qquad i \neq j, \tag{2}$$

where $\sigma > 0$ controls the distance scale. Our goal is therefore to learn $\mathbf{Z}$ whose induced affinities $\mathbf{A}$ match the observed matrix $\mathbf{S}$ while limiting the influence of erroneous relations.

3.2 Sensitivity of Binary Graph Matching

A natural approach is to interpret S_{ij} as a Bernoulli target and minimize the pairwise cross-entropy

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{Z}; \mathbf{S}) = -\frac{1}{N} \sum_{i<j} \left[S_{ij} \log A_{ij} + (1 - S_{ij}) \log(1 - A_{ij})\right]. \tag{3}$$

For the pairwise loss $\ell_{\mathrm{CE}}(A_{ij}, S_{ij}) = -S_{ij} \log A_{ij} - (1 - S_{ij}) \log(1 - A_{ij})$, its derivative with respect to the low-dimensional affinity is

$$\frac{\partial \ell_{\mathrm{CE}}(A_{ij}, S_{ij})}{\partial A_{ij}} = -\frac{S_{ij}}{A_{ij}} + \frac{1 - S_{ij}}{1 - A_{ij}}. \tag{4}$$

For a false-positive edge with $S_{ij} > 0$, the attractive loss and its affinity derivative become unbounded as $A_{ij} \to 0$, even when separating the pair is geometrically appropriate. Conversely, a false-negative non-edge has $S_{ij} = 0$ and contributes $-\log(1 - A_{ij})$, whose loss and derivative diverge as $A_{ij} \to 1$.

The practical optimization used by UMAP should be distinguished from (3). Damrich & Hamprecht (2021) has shown that UMAP's asymmetric negative-sampling updates are more accurately described by an effective objective with substantially reweighted attraction and repulsion, rather than by the nominal full cross-entropy itself. Specifically, the actual loss function of UMAP is

$$\mathcal{L}_{\mathrm{UMAP}}(\mathbf{Z}; \boldsymbol{\mu}) = -2 \sum_{i<j} \left[\mu_{ij} \log A_{ij} + \frac{m\,(d_i + d_j)}{2n} \log\left(1 - A_{ij}\right)\right] \tag{5}$$

where μ_{ij} is the high-dimensional UMAP affinity, $d_i = \sum_j \mu_{ij}$, m is the negative-sampling rate, and $A_{ij} \in (0,1)$ is the low-dimensional affinity. Thus, practical UMAP replaces the repulsive weight $1 - \mu_{ij}$ in the nominal full cross-entropy by the usually much smaller quantity $m\,(d_i + d_j)\,/(2n)$. Its negative sampling therefore implicitly rebalances attraction and repulsion and makes the effective target affinities nearly binary. However, $\mathcal{L}_{\text{UMAP}}$ remains a weighted cross-entropy: it only multiplies the positive and negative logarithmic terms by different coefficients and retains unbounded derivatives with respect to the affinity.

3.3 Asymmetric Bounded-influence Objective

The analysis in Section 3.2 motivates an objective that uniformly limits the influence of both observed edges and unobserved pairs. To this end, we propose to minimize the following robust loss

$$\mathcal{L}_{\text{RV}}(\mathbf{Z};\mathbf{S}) = \frac{1}{N}\left[\sum_{(i,j)\in\Omega^+} \ell_{ij}^+ + \sum_{(i,j)\in\Omega^-} \ell_{ij}^-\right], \tag{6}$$

where

$$\ell_{ij}^+ = -S_{ij}\log(A_{ij}+\alpha) - (1-S_{ij})\log(1-A_{ij}+\alpha), \tag{7}$$

$$\ell_{ij}^- = -\log(1-A_{ij}+\beta). \tag{8}$$

In the robust loss function, $\alpha, \beta > 0$ are two hyperparameters. Adding $\log(1+\alpha)$ or $\log(1+\beta)$ to the corresponding pairwise loss makes it nonnegative but does not change optimization for a fixed graph. The losses are uniformly bounded, i.e.,

$$\log\left(\frac{1}{1+\alpha}\right) \le \ell_{ij}^+ \le \log\left(\frac{1}{\alpha}\right), \qquad \log\left(\frac{1}{1+\beta}\right) \le \ell_{ij}^- \le \log\left(\frac{1}{\beta}\right). \tag{9}$$

Their derivatives are also uniformly bounded, i.e.,

$$-\frac{1}{\alpha} \le \frac{\partial \ell_{ij}^+}{\partial A_{ij}} \le \frac{1}{\alpha}, \qquad \frac{1}{1+\beta} \le \frac{\partial \ell_{ij}^-}{\partial A_{ij}} \le \frac{1}{\beta}. \tag{10}$$

Hence α controls the maximum attraction generated by a positive relation, while β controls the maximum repulsion generated by a negative relation. Because sparse graphs contain far more negative than positive pairs, we use $\alpha \ll \beta$. Our current default $\alpha = 10^{-2}$ and $\beta = 2n/K$ follows the heuristic scaling $\beta \propto |\Omega^-|/|\Omega^+|$. The factor 2 is empirical rather than theoretically optimal.

Theorem 3.1 (Bounded influence of graph errors)**.** *Let $\mathbf{S}^\star, \mathbf{S} \in [0,1]^{n\times n}$ be ideal and observed symmetric affinity matrices, respectively, and assume $0 < \alpha \le \beta$. Let $E = \{(i,j) : i < j, S_{ij} \ne S_{ij}^\star\}$. For any fixed embedding $\mathbf{Z}$,*

$$|\mathcal{L}_{\text{RV}}(\mathbf{Z};\mathbf{S}) - \mathcal{L}_{\text{RV}}(\mathbf{Z};\mathbf{S}^\star)| \le \frac{|E|}{N}\log\left(\frac{1+\beta}{\alpha}\right), \tag{11}$$

$$\|\nabla_{\mathbf{Z}}\mathcal{L}_{\text{RV}}(\mathbf{Z};\mathbf{S}) - \nabla_{\mathbf{Z}}\mathcal{L}_{\text{RV}}(\mathbf{Z};\mathbf{S}^\star)\|_F \le \frac{|E|}{N} C_{\text{grad}} G_{\max}(\mathbf{Z}), \tag{12}$$

where

$$C_{\text{grad}} = \max\left\{\frac{1}{\alpha} + \frac{1}{1+\alpha}, \frac{1}{\alpha} + \frac{1}{\beta}\right\}, \qquad G_{\max}(\mathbf{Z}) = \max_{i<j} \|\nabla_{\mathbf{Z}} A_{ij}\|_F.$$

For the affinity in equation 2,

$$\|\nabla_{\mathbf{Z}} A_{ij}\|_F = \sqrt{2}(l+1)\sigma^{-2} A_{ij}^{(l+3)/(l+1)} \|\mathbf{z}_i - \mathbf{z}_j\|_2 \le \frac{\sqrt{2}(l+1)}{\sigma}\frac{(l+2)^{(l+2)/2}}{(l+3)^{(l+3)/2}},$$

and consequently

$$G_{\max}(\mathbf{Z}) \le \frac{\sqrt{2}(l+1)}{\sigma}\frac{(l+2)^{(l+2)/2}}{(l+3)^{(l+3)/2}}.$$

The theorem controls the sensitivity of the objective and its gradient at a fixed embedding; it does not by itself guarantee recovery of the ideal embedding. Its key contrast with cross-entropy is uniformity: no analogous finite bound exists for (3) when a false-positive edge has $A_{ij} \to 0$ or a false-negative non-edge has $A_{ij} \to 1$. The proof is in Appendix M.1.

Theorem 3.2 (Affinity-level fidelity). *For a target affinity $w \in [0,1]$, let $\ell_w^+(A) = -w\log(A+\alpha) - (1-w)\log(1-A+\alpha)$, $A \in [0,1]$. Then ℓ_w^+ is strictly convex with unique minimizer $A_w^\star = \Pi_{[0,1]}(w + (2w-1)\alpha)$, $|A_w^\star - w| \le \alpha$. Moreover, the negative-pair loss $\ell^-(A) = -\log(1-A+\beta)$ is uniquely minimized at $A_-^\star = 0$. In particular, a binary positive relation has $A_1^\star = 1$. Hence, the robust loss preserves the binary pairwise targets exactly and perturbs a weighted positive target by at most α.*

The theorem shows that bounding the influence of individual relations does not substantially alter their preferred affinities. As all affinities are jointly induced by the same $\mathbf{Z}$, however, the pairwise optima need not be simultaneously realizable. Therefore, Theorem 3.2 characterizes the pairwise calibration of the loss rather than the global optimum over $\mathbf{Z}$. The numerical results in Appendix F show the superiority of the proposed robust loss over cross-entropy.

3.4 Importance-weighted negative-pair sampling

Each iteration in minimizing $\mathcal{L}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S})$ requires evaluating all N pairs and hence the complexity is $\mathcal{O}(n^2)$, which is too costly for large datasets. Let $\Omega_s^- \subseteq \Omega^-$ be sampled uniformly without replacement, we use the stochastic objective

$$\widehat{\mathcal{L}}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S}) = \frac{1}{N}\left[\sum_{(i,j)\in\Omega^+} \ell_{ij}^+ + \frac{|\Omega^-|}{|\Omega_s^-|}\sum_{(i,j)\in\Omega_s^-} \ell_{ij}^-\right]. \tag{13}$$

Proposition 3.3 (Accuracy of the sampled objective). *For any fixed embedding $\mathbf{Z}$,*

$$\mathbb{E}_{\Omega_s^-}\left[\widehat{\mathcal{L}}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S})\right] = \mathcal{L}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S}). \tag{14}$$

Moreover, whenever the pairwise losses are differentiable with respect to $\mathbf{Z}$,

$$\mathbb{E}_{\Omega_s^-}\left[\nabla_{\mathbf{Z}}\widehat{\mathcal{L}}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S})\right] = \nabla_{\mathbf{Z}}\mathcal{L}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S}). \tag{15}$$

Furthermore, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\widehat{\mathcal{L}}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S}) - \mathcal{L}_{\mathrm{RV}}(\mathbf{Z};\mathbf{S})\right| \le \frac{2|\Omega^-|}{n(n-1)}\log\left(\frac{1+\beta}{\beta}\right)\sqrt{\frac{\log(2/\delta)}{2|\Omega_s^-|}}. \tag{16}$$

Thus, the sampling preserves both the full objective and its gradient in expectation. For fixed δ, the sampling error decreases as $\mathcal{O}(|\Omega_s^-|^{-1/2})$. This sampled objective differs from UMAP-style negative sampling, which changes the effective attraction–repulsion weighting. If $|\Omega_s^-| = \mathcal{O}(|\Omega^+|)$ and the input graph is sparse with $|\Omega^+| = \mathcal{O}(nK)$, then each stochastic iteration requires $\mathcal{O}(nK)$ pairwise evaluations. In our experiments, we set $|\Omega_s^-| = |\Omega^+|$. In Appendix F, the numerical experiments show that $|\Omega_s^-|$ has little impact on the visualization performance.

For convenience, we call the proposed method *Robust Visualization* (RobustVis). Due to space limitations, we defer the implementation details to Appendix A.

3.5 Confidence-Weighted Extensions of RobustVis

Observed positive edges need not be equally reliable. We therefore introduce a confidence-weighted extension that retains the sparse binary graph support while allowing different positive edges to contribute different amounts of attraction.

Let $B_{ij} = \mathbb{1}\{S_{ij} > 0\}$ denote the binary support of the affinity graph, and define

$$\Omega_B^+ = \{(i,j) : i < j,\ B_{ij} = 1\}, \qquad \Omega_B^- = \{(i,j) : i < j,\ B_{ij} = 0\}. \tag{17}$$

The support matrix $\mathbf{B}$ determines which pairs are treated as positive and negative. For each positive pair $(i,j) \in \Omega_B^+$, we additionally compute a confidence score $c_{ij} > 0$. We normalize these scores over the positive edge set:

$$w_{ij} = \frac{c_{ij}}{\frac{1}{|\Omega_B^+|}\sum_{(u,v)\in\Omega_B^+} c_{uv}}, \qquad (i,j) \in \Omega_B^+. \tag{18}$$

Consequently, $\frac{1}{|\Omega_B^+|}\sum_{(i,j)\in\Omega_B^+} w_{ij} = 1$. This normalization preserves the average positive-pair contribution and therefore changes primarily the relative allocation of attraction among positive edges, rather than the overall attraction–repulsion balance.

The resulting confidence-weighted RobustVis objective is

$$\mathcal{L}_{\mathrm{CW}}(\mathbf{Z};\mathbf{B},\mathbf{W}) = \frac{1}{N}\left[-\sum_{(i,j)\in\Omega_B^+} w_{ij}\log(A_{ij}+\alpha) - \sum_{(i,j)\in\Omega_B^-}\log(1-A_{ij}+\beta)\right]. \tag{19}$$

Only the attractive loss of an observed positive pair is multiplied by its confidence. Every unobserved pair retains the same unit-weight negative loss, together with the importance correction when negative-pair sampling is used.

This construction differs from directly substituting a fractional affinity into the Bernoulli loss in equation 7. In the latter case, an edge with affinity $S_{ij} \in (0,1)$ contributes both an attractive term $-S_{ij}\log(A_{ij}+\alpha)$ and a repulsive term $-(1-S_{ij})\log(1-A_{ij}+\alpha)$. In equation 19, by contrast, every pair in Ω_B^+ remains a positive relation: its weight controls only how strongly the pair is attracted. This separation is appropriate when the graph support determines the type of the relation, while the soft weight represents confidence in that relation rather than a Bernoulli target probability.

We consider two concrete confidence constructions: RobustVis-UW and RobustVis-BR. They differ only in how the positive-edge weights are computed. The details of RobustVis-UW and RobustVis-BR are deferred to Appendix B.

4 Experiments

4.1 Baselines, Datasets, and Evaluation Metrics

Baselines We compare our RobustVis and its variants RobustVis-UW and RobustVis-BR with 10 baselines, including MDS (Shepard, 1962), Isomap (Tenenbaum et al., 2000), t-SNE (Van der Maaten & Hinton, 2008), TriMAP (Amid & Warmuth, 2019), UMAP (McInnes et al., 2018), PHATE (Moon et al., 2019), PaCMAP (Wang et al., 2021), UMATO (Jeon et al., 2025), FMDS (Dagès et al., 2025), and RobustVis w/o $\mathcal{L}_{\mathrm{RV}}$. Note that RobustVis w/o $\mathcal{L}_{\mathrm{RV}}$ is based on KNN graph and the loss function of UMAP, i.e., equation 5. Hence, it interpolates between UMAP and our RobustVis, and will verify that the gain of RobustVis is from the robust loss, rather than the binary graph.

Datasets We consider 3 synthetic datasets, including Swiss Roll, Swiss Roll with Hole, and Mammoth (Tenenbaum et al., 2000; Coenen & Pearce, 2019; Wang et al., 2021), and 13 real datasets. The real datasets include 4 image datasets COIL-20, COIL-100 (Nene et al., 1996), MNIST (LeCun et al., 1998), and Fashion-MNIST (Xiao et al., 2017), 6 tabular datasets HAR (Anguita et al., 2013), IoT (Sharmila & Nagapadma, 2023), Satimage, Texture (Brodatz, 1966), Gesture Phase (Madeo et al., 2016), Japanese Vowels (Kudo et al., 1999), and 3 single-cell RNA-sequencing datasets Allergic Memory, Glial Atlas, and Crohn's disease (Ordovas-Montanes et al., 2018; Jager et al., 2022; Zheng et al., 2021). Together, these datasets cover different sample sizes, ambient dimensions, numbers of classes, class-imbalance patterns, and noise characteristics. Dataset-specific preprocessing and statistics are provided in Appendix D.

Evaluation Metrics In addition to 2D visualizations, for synthetic datasets, we report Trustworthiness, Continuity (Venna & Kaski, 2001; Lee & Verleysen, 2009), Procrustes error, distance Stress, and Spearman correlation of pairwise distances, the definitions of which are in Appendix C. For real datasets, since there are no ground-truth low-dimensional coordinates, geometric recovery errors such as Procrustes error and stress cannot be computed, we therefore, in addition to Trustworthiness

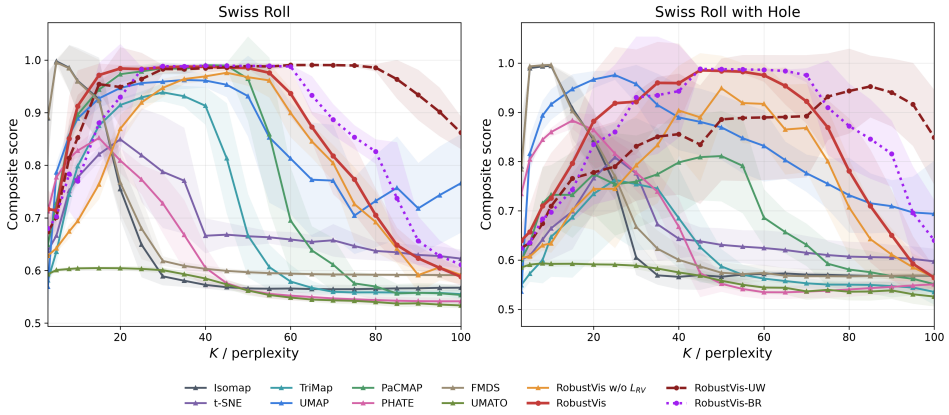


Figure 1: Neighborhood-size sensitivity on Swiss Roll (left) and Swiss Roll with Hole (right), with K or t-SNE perplexity varied from 3 to 100. Curves and bands show the mean and one standard deviation over 10 runs. The composite score averages Trustworthiness, Continuity, $1-\text{Procrustes error}$, $1-\text{Stress}$, and Spearman correlation (higher is better).

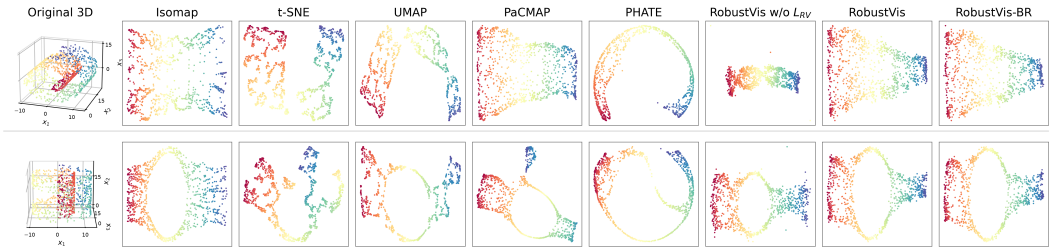


Figure 2: Embeddings of Swiss Roll (first row) and Swiss Roll with Hole (second row), with the original 3D data. Parameters are selected by the mean composite score over ten runs; displayed embeddings use seed 0. Settings for the two datasets, respectively, are Isomap $K = (5, 10)$, t-SNE perplexity $(20, 25)$, UMAP $K = (35, 25)$, PaCMAP $K = (40, 50)$, PHATE $K = (15, 15)$, RobustVis without $\mathcal{L}_{\text{RV}}$ $K = (45, 50)$, RobustVis $K = (35, 45)$, and RobustVis-BR $K = (45, 55)$. Note that the sampling density at one end of the Swiss roll is three times that at the other end. RobustVis and RobustVis-BR preserve this density variation, producing an embedding in which the denser end is approximately three times wider than the sparser end.

and Continuity, report the classification accuracy of logistic regression, and the normalized mutual information (NMI) and adjusted Rand index (ARI) of k-means clustering. Note that high Trustworthiness and Continuity do not guarantee correct global unfolding, relative cluster placement, or preservation of long-range transitions.

4.2 Results on Synthetic Datasets

Figure 1 reports neighborhood-size sensitivity on Swiss Roll and Swiss Roll with Hole. Isomap and FMDS perform well at small K but deteriorate as K increases, whereas UMAP, PaCMAP, and TriMap remain competitive over a broader range. RobustVis becomes stable at moderate K; RobustVis-UW and RobustVis-BR further reduce sensitivity, with RobustVis-BR generally providing the most reliable performance. Metric-wise results are provided in Appendix E.1.

Figure 2 compares embeddings at selected parameters. Despite high Trustworthiness and Continuity, t-SNE fragments the manifold, while RobustVis without $\mathcal{L}_{\text{RV}}$ strongly compresses the Swiss Roll. In contrast, RobustVis and RobustVis-BR recover more coherent global ordering and preserve the hole more clearly, illustrating that local neighborhood preservation alone does not guarantee faithful global recovery. Additional visualizations, including Mammoth, are provided in Appendix I.

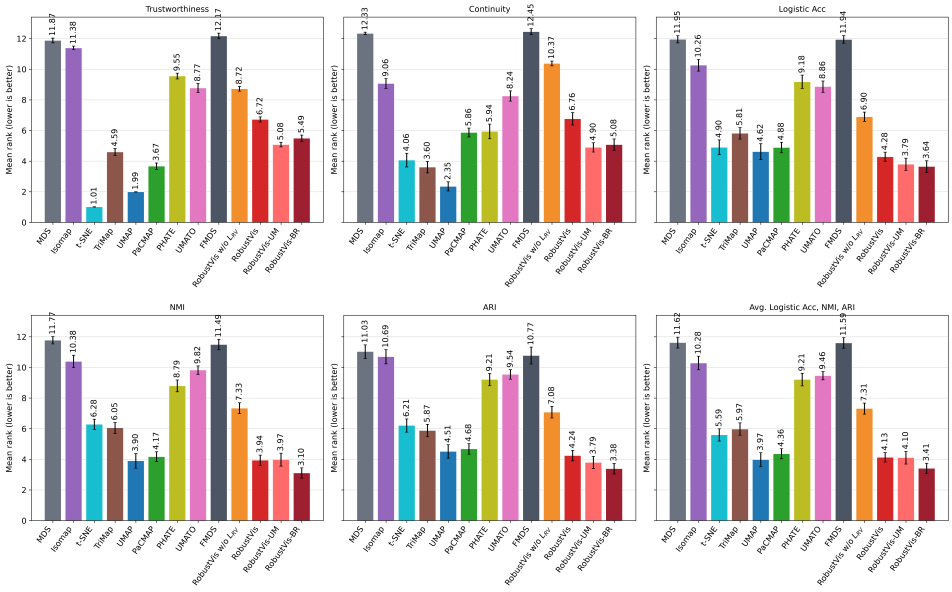


Figure 3: Average within-dataset ranks over 13 real datasets and 3 neighborhood-scale settings.

4.3 Results on Real Datasets

Figure 3 compares all methods using their average within-dataset ranks over 13 real-world datasets and three neighborhood-scale settings. Specifically, we use $K \in \{15, 30, 60\}$ for most datasets, $K \in \{5, 10, 20\}$ for COIL-20 and COIL-100, and perplexity $\in \{30, 60, 120\}$ for t-SNE. The ranking is computed separately for each dataset and parameter setting and then averaged over the resulting 39 dataset–setting combinations. Results for the three parameter settings separately are provided in Appendix H. The detailed values of the five metrics on the 13 datasets are in Tables 2–14 of the appendix. RobustVis-BR achieves the best overall downstream-task performance, with an average rank of 3.41 over logistic accuracy, NMI, and ARI. It also obtains the best individual ranks for logistic accuracy, NMI, and ARI. t-SNE leads in Trustworthiness, while UMAP leads in Continuity.

Figure 4 illustrates the different geometric behavior of the methods. On COIL-20, the curved structures correspond to continuous viewpoint changes of individual objects, and RobustVis-BR preserves these within-object trajectories while separating most object identities. On Fashion-MNIST, UMAP and PaCMAP create more isolated islands, whereas RobustVis retains broad transitions among visually related clothing categories. On HAR, the RobustVis embeddings form coherent regions while preserving proximity between related activities. On Crohn, the proposed objective recovers stable major cell-type compartments together with smaller branches and intermediate regions. We do not assess an embedding only by the amount of empty space between colored clusters; intermediate samples are also checked against their neighborhoods in the original feature space.

On these datasets, the performance of all graph-based objectives is partly limited by the quality of the constructed similarity graph. To better isolate the effect of the optimization objective from graph construction, we further evaluate RobustVis on naturally occurring graphs in Section 4.4.

4.4 Quantitative Evaluation of Graph Visualization

We evaluate RobustVis on five naturally occurring graphs: Cora, CiteSeer, and PubMed (Yang et al., 2016), Chameleon (Pei et al., 2020), and Amazon Photo (Shchur et al., 2018). The observed graph is used directly, without constructing a kNN similarity graph. All methods use graph structure only; methods requiring node attributes are given standardized node degree as input. Figure 5 reports the average within-dataset ranks and mean running times over the five graphs. RobustVis achieves the best ranks for logistic accuracy, ARI, and edge AUC, and obtains the best combined rank of 1.30, compared with 2.90 for node2vec and 3.10 for RobustVis without $\mathcal{L}_{\mathrm{RV}}$. Moreover, RobustVis is much more efficient than other methods.

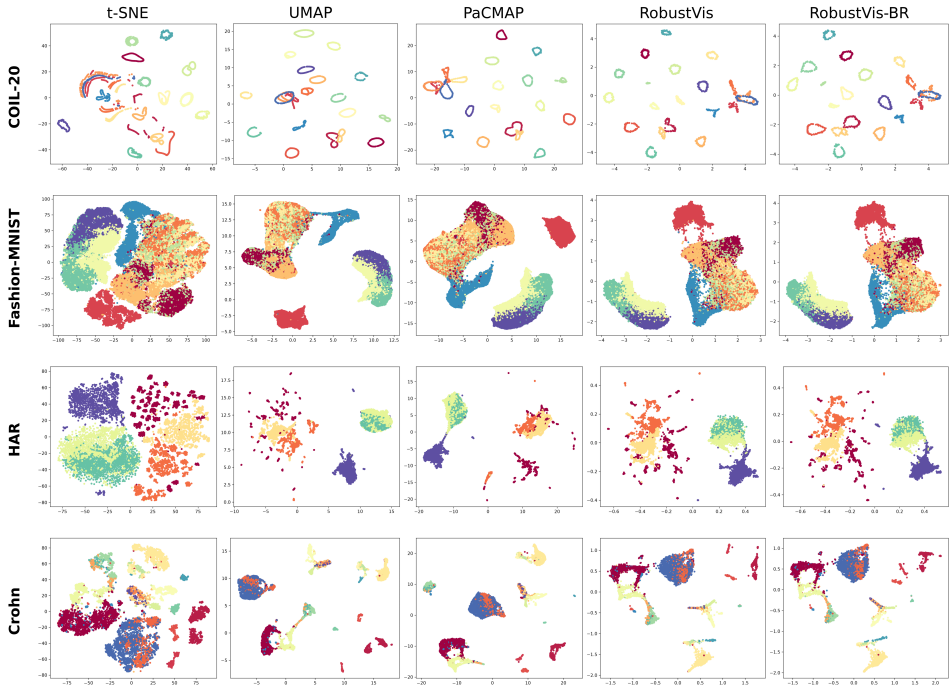


Figure 4: Two-dimensional embeddings on four representative real-world datasets. UMAP and PaCMAP often produce larger empty gaps between labeled groups, whereas RobustVis and RobustVis-BR retain more continuous transitions between related or ambiguous samples. Such transitions are not necessarily projection errors, because class labels need not define disconnected manifolds in the original feature space.

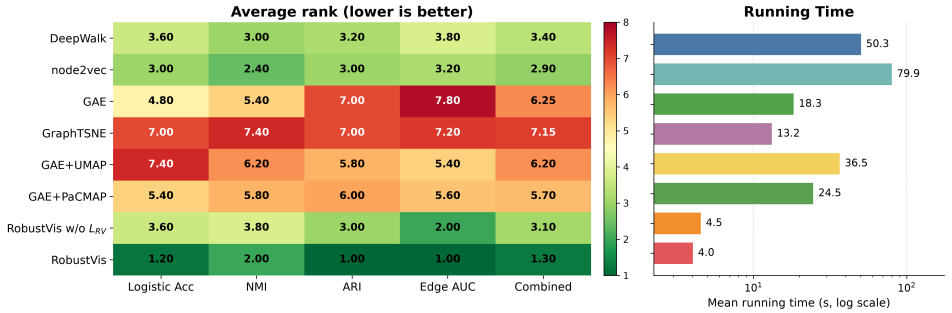


Figure 5: Average ranks and running times on five graph datasets using graph structure only.

4.5 More Results

Due to space limitations, we report the hyperparameter sensitivity, robustness to graph errors, computational cost, additional real-data visualization, the analysis of statistical significance, and the visualization results of graph datasets in Appendices E, F, G, I, J, and K, respectively.

5 Conclusion

We introduced RobustVis for visualization under imperfect neighborhood graphs. Its asymmetric objective uniformly bounds pairwise losses and affinity derivatives, providing objective- and gradient-perturbation guarantees under graph corruption. Importance-weighted negative sampling enables practical optimization, while confidence-weighted extensions account for variation in positive-edge reliability. Experiments on synthetic and real-world datasets demonstrate competitive geometric recovery and downstream performance, together with improved robustness to false-positive and false-negative graph perturbations. Future work will investigate more reliable and locally adaptive graph-construction methods and their integration with the robust objective.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language. We have reviewed all AI-assisted work.

REFERENCES

Ehsan Amid and Manfred K. Warmuth. TriMap: Large-scale Dimensionality Reduction Using Triplets. *arXiv preprint arXiv:1910.00204*, 2019.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pp. 3–4, 2013.

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

Noah Bergam, Szymon Snoeck, and Nakul Verma. t-sne exaggerates clusters, provably. *arXiv preprint arXiv:2510.07746*, 2025.

Phil Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover Publications, New York, 1966. ISBN 978-0-486-21669-0.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, 2023.

Andy Coenen and Adam Pearce. Understanding umap, 2019. URL `https://pair-code.github.io/understanding-umap/`.

Thomas Dagès, Simon Weber, Ya-Wei Eileen Lin, Ronen Talmon, Daniel Cremers, Michael Lindenbaum, Alfred M Bruckstein, and Ron Kimmel. Finsler multi-dimensional scaling: Manifold learning for asymmetric dimensionality reduction and embedding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25842–25853, 2025.

Ayush Dalmia and Suzanna Sia. Clustering with umap: Why and how connectivity matters. *arXiv preprint arXiv:2108.05525*, 2021.

Sebastian Damrich and Fred A Hamprecht. On umap's true loss function. *Advances in Neural Information Processing Systems*, 34:5798–5809, 2021.

David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

Jicong Fan, Tommy WS Chow, Mingbo Zhao, and John KL Ho. Nonlinear dimensionality reduction for data with disconnected neighborhood graph. *Neural Processing Letters*, 47(2):697–716, 2018.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016. doi: 10.1145/2939672.2939754.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Daniel Jiwoong Im, Nakul Verma, and Kristin Branson. Stochastic neighbor embedding under f-divergences. *arXiv preprint arXiv:1811.01247*, 2018.

Sara Elgaard Jager, Lone Tjener Pallesen, Lin Lin, Francesca Izzi, Alana Miranda Pinheiro, Sara Villa-Hernandez, Paolo Cesare, Christian Bjerggaard Vaegter, and Franziska Denk. Comparative transcriptional analysis of satellite glial cell injury response. *Wellcome open research*, 7:156, 2022.

Hyeon Jeon, Kwon Ko, Soohyun Lee, Jake Hyun, Taehyun Yang, Gyehun Go, Jaemin Jo, and Jinwook Seo. Umato: Bridging local and global structures for reliable visual analytics with dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Mineichi Kudo, Jun Toyama, and Masaru Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20(11–13):1103–1111, 1999. doi: 10.1016/S0167-8655(99)00077-X. URL https://doi.org/10.1016/S0167-8655(99)00077-X.

Noël Kury, Dmitry Kobak, and Sebastian Damrich. DREAMS: Preserving both local and global structure in dimensionality reduction. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=xpGu3Sichc. Expert Certification.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009. doi: 10.1016/j.neucom.2008.12.017.

Yao Yang Leow, Thomas Laurent, and Xavier Bresson. GraphTSNE: A visualization technique for graph-structured data. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Renata Cristina Barros Madeo, Sarajane Marques Peres, and Clodoaldo Aparecido de Moraes Lima. Gesture phase segmentation using support vector machines. *Expert Systems with Applications*, 56:100–115, 2016. doi: 10.1016/j.eswa.2016.02.021. URL https://doi.org/10.1016/j.eswa.2016.02.021.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Kevin R Moon, David Van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature biotechnology*, 39(6):765–774, 2021.

SA Nene, SK Nayar, and H Murase. Columbia university image library (coil-20). *Technical Report CUCS-005-96*, 1996.

Jose Ordovas-Montanes, Daniel F Dwyer, Sarah K Nyquist, Kathleen M Buchheit, Marko Vukovic, Chaarushena Deb, Marc H Wadsworth, Travis K Hughes, Samuel W Kazer, Eri Yoshimoto, et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature*, 560 (7720):649–654, 2018.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, 2014. doi: 10.1145/2623330.2623732.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Tim Sainburg, Leland McInnes, and Timothy Q Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

BS Sharmila and Rohini Nagapadma. Quantized autoencoder (qae) intrusion detection system for anomaly detection in resource-constrained iot devices using rt-iot2022 dataset. *Cybersecurity*, 6 (1):41, 2023.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *NeurIPS Workshop on Relational Representation Learning*, 2018.

Roger N Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

Kiran Smelser, Jacob Miller, and Stephen Kobourov. "normalized stress" is not normalized: How to interpret stress correctly. In *2024 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 41–50. IEEE, 2024.

Ashwin Srinivasan. Statlog (Landsat Satellite). UCI Machine Learning Repository, 1993. URL `https://doi.org/10.24432/C55887`.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pp. 287–297, 2016.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Laurens J. P. van der Maaten, Eric O. Postma, and H. Jaap van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University, Tilburg, The Netherlands, oct 2009.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks—ICANN 2001*, volume 2130 of *Lecture Notes in Computer Science*, pp. 485–491. Springer, 2001. doi: 10.1007/3-540-44668-0_68.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

Yingfan Wang, Yiyang Sun, Haiyang Huang, and Cynthia Rudin. Dimension reduction with locally adjusted graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21357–21365, 2025.

Koshi Watanabe, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Starmap: Global neighbor embedding for faithful data visualization. *arXiv preprint arXiv:2502.03776*, 2025.

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

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48. PMLR, 2016.

Zhirong Yang, Yuwei Chen, and Jukka Corander. T-sne is not optimized to reveal clusters in data. *arXiv preprint arXiv:2110.02573*, 2021.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing*, 26(1):313–338, 2004.

HB Zheng, BA Doran, K Kimler, A Yu, V Tkachev, V Niederlova, K Cribbin, R Fleming, B Bratrude, K Betz, et al. A treatment-naïve cellular atlas of pediatric crohn's disease predicts disease severity and therapeutic response. *medrxiv*, 2021.

A ALGORITHMIC IMPLEMENTATION

In RobustVis, we first construct a mutual K-NN graph, which can be either binary or weighted; other affinity graph construction methods can also be used. For synthetic data, where the ambient-space distances directly reflect the underlying geometry, we use Cauchy-weighted edges to retain local metric information. For real high-dimensional data, we use binary edges, as neighborhood membership is generally more reliable than the precise magnitude of ambient-space distances. This convention is fixed according to the data type rather than tuned for individual datasets, and its effect is further examined in the experiments.

The use of binary affinities for real high-dimensional data is also motivated by distance concentration. Under standard product distributions with independent coordinates and finite second moments, normalized Euclidean distances concentrate around a common value as the ambient dimension increases, reducing the relative separation between neighboring candidates. Moreover, for $1 \le p < q$, $\|\mathbf{x}\|_q \le \|\mathbf{x}\|_p \le d^{1/p-1/q}\|\mathbf{x}\|_q$, so the potential distortion between different ℓ_p metrics increases with dimension. Consequently, when the margin between the Kth and $(K+1)$th neighbors is small, even moderate metric perturbations can change the resulting neighborhood. This motivates retaining only neighborhood membership for high-dimensional real data, while distance-weighted affinities remain useful for low-dimensional synthetic data where the ambient metric directly represents the generating geometry.

The resulting graph may contain multiple components, especially when K is small. In such cases, the relative locations of the components cannot be reliably determined from the graph alone. We therefore perform a component-level graph repair. The distance between two components is defined as the minimum pairwise distance between their constituent points, and a minimum spanning tree (MST) is constructed over the components using these distances. For each component pair selected by the MST, we add $1 \le K' < K$ short and spatially diverse cross-component edges, preferring edges with distinct endpoints to provide stronger constraints on their relative positions and orientations. In this way, the MST provides a sparse connectivity backbone, while the multiple point-level repair edges help preserve the relative geometry between connected components. To avoid assigning excessive confidence to auxiliary repair edges, we weight them according to their original-space distances using a globally scaled Cauchy affinity: $S_{ij}^{\mathrm{add}} = \left(1 + \|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\tau^2\right)^{-1}$, where τ is the median distance from each sample to its Kth nearest neighbor. The common scale τ ensures that all repair edges are evaluated using the same reference: shorter inter-component connections receive larger affinities, whereas longer connections are assigned lower confidence.

Ablation studies of the two repair operations are reported in Figure 12, Figure 13, and Table 1. Without these operations, the locations of isolated points and the relative placement of disconnected components are underdetermined.

A suitable initialization of $\mathbf{Z}$ is important in practice. We initialize $\mathbf{Z}$ using PCA, which provides a deterministic low-dimensional configuration that preserves part of the global variance structure of the data. The scale parameter σ in (2) controls the spread of the low-dimensional affinities. Letting $c > 0$ be a constant, we set

$$\sigma = c \cdot \mathrm{mean}\left(\left\{\|\mathbf{z}_i - \mathbf{z}_j\|_2 : (i,j) \in \Omega^+\right\}\right). \tag{20}$$

We use a fixed value $c = 100$ across all datasets. Empirically, this initialization and scale calibration provide stable optimization across the datasets considered in this paper.

The objective in equation 6 is optimized by gradient descent. For the Cauchy kernel in equation 2, the derivative of $A_{i,j}$ with respect to $\mathbf{z}_i$ is

$$\frac{\partial A_{i,j}}{\partial \mathbf{z}_i} = -\frac{l+1}{\sigma^2}\left(1 + \frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\sigma^2}\right)^{-(l+3)/2}(\mathbf{z}_i - \mathbf{z}_j) = -(l+1)\sigma^{-2}A_{i,j}^{\frac{l+3}{l+1}}(\mathbf{z}_i - \mathbf{z}_j). \tag{21}$$

Therefore, the gradient of $\mathcal{L}(\mathbf{Z};\mathbf{S})$ with respect to $\mathbf{z}_i$ can be written as

$$\frac{\partial \mathcal{L}(\mathbf{Z};\mathbf{S})}{\partial \mathbf{z}_i} = -r(l+1)\sigma^{-2}\sum_{j\neq i} C_{i,j}A_{i,j}^{\frac{l+3}{l+1}}(\mathbf{z}_i - \mathbf{z}_j) \tag{22}$$

where $r = \frac{1}{|\Omega^+|+|\Omega^-|}$, $C_{i,j} = H_{i,j}$ if $(i,j) \in \Omega^+$ or $(j,i) \in \Omega^+$, $C_{i,j} = W_{i,j}$ if $(i,j) \in \Omega^-$ or $(j,i) \in \Omega^-$, $H_{i,j} = -\frac{S_{i,j}}{A_{i,j}+\alpha} + \frac{1-S_{i,j}}{1-A_{i,j}+\alpha}$, and $W_{i,j} = \frac{1}{1-A_{i,j}+\beta}$. We use Adam (Kingma & Ba,

2014) because it is effective and robust in this setting. In all experiments, we set the learning rate to σ, which provides stable optimization in practice.

The overall procedure is summarized in Algorithm 1.

Algorithm 1 RobustVis

Input: Data matrix $\mathbf{X}$; embedding dimension l (default: 2); neighborhood size K (default: 30); robustness parameters α (default: 10^{-2}) and β (default: $2n/K$); maximum number of iterations maxiter (default: 200); initialization strategy init_Z (default: PCA).

Construct a mutual KNN graph (binary or weighted) and get $\mathbf{S}$.
Connect disconnected components of $\mathbf{S}$ (if necessary).
if init_Z is PCA **then**
  Initialize $\mathbf{Z} = \text{PCA}(\mathbf{X}; l)$.
  Compute σ using (20) with $c = 100$.
else
  Sample $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_l)$ for all $i \in [n]$.
  Compute σ using (20) with $c = 0.1$.
end if
Set the optimizer to Adam with learning rate σ.
for $t = 1, \ldots,$ maxiter **do**
  Compute the low-dimensional similarities $\mathbf{A}$ using (2).
  Compute $\mathcal{L}_{\text{RV}}(\mathbf{Z}; \mathbf{S})$ using (6) or $\widehat{\mathcal{L}}_{\text{RV}}(\mathbf{Z}; \mathbf{S})$ using (13).
  Update $\mathbf{Z}$ using Adam.
end for

Output: Low-dimensional embedding $\mathbf{Z}$.

B RobustVis-UW and RobustVis-BR

RobustVis-UW uses locally scaled fuzzy memberships to quantify positive edge confidence. Let d_{ij} be the input-space distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Following the graph construction of UMAP (McInnes et al., 2018), a directed membership strength is defined as

$$\mu_{j|i} = \begin{cases} \exp\left(-\dfrac{\max(0, d_{ij} - \rho_i)}{\sigma_i}\right), & j \in \mathcal{N}_K(i), \\ 0, & j \notin \mathcal{N}_K(i), \end{cases} \tag{23}$$

where ρ_i represents a local connectivity radius and σ_i is a sample-specific scale determined from the distances in $\mathcal{N}_K(i)$. The two directed memberships are combined using the fuzzy union

$$\mu_{ij} = \mu_{j|i} + \mu_{i|j} - \mu_{j|i}\mu_{i|j}. \tag{24}$$

We use these memberships only to weight edges already present in the RobustVis support:

$$c_{ij}^{\text{UW}} = \mu_{ij}, \qquad (i, j) \in \Omega_B^+, \tag{25}$$

followed by the normalization in equation 18. Auxiliary graph-repair edges that do not appear in the UMAP fuzzy graph retain unit raw weight. No additional UMAP edge is introduced into $\mathbf{B}$. Therefore, RobustVis-UW and the original RobustVis use the same positive and negative pair sets; RobustVis-UW merely assigns stronger attraction to locally well-supported positive relations and weaker attraction to less certain ones.

The local scales ρ_i and σ_i allow the confidence to adapt to heterogeneous sampling density. However, the resulting weights depend on the magnitudes of the input-space distances and may consequently be affected by distance concentration or metric misspecification in very high dimensions.

RobustVis-BR avoids using distance magnitudes and derives confidence only from neighbor ranks. Let $r_{i\to j} \in \{1, \ldots, K\}$ denote the rank of $\mathbf{x}_j$ among the K nearest neighbors of $\mathbf{x}_i$, with smaller values indicating stronger local evidence. For an edge $(i, j) \in \Omega_B^+$, we define

$$c_{ij}^{\text{BR}} = \frac{1}{2}\left[\exp\left(-\frac{r_{i\to j} - 1}{\tau}\right) + \exp\left(-\frac{r_{j\to i} - 1}{\tau}\right)\right], \tag{26}$$

where $\tau > 0$ controls the rate of rank decay. We use

$$\tau = \frac{K}{2} \tag{27}$$

throughout the experiments. If a directed rank is unavailable for an auxiliary graph-repair edge, it is set to K. The raw scores in equation 26 are subsequently normalized using equation 18.

The bidirectional construction assigns the largest confidence to pairs that are highly ranked from both directions. A pair receives a smaller weight when one or both endpoints regard the other as a boundary neighbor. In contrast to distance-based affinities, the rank weights are invariant to monotone transformations of the input distances and are less directly affected by differences in local distance scale. RobustVis-BR is therefore particularly suited to high-dimensional data for which relative neighbor order may be more reliable than the numerical distance values.

Both RobustVis-UW and RobustVis-BR retain the sparse graph structure, negative-pair sampling procedure, low-dimensional affinity, and robust positive- and negative-pair losses of the original RobustVis. Their additional cost is limited to computing the positive-edge confidence scores. The two variants therefore isolate the effect of positive-edge reliability from changes in graph support or optimization.

C Definitions of Quantitative Evaluation Metrics

For the synthetic datasets considered in Section 4, the underlying low-dimensional coordinates are known from the data-generation process. This makes it possible to quantitatively evaluate how well a visualization method recovers the ground-truth manifold unfolding.

Let

$$D = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$$

be a synthetic dataset, and suppose that each sample $\mathbf{x}_i$ is associated with a ground-truth latent coordinate

$$\mathbf{y}_i \in \mathbb{R}^2.$$

Write

$$\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\} \subset \mathbb{R}^2$$

for the collection of latent coordinates. Let

$$\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}, \qquad \mathbf{z}_i \in \mathbb{R}^2,$$

denote the two-dimensional embedding produced by a visualization method.

Since such an embedding is only identifiable up to translation, rotation, reflection, and global scaling, the evaluation criteria should ideally be invariant to these transformations. Therefore, we calculate the following three metrics.

Normalized Procrustes Error (NPE). Our first metric measures the discrepancy between the learned embedding and the ground-truth latent coordinates after optimal similarity alignment. Let

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i, \qquad \bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$$

be the sample means. We define the normalized Procrustes error by

$$\mathrm{NPE}(\mathbf{Z}, \mathbf{Y}) = \min_{s>0,\, \mathbf{R}\in O(2)} \frac{\sum_{i=1}^{n} \|s\mathbf{R}(\mathbf{z}_i - \bar{\mathbf{z}}) - (\mathbf{y}_i - \bar{\mathbf{y}})\|_2^2}{\sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|_2^2},$$

where $O(2)$ denotes the set of 2×2 orthogonal matrices. The orthogonal transformation $\mathbf{R}$ accounts for rotation and reflection, while the scalar s accounts for global rescaling. The range of NPE is $[0, 1]$. A smaller value of $\mathrm{NPE}(\mathbf{Z}, \mathbf{Y})$ indicates that the learned embedding more accurately matches the true unfolded manifold.

Normalized Distance Stress (NDS). The second metric, motivated by (Kruskal, 1964), evaluates how well pairwise distances in the embedding agree with those in the ground-truth latent space. Define

$$d_{ij}^{(Z)} = \|\mathbf{z}_i - \mathbf{z}_j\|_2, \qquad d_{ij}^{(Y)} = \|\mathbf{y}_i - \mathbf{y}_j\|_2, \qquad 1 \le i < j \le n.$$

Smelser et al. (2024) defined the normalized distance stress by

$$\mathrm{NDS}(\mathbf{Z}, \mathbf{Y}) = \min_{s>0} \frac{\sum_{1\le i<j\le n} \left(s d_{ij}^{(Z)} - d_{ij}^{(Y)}\right)^2}{\sum_{1\le i<j\le n} \left(d_{ij}^{(Y)}\right)^2}.$$

Here, the scale factor s again removes the effect of a global rescaling. This metric is less strict than Procrustes error, since it does not require pointwise alignment of the two configurations; instead, it measures how well the intrinsic metric structure of the latent manifold is preserved. The range of NDS is $[0, 1]$. Smaller values indicate better performance.

Spearman Correlation of Pairwise Distances (SCP). Our third metric is rank-based. It evaluates whether the relative ordering of pairwise distances is preserved by the visualization. Specifically, let

$$\mathcal{D}_Z = \{d_{ij}^{(Z)} : 1 \le i < j \le n\}, \qquad \mathcal{D}_Y = \{d_{ij}^{(Y)} : 1 \le i < j \le n\}.$$

We define

$$\mathrm{SCP}(\mathbf{Z}, \mathbf{Y}) = \mathrm{Spearman}(\mathcal{D}_Z, \mathcal{D}_Y),$$

namely, the Spearman rank correlation between the pairwise distances in the learned embedding and those in the ground-truth latent space. Unlike the previous two metrics, this criterion depends only on the ordering of distances and is therefore insensitive to monotone distortions of the distance scale. The range of SCP is $[-1, 1]$. Larger values of SCP indicate better preservation of the ordinal geometry.

These three metrics capture complementary aspects of visualization quality on synthetic data. The normalized Procrustes error measures exact recovery of the ground-truth unfolding after removing trivial similarity transformations. The normalized distance stress measures preservation of latent pairwise distances. The Spearman correlation focuses on ordinal geometric fidelity. Reporting all three metrics provides a more complete quantitative assessment of whether a nonlinear dimensionality reduction method faithfully recovers the true manifold structure.

Trustworthiness and Continuity. Trustworthiness and Continuity evaluate complementary aspects of local neighborhood preservation (Venna & Kaski, 2001; Lee & Verleysen, 2009). Consider an evaluation set of m samples, reindexed as $1, \dots, m$. Let $r_{ij}^{(X)}$ and $r_{ij}^{(Z)}$ denote the ranks of sample j among the other samples ordered by increasing Euclidean distance from sample i in the input and embedding spaces, respectively. Self-neighbors are excluded, ranks start at one, and distance ties are resolved consistently. Denote the corresponding k-nearest-neighbor sets by $\mathcal{N}_k^{(X)}(i)$ and $\mathcal{N}_k^{(Z)}(i)$, and define

$$\mathcal{U}_k(i) = \mathcal{N}_k^{(Z)}(i) \setminus \mathcal{N}_k^{(X)}(i), \qquad \mathcal{V}_k(i) = \mathcal{N}_k^{(X)}(i) \setminus \mathcal{N}_k^{(Z)}(i).$$

Here, $\mathcal{U}_k(i)$ contains neighbors introduced by the embedding, whereas $\mathcal{V}_k(i)$ contains input neighbors omitted from the embedding neighborhood. For $1 \le k < m/2$, the two scores are

$$\mathrm{Trustworthiness}(k) = 1 - \frac{2}{mk(2m - 3k - 1)} \sum_{i=1}^{m} \sum_{j \in \mathcal{U}_k(i)} \left(r_{ij}^{(X)} - k\right), \tag{28}$$

$$\mathrm{Continuity}(k) = 1 - \frac{2}{mk(2m - 3k - 1)} \sum_{i=1}^{m} \sum_{j \in \mathcal{V}_k(i)} \left(r_{ij}^{(Z)} - k\right). \tag{29}$$

Both scores lie in $[0, 1]$, with larger values indicating better preservation and a value of one indicating identical k-nearest-neighbor sets in the two spaces.

We use $k = 5$, independently of the neighborhood size K used to construct an embedding. For datasets with more than 10,000 samples, we evaluate both scores on the same fixed random subset of 10,000 samples across methods. Neighborhoods and ranks are computed within this subset, and m denotes its size.

These scores are rank-weighted diagnostics rather than neighborhood-overlap proportions: replacing an input neighbor with a sample ranked just beyond k incurs only a small penalty. Their reference is the observed input geometry, not the unknown latent structure. Consequently, their interpretation depends on the input representation and distance, whose neighborhoods may be affected by noise, irrelevant features, or metric misspecification. Moreover, a fixed k probes only one local scale, whose spatial extent varies with sampling density. High scores therefore do not establish correct global unfolding, long-range distance preservation, or relative cluster placement. Conversely, lower scores alone do not demonstrate better latent recovery; this requires additional evidence, such as the ground-truth-based metrics used for synthetic data.

Logistic-regression Accuracy. We use logistic-regression accuracy as a linear-probe measure of how well class information is separated in the two-dimensional embedding. Let $\mathbf{z}_i \in \mathbb{R}^2$ denote the embedding of sample i and $y_i \in 1, \ldots, C$ its ground-truth class label. For multiclass data, logistic regression models

$$P(y_i = c \mid \mathbf{z}_i) = \frac{\exp(\mathbf{w}_c^\top \mathbf{z}_i + b_c)}{\sum_{r=1}^{C} \exp(\mathbf{w}_r^\top \mathbf{z}_i + b_r)}, \qquad c = 1, \ldots, C.$$

The classifier is fitted by minimizing the regularized cross-entropy

$$-\frac{1}{m} \sum_{i=1}^{m} \log P(y_i \mid \mathbf{z}_i) + \lambda \sum_{c=1}^{C} \|\mathbf{w}_c\|_2^2,$$

where m is the number of training samples and λ controls ℓ_2 regularization. For binary data, this reduces to standard binary logistic regression.

For each embedding, we perform five stratified random 80/20 train–test splits. The regularization parameter λ is selected independently within each training split using stratified cross-validation, without access to the corresponding test labels. Predictions are made by

$$\widehat{y}_i = \arg\max_c P(y_i = c \mid \mathbf{z}_i),$$

and the test accuracy is

$$\mathrm{Acc} = \frac{1}{n_{\mathrm{test}}} \sum_{i \in \mathcal{I}_{\mathrm{test}}} \mathbb{1}\{\widehat{y}_i = y_i\}.$$

We report the accuracy averaged over the five test splits.

This metric measures *linear class separability* of the learned two-dimensional representation: a high value indicates that the class labels can be recovered by simple linear decision boundaries in the embedding. It does not directly measure geometric fidelity, and an embedding can achieve high logistic-regression accuracy while distorting within-class geometry or inter-class distances. Moreover, because the unsupervised embedding is computed once using the complete dataset before the labeled train–test splits are formed, the score should be interpreted as a transductive linear-probe diagnostic rather than an estimate of inductive generalization to previously unseen samples. The labels are not used to construct or optimize the embedding.

NMI and ARI. For clustering evaluation, we apply K-means to the embedding, setting the number of clusters to the number of ground-truth classes and using 50 random initializations. Let $\mathcal{C} = \{C_1, \ldots, C_q\}$ and $\widehat{\mathcal{C}} = \{\widehat{C}_1, \ldots, \widehat{C}_q\}$ denote the ground-truth and predicted partitions of the n samples. Define the contingency counts and their marginals by

$$n_{ab} = |C_a \cap \widehat{C}_b|, \qquad n_a = \sum_b n_{ab}, \qquad \widehat{n}_b = \sum_a n_{ab}.$$

Normalized mutual information (NMI) measures the information shared by the two partitions (**?**). Their mutual information is

$$I(\mathcal{C};\widehat{\mathcal{C}}) = \sum_{a=1}^{q}\sum_{b=1}^{q} \frac{n_{ab}}{n} \log\left(\frac{n\, n_{ab}}{n_a \widehat{n}_b}\right), \tag{30}$$

and their entropies are

$$H(\mathcal{C}) = -\sum_{a=1}^{q} \frac{n_a}{n} \log \frac{n_a}{n}, \qquad H(\widehat{\mathcal{C}}) = -\sum_{b=1}^{q} \frac{\widehat{n}_b}{n} \log \frac{\widehat{n}_b}{n}.$$

Zero-count terms are taken as zero. Using arithmetic-mean normalization, we define

$$\mathrm{NMI}(\mathcal{C},\widehat{\mathcal{C}}) = \frac{2I(\mathcal{C};\widehat{\mathcal{C}})}{H(\mathcal{C}) + H(\widehat{\mathcal{C}})}. \tag{31}$$

NMI lies in $[0, 1]$, with one indicating identical partitions up to label permutation. It is not corrected for chance, so independent random partitions need not have zero empirical NMI.

Adjusted Rand index (ARI) measures pairwise agreement after correcting for chance. Let

$$P_{\mathrm{true}} = \sum_{a} \binom{n_a}{2}, \qquad P_{\mathrm{pred}} = \sum_{b} \binom{\widehat{n}_b}{2}, \qquad P_{\mathrm{both}} = \sum_{a,b} \binom{n_{ab}}{2},$$

which count unordered pairs assigned to the same group in the ground-truth partition, the predicted partition, and both, respectively. Then

$$\mathrm{ARI}(\mathcal{C},\widehat{\mathcal{C}}) = \frac{P_{\mathrm{both}} - P_{\mathrm{true}}P_{\mathrm{pred}}/\binom{n}{2}}{\frac{1}{2}(P_{\mathrm{true}} + P_{\mathrm{pred}}) - P_{\mathrm{true}}P_{\mathrm{pred}}/\binom{n}{2}}. \tag{32}$$

The correction term is the expected value of P_{both} under random permutations preserving the cluster sizes of both partitions. ARI equals one for identical partitions, has expectation zero under this model for nondegenerate partitions, and can be negative when agreement is below chance.

Both metrics are invariant to class and cluster label permutations, and larger values indicate better agreement. Identical degenerate partitions are assigned a score of one. Neither metric is used to optimize the embedding.

Discussion. Together, these metrics evaluate three distinct aspects of embedding quality: NPE, NDS, and SCP assess recovery relative to known latent coordinates; Trustworthiness and Continuity assess local preservation relative to the observed input geometry; and logistic-regression accuracy, NMI, and ARI assess label-related utility. These criteria are complementary rather than interchangeable: preserving input neighborhoods does not ensure semantic usefulness, while strong classification or clustering performance does not establish faithful within-class or global geometry.

D More Information of Real-world Datasets

The detailed information of the 13 real-world datasets are as follows.

- **COIL20** (Nene et al., 1996) contains 1440 gray-scale images of 20 objects.
- **COIL100** (Nene et al., 1996) contains 7200 gray-scale images of 100 objects.
- **MNIST** (LeCun et al., 1998) is composed of 70,000 images of handwritten digits.
- **Fashion-MNIST** (Xiao et al., 2017) is composed of 70,000 images of fashion products from 10 categories.
- **HAR** (Anguita et al., 2013) is a dataset for human activity recognition using smartphones. It contains 10,299 samples and 561 features. The number of classes (activities) is 6.
- **IoT** (Sharmila & Nagapadma, 2023) is a dataset that captures real-time network traffic from diverse IoT devices (such as ThingSpeak-LED and MQTT-Temp) alongside simulated cyberattacks including DDoS, brute-force SSH, and Nmap scans. It provides labeled normal and adversarial behaviors. There are 123,117 samples with 83 features, from 12 classes.

- **Gesture Phase** (Madeo et al., 2016) is designed for segmenting naturally performed gestures into five phases: rest, preparation, stroke, hold, and retraction. We use the processed tabular version, which contains 9,873 samples and 32 numerical motion features extracted from seven videos.
- **Japanese Vowels** (Kudo et al., 1999) is a speech dataset in which nine male speakers successively utter two Japanese vowels. The original recordings comprise 640 variable-length utterances represented by 12 LPC cepstral coefficients. We use the OpenML frame-level representation, which contains 9,961 samples and 14 numerical attributes, including the 12 cepstral coefficients and two sequence-index attributes. The nine classes correspond to the identities of the speakers.
- **Satimage** (Srinivasan, 1993) is the Statlog Landsat Satellite dataset. Each sample describes a 3×3 neighborhood of satellite-image pixels observed in four spectral bands, resulting in 36 numerical features. The OpenML version used in our experiments contains 6,430 samples from six land-cover classes.
- **Texture** (Brodatz, 1966) contains 5,500 samples from 11 texture classes, with 500 samples per class. Each sample is represented by 40 features obtained from fourth-order modified moments measured at four orientations: $0°$, $45°$, $90°$, and $135°$.
- **Allergic Memory** (Ordovas-Montanes et al., 2018) is a single-cell RNA sequencing dataset with 18,036 cells and 22,575 genes. There are 10 classes of cells, where each class corresponds to a cell type.
- **Glial Atlas** (Jager et al., 2022) is a single-cell RNA sequencing dataset with 10,901 cells (we use the SNI subset) and 31,053 genes. There are 10 classes of cells, where each class corresponds to a cell type.
- **Crohn's disease** (Zheng et al., 2021) is a single-cell RNA sequencing dataset with 50,000 cells (sampled from the original data) and 22,431 genes. There are 26 classes of cells, where each class corresponds to a cell type.

Note that for the single-cell datasets, we use the log-transformed values instead of the raw counts. To improve efficiency, we reduce the dimension to 30 using PCA.

E Hyperparameter Sensitivity Analysis

E.1 Metric-wise Neighborhood Sensitivity

Figure 6 provides the metric-wise results corresponding to the composite-score comparison in Figure 1. Each row reports one evaluation criterion—Trustworthiness, Continuity, Procrustes error, Stress, and Spearman correlation—for Swiss Roll and Swiss Roll with Hole. The results confirm that the conclusions from the composite score are not driven by a single metric.

In particular, Isomap and FMDS tend to achieve low Procrustes error and Stress only for relatively small neighborhood sizes, while their geometric recovery deteriorates as K increases. UMAP and t-SNE generally preserve local neighborhoods well, as reflected by strong Trustworthiness and Continuity, but their global distance-recovery metrics can be less stable. PaCMAP maintains competitive performance across several criteria and across a broad parameter range. RobustVis is less reliable for very small K. However, its performance improves rapidly for moderate K values. RobustVis-UW and RobustVis-BR show more stable behavior across the five metrics, indicating that weighting positive-pair constraints can improve the trade-off between local-neighborhood preservation and global geometric recovery. The same trends are observed on Swiss Roll with Hole, where preserving the hole structure makes excessive non-local neighborhood connections particularly harmful.

E.2 Sensitivity to Other Hyperparameters

Synthetic data Figures 7 and 8 show the visualization performance of RobustVis for Swiss Roll when different α and β are used. As $\alpha, \beta \to 0$, the objective approaches ordinary cross-entropy and loses its uniform influence bounds.

Figure 9 shows the visualization performance of RobustVis for Swiss Roll at different iteration numbers, where both PCA initialization and random initialization are tested.

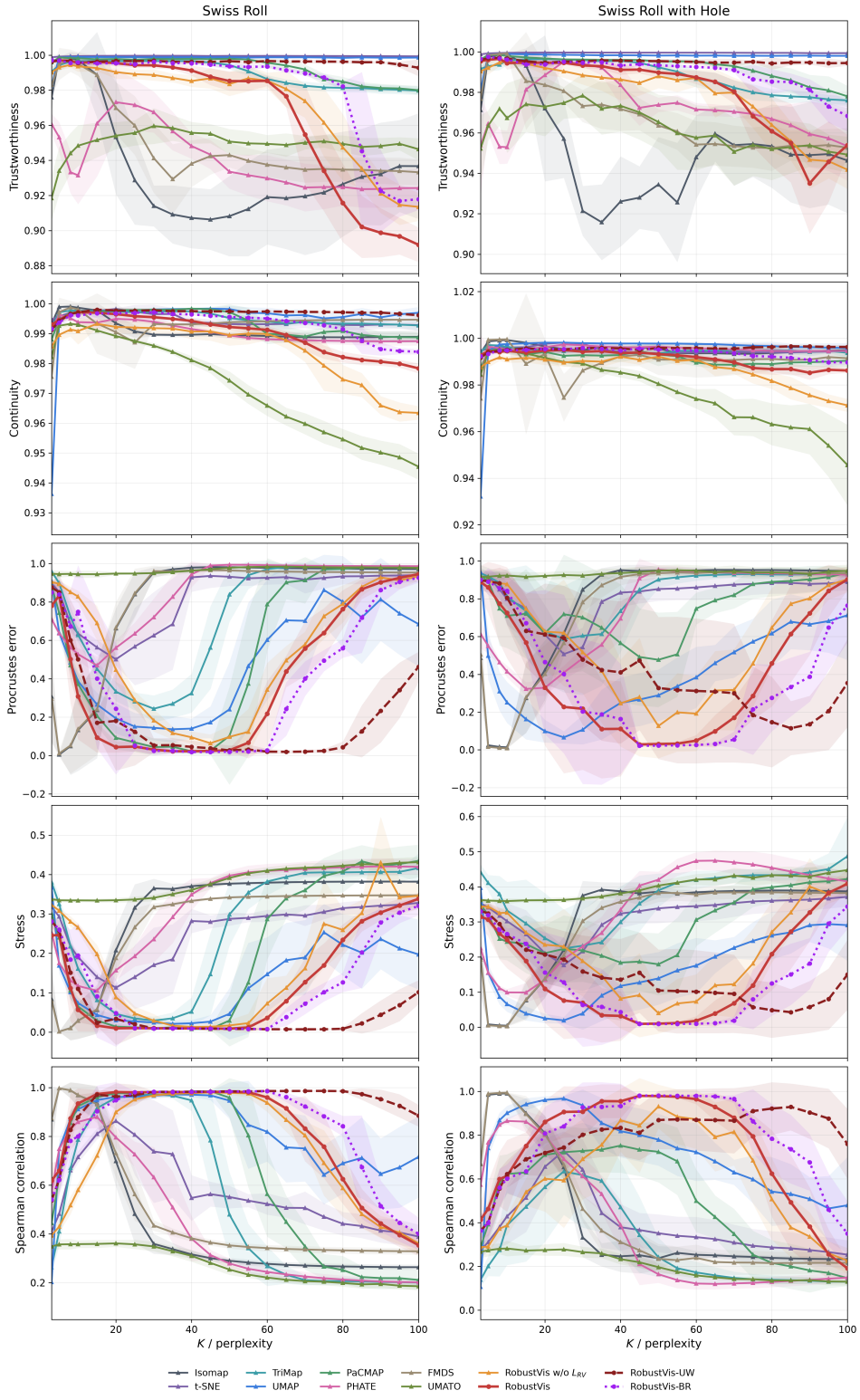
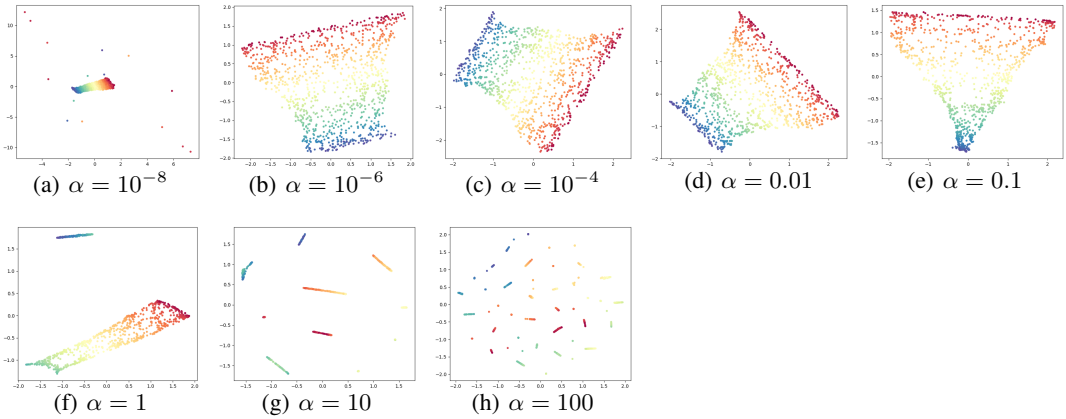


Figure 6: Metric-wise neighborhood sensitivity results on Swiss Roll and Swiss Roll with Hole corresponding to the composite-score comparison in Figure 1.

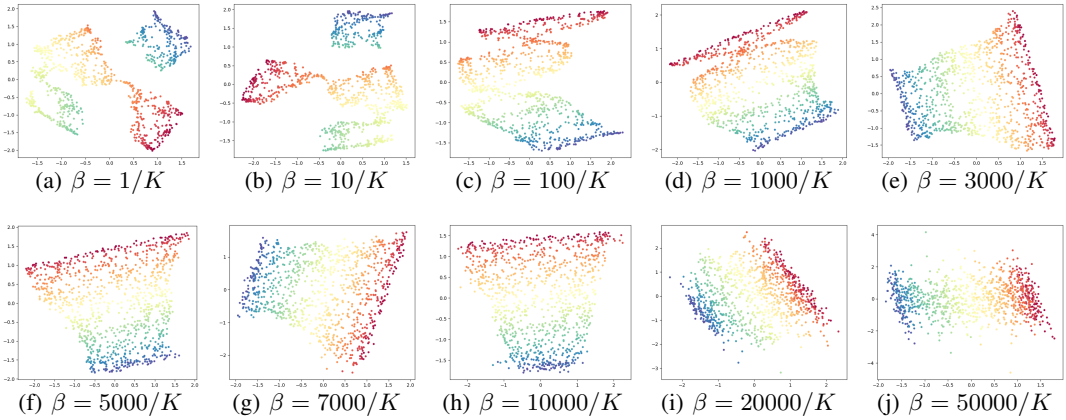
In sum, the results demonstrated the effectiveness of the hyperparameters and rationality of their default values. Moreover, our RobustVis works well across a wide range of hyperparameters.



(a) $\alpha = 10^{-8}$ (b) $\alpha = 10^{-6}$ (c) $\alpha = 10^{-4}$ (d) $\alpha = 0.01$ (e) $\alpha = 0.1$

(f) $\alpha = 1$ (g) $\alpha = 10$ (h) $\alpha = 100$

Figure 7: Influence of α on the performance of RobustVis on Swiss roll (default: $\alpha = 10^{-2}$).



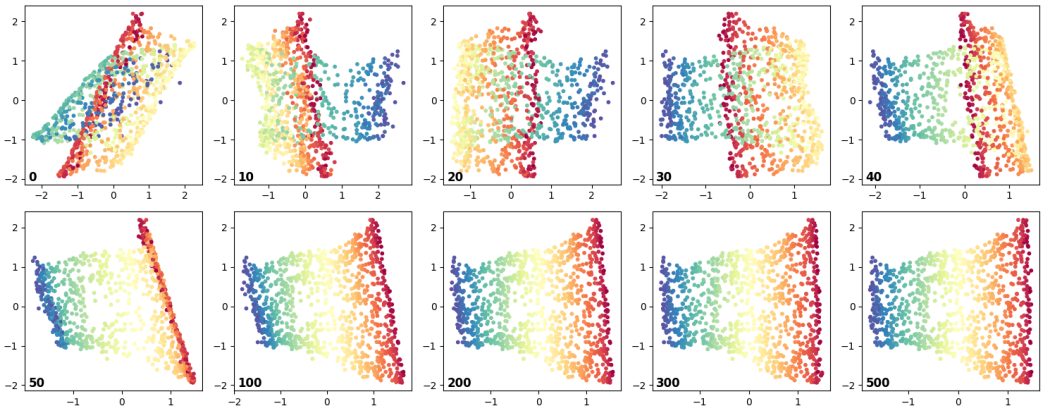
(a) $\beta = 1/K$ (b) $\beta = 10/K$ (c) $\beta = 100/K$ (d) $\beta = 1000/K$ (e) $\beta = 3000/K$

(f) $\beta = 5000/K$ (g) $\beta = 7000/K$ (h) $\beta = 10000/K$ (i) $\beta = 20000/K$ (j) $\beta = 50000/K$

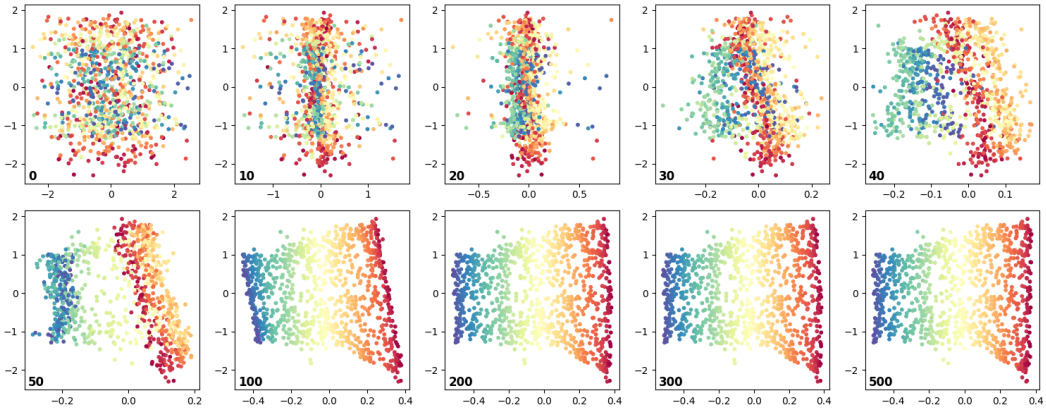
Figure 8: Influence of β on the performance of RobustVis on Swiss roll (default: $\beta = 2n/K$). Here, we have set $K = 40$.

Real datasets Figure 10 examines $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ while fixing $\beta = 2n/K$. All three RobustVis variants remain stable over $\alpha \in [10^{-4}, 10^{-2}]$, with only minor and statistically insignificant differences. The original RobustVis also remains competitive at $\alpha = 10^{-1}$, whereas the two positive-edge-weighted variants begin to deteriorate. In particular, RobustVis-UM shows a significant reduction in logistic accuracy. Setting $\alpha = 1$ substantially degrades all three metrics for every variant across all 13 datasets, indicating that excessive smoothing weakens the distinction between positive and negative pairs. Overall, $\alpha = 10^{-2}$ provides a stable default and achieves the best average performance for RobustVis-BR.

Figure 11 studies $\beta = cn/K$ with $c \in \{0.25, 0.5, 1, 2, 4, 8\}$ over the 13 real-world datasets. A very small coefficient ($c = 0.25$) consistently degrades logistic accuracy, NMI, and ARI for all three RobustVis variants, indicating that insufficient negative-pair repulsion leads to poorly separated embeddings. Performance improves rapidly up to $c = 1$–2 and then largely saturates. Larger values slightly improve logistic accuracy, whereas NMI and ARI remain stable or decrease, suggesting that excessive repulsion may improve linear separability at the expense of cluster compactness. RobustVis and RobustVis-BR obtain their highest numerical average scores at $c = 8$, but their improvements over $c = 2$ are small and not statistically significant. RobustVis-UM peaks near $c = 4$



(a) PCA initialization



(b) random initialization

Figure 9: Performance of RobustVis on Swiss roll in different iterations.

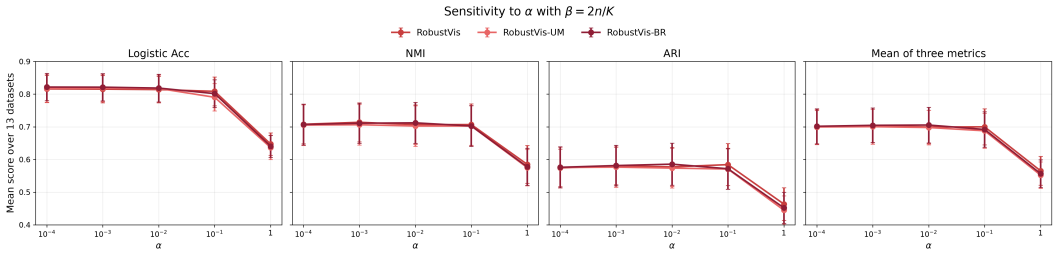


Figure 10: Average sensitivity of RobustVis to α over 13 real-world datasets

and deteriorates at $c = 8$, particularly in ARI. Overall, $c = 2$ provides a stable and well-balanced default across objectives and weighting schemes.

E.3 Handling Disconnected Graph

Here, we use a few examples to show the impact of connecting isolated points and bridging disconnected components.

Figure 12 illustrates the effect of bridging disconnected components on COIL20, where one original image from each class is displayed in the plots. Without this operation, the similarity graph becomes disconnected and splits into 15 components. Since the similarities between different components are

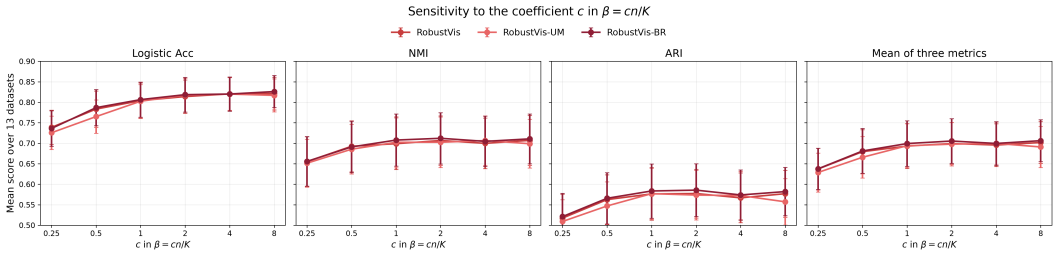
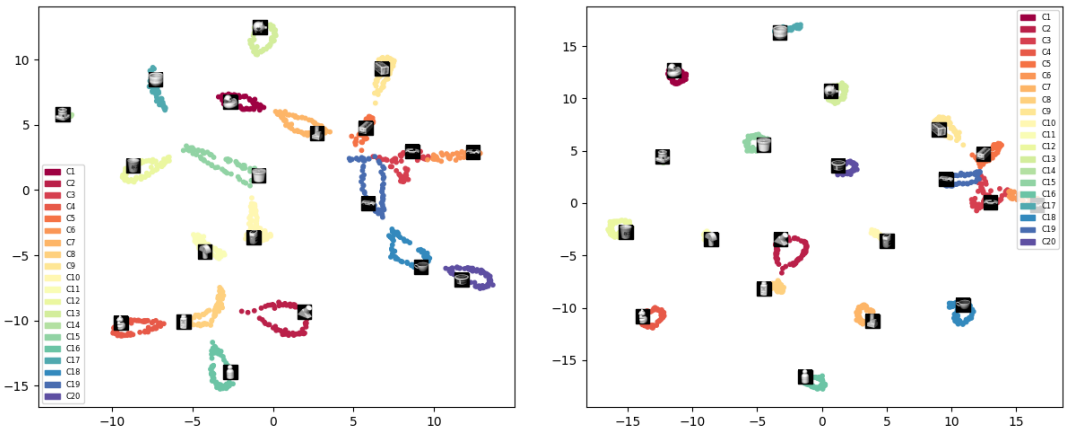


Figure 11: Average sensitivity of RobustVis to β over 13 real-world datasets

undefined, they are effectively treated as zero. As a result, the distances between these components in the low-dimensional space are driven to be as large as possible, which is not meaningful. This effect is evident in Figure 12(b), where some visually similar objects, such as classes 18 and 20, are placed far apart. By contrast, after bridging the components, class 20 is positioned closest to class 18, which is much more consistent with their visual similarity.

Figure 13 demonstrates the effectiveness of connecting isolated points on MNIST. In Figure 13(b), where this operation is not applied, the graph contains 44 isolated points, leading to many isolated samples in the 2-D visualization. Moreover, the locations of these isolated points, which may be viewed as outliers, are not reliable. In contrast, this issue is avoided in Figure 13(a) after connecting the isolated points.

In addition to the visualizations above, we provide quantitative classification and clustering results to further evaluate the operation of bridging components. Table 1 shows that they have almost no impact on the performance of logistic regression. Ablating component bridging from RobustVis slightly reduces the clustering accuracy on MNIST, but can increase the clustering accuracy on COIL20. The reason is that, on COIL20, omitting component bridging increases the distances between clusters; on MNIST, there is only one major component even without component bridging.

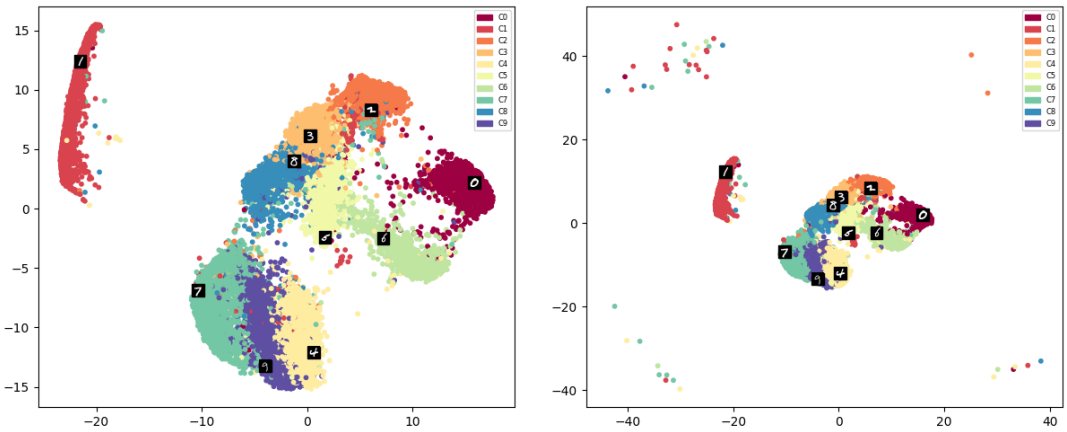


(a) Complete RobustVis (b) RobustVis without component bridging

Figure 12: Ablation study of component bridging on COIL20.

F Robustness to Graph Errors and Negative Sampling

To investigate the effects of graph errors and negative sampling, we compare Robust loss and cross-entropy (CE) loss on Swiss Roll and COIL20 using both full and sampled objectives. We introduce random false-positive (FP) edges at rates of 1%–10% and remove randomly selected edges to simulate false negatives (FN) at rates of 10%–50%, with both rates defined relative to the original number of undirected edges. For each dataset and perturbation realization, all loss variants share



(a) Complete RobustVis (b) RobustVis without isolated-point connection

Figure 13: Ablation study of isolated-point connection on MNIST.

Table 1: Ablation study on connecting isolated points and bridging components. The graph of COIL20 has no isolated points but has 15 components, whereas the graph of MNIST has 44 isolated points but only one major component. Therefore, we use COIL20 to show the impact of component bridging and MNIST to show the impact of isolated-point connection.

Dataset	Metric	Complete RobustVis	RobustVis without graph repair
COIL20	Acc of logistic reg.	0.997	0.997
	NMI of K-means	0.671	0.695
MNIST	Acc of logistic reg.	0.953	0.953
	NMI of K-means	0.744	0.743

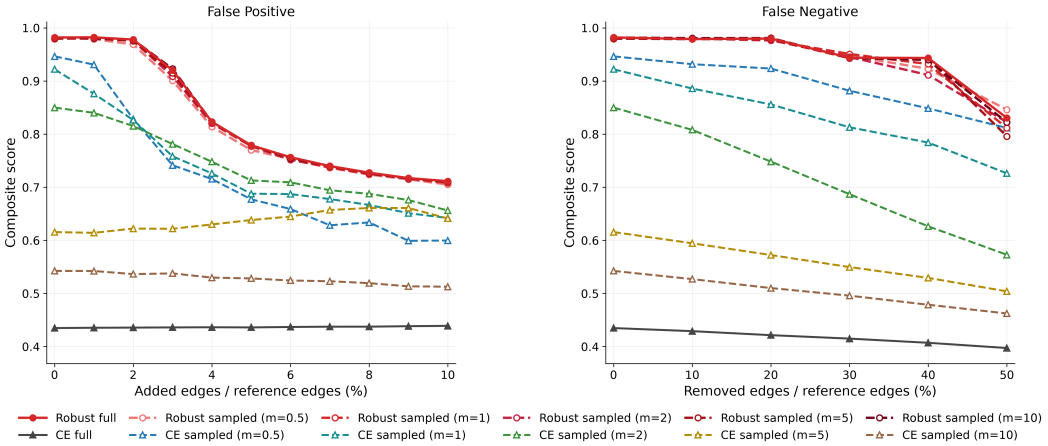
the same graph, initialization, and optimization settings. We evaluate negative-sampling budgets of $m \in \{0.5, 1, 2, 5, 10\}$, where m specifies the number of negative samples relative to the original number of positive edges. Each experiment is repeated ten times. Figure 14 reports the mean composite score: the average of $1 -$ Procrustes error, $1 -$ Stress, and Spearman correlation for Swiss Roll, and the average of k-NN classification accuracy, logistic regression accuracy, NMI, and clustering accuracy for COIL20. Robust loss exhibits relatively consistent performance across sampling budgets, with sampled variants closely matching the full objective, whereas CE is substantially more sensitive to the sampling budget. Robust loss performs better under FP perturbations on Swiss Roll and FN perturbations on COIL20, although its advantage is not universal: CE with larger sampling budgets outperforms it under severe FP perturbations on COIL20. Furthermore, directly optimizing the full CE objective produces substantially poorer embeddings than its sampled variants under the tested settings. Since sampled CE does not apply a sampling correction to the negative term, sampling changes the balance between attraction and repulsion rather than merely accelerating optimization. These results therefore suggest that the effectiveness of sampled CE cannot be attributed to the unweighted full CE objective alone, while highlighting the reduced dependence of Robust loss on the negative-sampling budget.

G Computational Cost Analysis

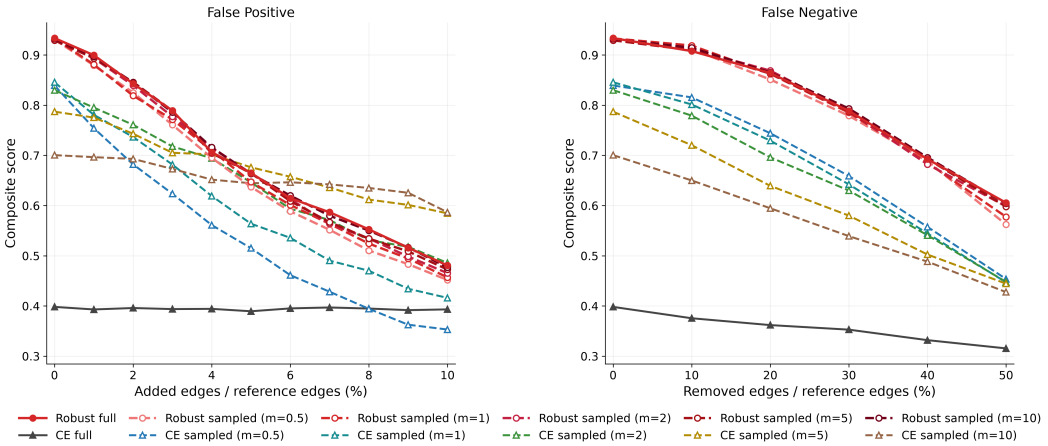
We evaluate the computational cost of RobustVis against several representative visualization methods. In our experiments, t-SNE was implemented using `scikit-learn`, UMAP using the official `umap-learn` package[1], and PaCMAP using its official implementation[2]. All experiments were

[1] `https://umap-learn.readthedocs.io/en/latest/`

[2] `https://github.com/YingfanWang/PaCMAP`



(a) Swiss roll



(b) COIL20

Figure 14: Robustness to graph errors

conducted on a machine with a 12th Gen Intel Core i9-12900K CPU running at 3200 MHz, without GPU acceleration.

Figure 15 reports the fit-only wall-clock time of the six methods on the 13 real-world datasets, ordered by sample size. Data loading, preprocessing, evaluation, and plotting are excluded. PaCMAP and TriMap are consistently the fastest methods, with geometric mean runtimes of 3.01 and 3.95 seconds, respectively. RobustVis has a geometric mean runtime of 10.65 seconds, which is nearly identical to UMAP (10.69 seconds) and lower than t-SNE (13.54 seconds). RobustVis is faster than t-SNE on 11 of the 13 datasets, although its runtime increases on MNIST and Fashion-MNIST. The pronounced peak on COIL-100 is mainly caused by its highly disconnected neighborhood graph: the graph contains many connected components, and RobustVis requires additional time to identify and bridge these components before optimization. Thus, this peak reflects the cost of graph-connectivity repair rather than the embedding optimization itself. Overall, RobustVis provides computational efficiency comparable to UMAP and generally better than t-SNE, while PaCMAP and TriMap remain faster.

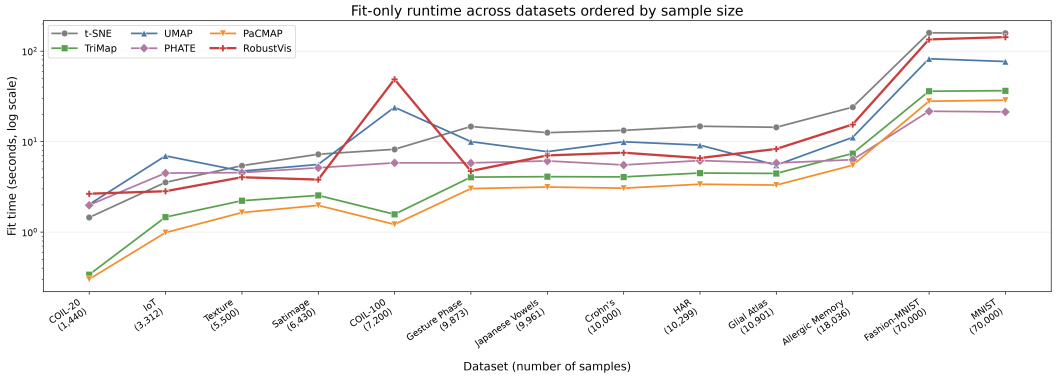


Figure 15: Fit-only runtime of t-SNE, TriMap, UMAP, PHATE, PaCMAP, and RobustVis on 13 real-world datasets ordered by sample size. The reported time includes graph construction and embedding optimization, but excludes data loading, preprocessing, evaluation, and plotting. Each method is run once using the default parameter setting. The runtime peak of RobustVis on COIL-100 is mainly caused by the additional cost of bridging the many connected components in its disconnected neighborhood graph.

H More numerical results of real datasets

H.1 Results under Individual Neighborhood-Scale Settings

Figures 16, 17, and 18 reports the rankings under the default, $2\times$, and $4\times$ neighborhood settings separately. RobustVis-BR ranks first on the combined logistic accuracy, NMI, and ARI criterion in all three settings, with average ranks of 3.23, 3.62, and 3.38, respectively. RobustVis-UW becomes more competitive when the neighborhood size increases, whereas the original RobustVis performs best under the default setting. UMAP benefits from larger neighborhoods, particularly in Continuity, while PaCMAP deteriorates under the $4\times$ setting. These results confirm that RobustVis-BR provides the most stable downstream performance across different neighborhood scales.

H.2 Dataset-Level Results under the Default Setting

Tables 2–14 report the complete dataset-level results under the default parameter setting. Higher values are better. The best and second-best result in each column are shown in bold and underlined, respectively.

Table 2: Results on the COIL-20 dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.920	0.930	0.440	0.650	0.470
Isomap	0.890	0.990	0.540	0.640	0.370
t-SNE	**1.000**	**1.000**	0.780	0.860	0.750
TriMap	0.990	**1.000**	0.920	0.930	0.850
UMAP	**1.000**	0.990	0.830	0.920	0.770
PaCMAP	0.990	0.990	<u>0.950</u>	<u>0.960</u>	<u>0.900</u>
PHATE	0.910	0.990	0.630	0.740	0.500
UMATO	0.930	0.980	0.580	0.690	0.480
FMDS	0.910	0.980	0.580	0.680	0.460
RobustVis w/o L_{RV}	0.970	0.990	0.890	0.890	0.790
RobustVis	0.990	0.990	<u>0.950</u>	0.950	<u>0.900</u>
RobustVis-UW	0.985	<u>0.996</u>	0.928	0.926	0.846
RobustVis-BR	<u>0.991</u>	0.996	**0.957**	**0.966**	**0.926**

Table 3: Results on the COIL-100 dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.910	0.953	0.377	0.687	0.375
Isomap	0.821	0.994	0.299	0.626	0.218
t-SNE	**0.998**	0.994	0.593	0.841	0.602
TriMap	0.963	0.991	0.718	0.843	0.528
UMAP	0.989	0.995	0.709	0.893	0.672
PaCMAP	0.978	0.993	0.817	**0.920**	**0.763**
PHATE	0.827	0.988	0.360	0.643	0.220
UMATO	0.938	0.990	0.401	0.696	0.360
FMDS	0.539	0.596	0.013	0.287	0.041
RobustVis w/o L_{RV}	0.939	0.991	0.681	0.832	0.569
RobustVis	0.966	**0.996**	0.799	0.909	0.725
RobustVis-UW	0.966	0.995	**0.822**	0.893	0.687
RobustVis-BR	0.967	0.996	0.820	0.910	0.718

Table 4: Results on the MNIST dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.739	0.811	0.294	0.238	0.145
Isomap	0.756	0.959	0.459	0.384	0.220
t-SNE	**0.986**	0.974	0.956	0.836	**0.819**
TriMap	0.939	0.973	0.935	0.789	0.676
UMAP	0.956	0.975	0.965	0.850	0.777
PaCMAP	0.950	0.972	**0.969**	**0.863**	0.788
PHATE	0.868	**0.976**	0.702	0.634	0.474
UMATO	0.894	0.972	0.719	0.653	0.539
FMDS	0.603	0.585	0.155	0.071	0.052
RobustVis w/o L_{RV}	0.906	0.964	0.911	0.770	0.672
RobustVis	0.931	0.971	0.951	0.841	0.763
RobustVis-UW	0.936	0.971	0.952	0.842	0.765
RobustVis-BR	0.934	0.971	0.954	0.847	0.771

Table 5: Results on the Fashion-MNIST dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.919	0.953	0.494	0.424	0.272
Isomap	0.922	0.981	0.566	0.486	0.327
t-SNE	**0.990**	0.985	0.665	0.564	0.400
TriMap	0.967	**0.990**	0.703	0.588	0.413
UMAP	0.974	0.988	0.699	0.620	**0.465**
PaCMAP	0.970	0.987	0.706	0.614	0.427
PHATE	0.945	0.988	0.645	0.571	0.390
UMATO	0.959	0.985	0.678	0.523	0.376
FMDS	0.901	0.844	0.369	0.418	0.240
RobustVis w/o L_{RV}	0.937	0.976	0.639	0.582	0.425
RobustVis	0.960	0.987	0.706	**0.630**	0.441
RobustVis-UW	0.961	0.987	0.707	0.620	0.437
RobustVis-BR	0.960	0.987	**0.707**	0.629	0.442

Table 6: Results on the Gesture Phase dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.640	0.869	0.273	0.060	0.027
Isomap	0.661	0.955	0.401	0.066	0.027
t-SNE	**0.960**	0.960	0.440	0.070	0.060
TriMap	0.740	0.954	0.427	0.087	0.068
UMAP	0.920	0.890	0.390	0.050	0.040
PaCMAP	0.800	0.930	0.420	0.080	**0.080**
PHATE	0.657	0.956	**0.442**	**0.092**	0.076
UMATO	0.717	0.942	0.330	0.057	0.054
FMDS	0.546	0.654	0.294	0.061	0.033
RobustVis w/o L_{RV}	0.683	0.937	0.418	0.074	0.064
RobustVis	0.680	**0.970**	0.430	0.080	0.070
RobustVis-UW	0.736	0.966	0.418	0.072	0.069
RobustVis-BR	0.734	0.967	0.419	0.075	0.070

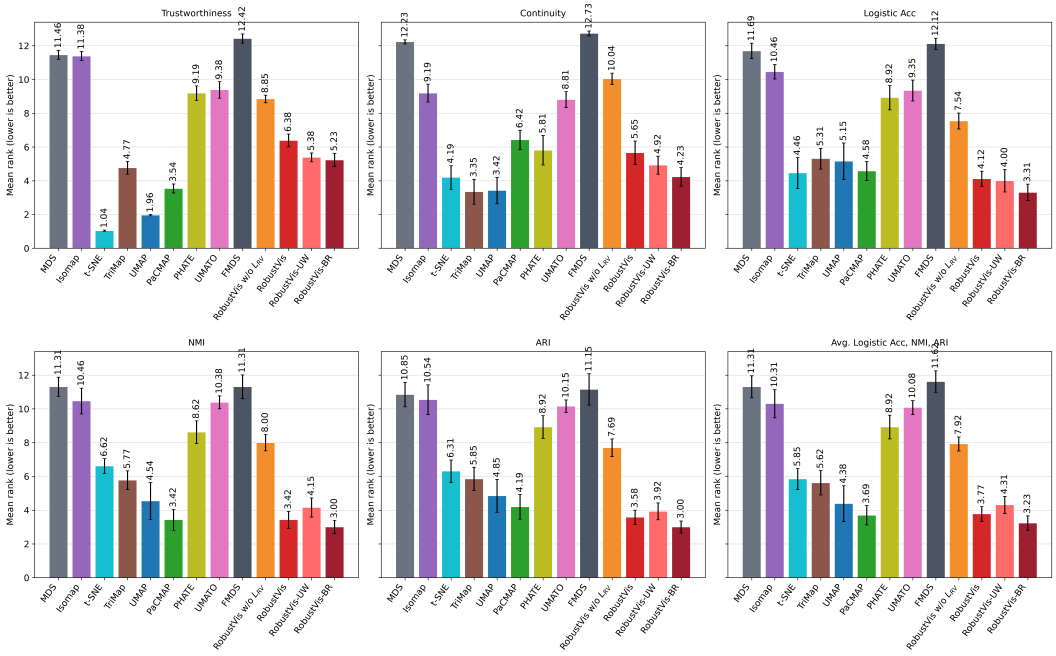


Figure 16: Average within-dataset ranks of all methods over 13 real-world datasets under the default neighborhood setting: $K = 15$ for most datasets, $K = 5$ for COIL-20 and COIL-100, and perplexity 30 for t-SNE. The first five panels report Trustworthiness, Continuity, logistic-regression accuracy, NMI, and ARI, respectively, while the final panel reports the average rank over logistic accuracy, NMI, and ARI. Lower values indicate better performance.

Table 7: Results on the HAR dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.881	0.871	0.505	0.434	0.266
Isomap	0.897	0.958	0.708	0.583	0.451
t-SNE	**0.989**	0.983	**0.835**	0.692	0.615
TriMap	0.955	0.985	0.774	0.650	0.537
UMAP	0.961	**0.986**	0.716	0.710	0.600
PaCMAP	0.954	0.984	0.804	0.718	0.602
PHATE	0.923	0.973	0.767	0.607	0.481
UMATO	0.942	0.981	0.798	0.635	0.535
FMDS	0.881	0.875	0.566	0.440	0.280
RobustVis w/o L_{RV}	0.917	0.967	0.780	0.628	0.546
RobustVis	0.946	0.985	0.807	0.740	0.632
RobustVis-UW	0.947	0.984	0.821	**0.761**	**0.649**
RobustVis-BR	0.947	0.985	0.785	0.729	0.623

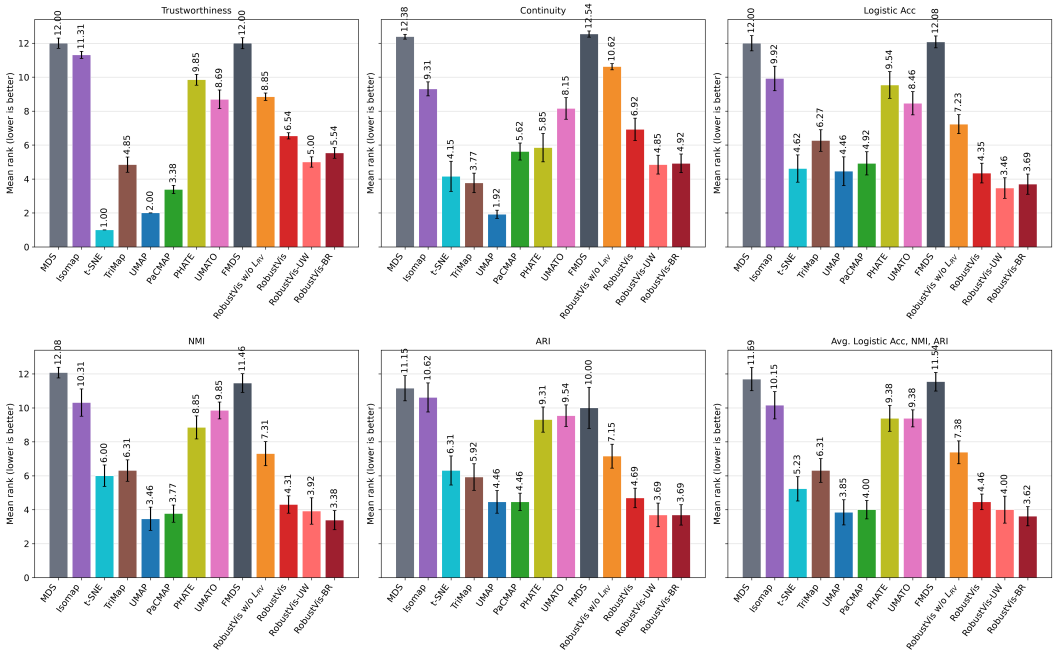


Figure 17: Average within-dataset ranks of all methods over 13 real-world datasets when the neighborhood-related parameters are doubled: $K = 30$ for most datasets, $K = 10$ for COIL-20 and COIL-100, and perplexity 60 for t-SNE. The panels report Trustworthiness, Continuity, logistic-regression accuracy, NMI, ARI, and the average rank over the three label-informed metrics. Lower values indicate better performance.

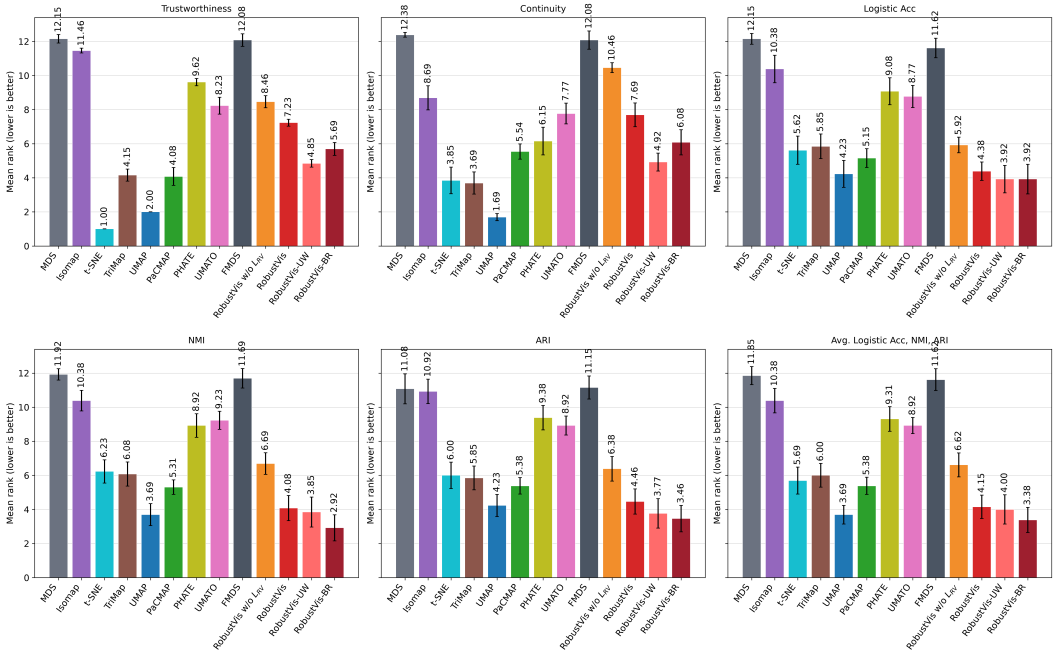


Figure 18: Average within-dataset ranks of all methods over 13 real-world datasets when the neighborhood-related parameters are increased to four times their default values: $K = 60$ for most datasets, $K = 20$ for COIL-20 and COIL-100, and perplexity 120 for t-SNE. The panels report Trustworthiness, Continuity, logistic-regression accuracy, NMI, ARI, and the average rank over logistic accuracy, NMI, and ARI. Lower values indicate better performance.

Table 8: Results on the IoT dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.971	0.947	0.640	0.580	0.506
Isomap	0.976	0.993	0.645	0.633	**0.591**
t-SNE	**0.996**	**0.999**	**0.755**	0.630	0.474
TriMap	0.985	0.996	0.548	0.548	0.376
UMAP	0.994	0.997	0.406	0.338	0.170
PaCMAP	0.986	0.997	0.715	0.552	0.367
PHATE	0.979	0.995	0.488	0.469	0.210
UMATO	0.950	0.982	0.607	0.443	0.262
FMDS	0.975	0.928	0.506	0.589	0.554
RobustVis w/o L_{RV}	0.979	0.990	0.489	0.302	0.164
RobustVis	0.989	0.996	0.684	**0.634**	0.443
RobustVis-UW	0.988	0.996	0.706	0.535	0.374
RobustVis-BR	0.990	0.997	0.693	0.566	0.389

Table 9: Results on the Japanese Vowels dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.839	0.941	0.320	0.209	0.117
Isomap	0.867	0.983	0.491	0.324	0.161
t-SNE	**0.998**	0.986	0.842	0.767	0.709
TriMap	0.970	0.990	0.829	0.674	0.508
UMAP	0.987	0.990	**0.948**	**0.880**	**0.887**
PaCMAP	0.979	0.988	0.917	0.760	0.692
PHATE	0.925	0.990	0.634	0.570	0.377
UMATO	0.917	0.983	0.434	0.364	0.193
FMDS	0.565	0.577	0.163	0.065	0.036
RobustVis w/o L_{RV}	0.918	0.976	0.860	0.731	0.698
RobustVis	0.960	0.989	0.921	0.798	0.759
RobustVis-UW	0.964	**0.991**	0.939	0.844	0.816
RobustVis-BR	0.961	0.990	0.934	0.864	0.868

Table 10: Results on the Satimage dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.937	0.890	0.499	0.469	0.364
Isomap	0.952	0.988	0.801	0.596	0.516
t-SNE	**0.996**	0.990	0.810	0.615	0.513
TriMap	0.971	**0.992**	0.823	0.622	0.545
UMAP	0.986	0.991	0.828	**0.637**	**0.548**
PaCMAP	0.976	0.989	0.815	0.636	0.541
PHATE	0.961	0.991	0.806	0.608	0.502
UMATO	0.970	0.990	0.809	0.508	0.395
FMDS	0.944	0.877	0.564	0.480	0.385
RobustVis w/o L_{RV}	0.959	0.979	0.823	0.616	0.524
RobustVis	0.974	0.990	0.830	0.620	0.530
RobustVis-UW	0.975	0.990	0.830	0.625	0.534
RobustVis-BR	0.976	0.990	**0.830**	0.625	0.534

Table 11: Results on the Texture dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.963	0.959	0.564	0.523	0.365
Isomap	0.953	0.991	0.628	0.551	0.375
t-SNE	**0.999**	0.996	0.913	0.829	0.748
TriMap	0.991	**0.996**	0.931	0.771	0.663
UMAP	0.995	0.996	**0.967**	**0.881**	**0.813**
PaCMAP	0.993	0.994	0.924	0.833	0.740
PHATE	0.976	0.991	0.719	0.647	0.469
UMATO	0.962	0.992	0.558	0.490	0.286
FMDS	0.939	0.890	0.436	0.561	0.358
RobustVis w/o L_{RV}	0.979	0.986	0.920	0.815	0.739
RobustVis	0.988	0.994	0.936	0.853	0.765
RobustVis-UW	0.990	0.995	0.952	0.870	0.796
RobustVis-BR	0.990	0.995	0.949	0.870	0.796

Table 12: Results on the Crohn dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.870	0.884	0.477	0.452	0.277
Isomap	0.866	0.978	0.585	0.501	0.258
t-SNE	**0.997**	0.994	**0.814**	0.710	0.389
TriMap	0.986	0.994	0.805	0.749	0.483
UMAP	0.989	**0.994**	0.812	0.733	0.441
PaCMAP	0.986	0.992	0.790	0.726	0.424
PHATE	0.965	0.992	0.749	0.678	0.428
UMATO	0.962	0.989	0.605	0.578	0.277
FMDS	0.608	0.606	0.196	0.196	0.134
RobustVis w/o L_{RV}	0.962	0.980	0.709	0.675	0.376
RobustVis	0.981	0.992	0.802	**0.754**	**0.488**
RobustVis-UW	0.982	0.992	0.799	0.741	0.480
RobustVis-BR	0.982	0.992	0.801	0.753	0.487

Table 13: Results on the Allergic dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.831	0.862	0.516	0.323	0.240
Isomap	0.870	0.978	0.627	0.446	0.306
t-SNE	**0.996**	0.986	0.851	0.649	0.480
TriMap	0.972	**0.990**	**0.853**	0.705	**0.549**
UMAP	0.985	0.987	0.843	0.714	0.507
PaCMAP	0.978	0.986	0.821	**0.721**	0.515
PHATE	0.940	0.986	0.689	0.614	0.445
UMATO	0.959	0.986	0.818	0.595	0.466
FMDS	0.580	0.578	0.242	0.084	0.074
RobustVis w/o L_{RV}	0.947	0.975	0.824	0.630	0.459
RobustVis	0.966	0.986	0.814	0.684	0.521
RobustVis-UW	0.968	0.986	0.810	0.689	0.526
RobustVis-BR	0.967	0.987	0.832	0.713	0.516

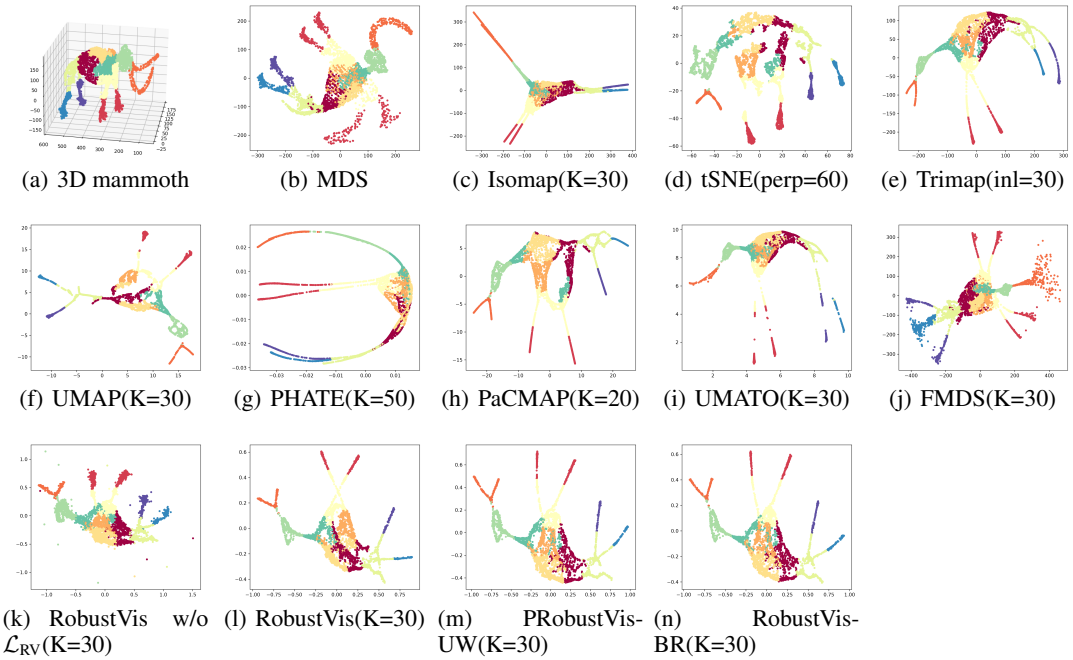
I VISUALIZATION OF MORE DATASETS

Figure 19 compares the embeddings of the Mammoth dataset using $K = 30$ and t-SNE perplexity 60. MDS and FMDS preserve the coarse arrangement of the body and extremities but produce substantial overlap in the central region. Isomap severely stretches the extremities, while t-SNE fragments the connected surface into separate islands. TriMap and UMATO recover a recognizable branching structure but introduce noticeable scale distortions. UMAP and PaCMAP preserve the dominant extremities but compress the body and fail to reveal the tail as a distinct branch. PHATE over-smooths the structure, and RobustVis without $\mathcal{L}_{\mathrm{RV}}$ produces a crowded center. In comparison, RobustVis and its weighted variants preserve a more coherent central structure together with continuous and distinguishable extremities, including the thin tail. This result indicates that the robust objective better preserves both dominant structures and weaker geometric branches.

Figures 20–23 provide additional qualitative comparisons on four real-world datasets. MDS and FMDS generally produce substantial class mixing, while Isomap is sensitive to the data geometry.

Table 14: Results on the Atlas dataset under the default parameter setting.

Method	Trust.	Cont.	Logistic Acc	NMI	ARI
MDS	0.956	0.925	0.689	0.372	0.314
Isomap	0.904	0.990	0.711	0.320	0.171
t-SNE	**0.998**	0.995	**0.976**	0.667	0.398
TriMap	0.987	**0.996**	0.958	0.657	0.408
UMAP	0.992	0.995	0.976	0.732	0.444
PaCMAP	0.991	0.995	0.939	**0.733**	**0.485**
PHATE	0.986	0.994	0.887	0.544	0.344
UMATO	0.927	0.986	0.626	0.370	0.218
FMDS	0.973	0.974	0.832	0.535	0.422
RobustVis w/o L_{RV}	0.973	0.985	0.930	0.647	0.403
RobustVis	0.989	0.995	0.950	0.695	0.468
RobustVis-UW	0.985	0.995	0.940	0.713	0.478
RobustVis-BR	0.986	0.995	0.959	0.710	0.474



(a) 3D mammoth (b) MDS (c) Isomap(K=30) (d) tSNE(perp=60) (e) Trimap(inl=30)

(f) UMAP(K=30) (g) PHATE(K=50) (h) PaCMAP(K=20) (i) UMATO(K=30) (j) FMDS(K=30)

(k) RobustVis w/o $\mathcal{L}_{\mathrm{RV}}$(K=30) (l) RobustVis(K=30) (m) PRobustVis-UW(K=30) (n) RobustVis-BR(K=30)

Figure 19: Visualization of the 3D mammoth data (5000 samples) in 2D space.

Methods such as t-SNE, UMAP, and PaCMAP often create larger inter-class gaps, but may fragment related samples or exaggerate class separation. This behavior is particularly visible on MNIST, where the RobustVis embeddings retain transitions between visually ambiguous digits instead of forcing every class into an isolated component.

On COIL-100, RobustVis and its weighted variants recover numerous coherent curves associated with changes in object viewpoint, while maintaining separation among different objects. They also produce well-organized class structures on Japanese Vowels and Texture. In contrast, RobustVis without $\mathcal{L}_{\mathrm{RV}}$ consistently yields more compact and overlapping embeddings, showing that the graph support alone does not explain the observed performance. RobustVis-UW and RobustVis-BR largely preserve the structure recovered by RobustVis while improving the organization and separation of several classes. Overall, the visualizations support the quantitative results: the robust loss provides a useful balance between semantic separation and the preservation of continuous within-class structure.

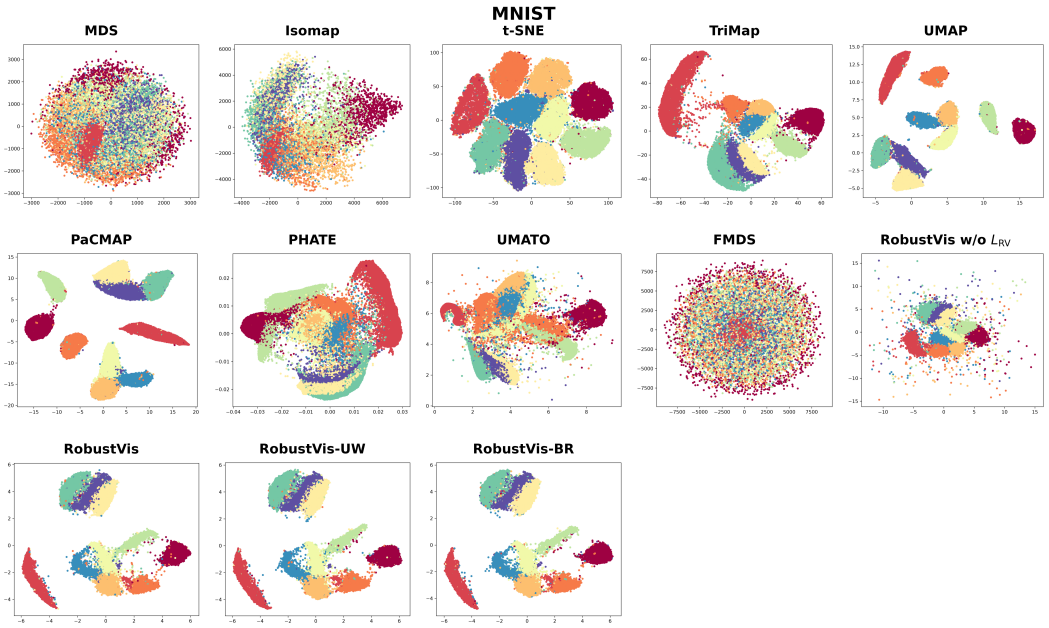


Figure 20: Two-dimensional embeddings of MNIST produced by the 13 methods under the default parameter setting. Colors indicate digit labels. The neighborhood size is $K = 15$ for neighborhood-based methods, and the t-SNE perplexity is 30.

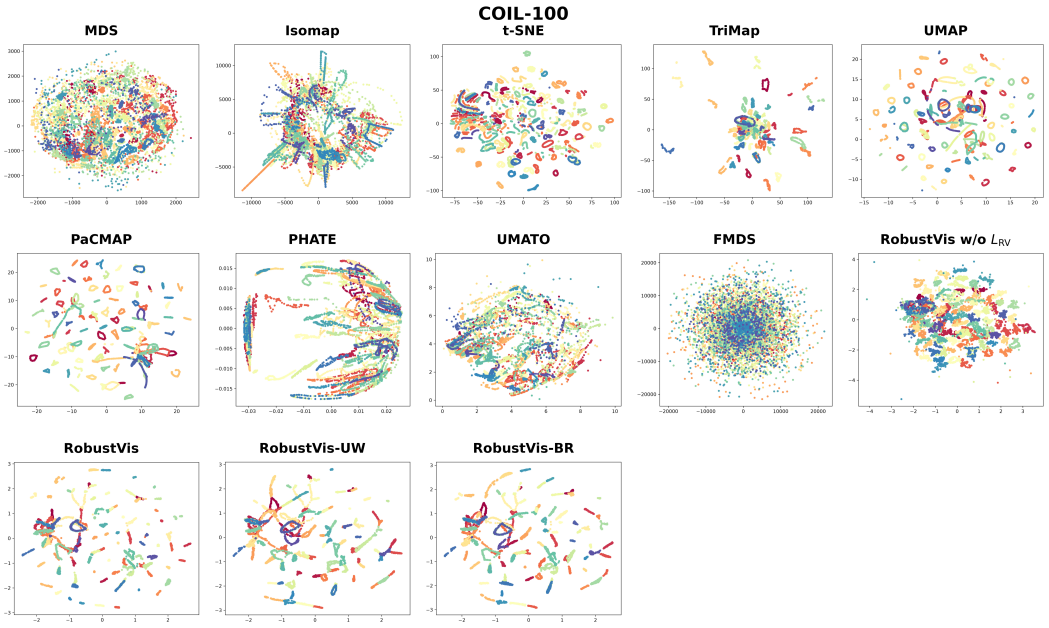


Figure 21: Two-dimensional embeddings of COIL100 produced by the 13 methods under the default parameter setting. Colors indicate object identity labels. The neighborhood size is $K = 15$ for neighborhood-based methods, and the t-SNE perplexity is 30.

J Statistical Significance Analysis

Statistical significance of the robust objective. We further examine whether the improvement of the proposed robust objective over the UMAP-style loss is statistically significant. Each of the 13

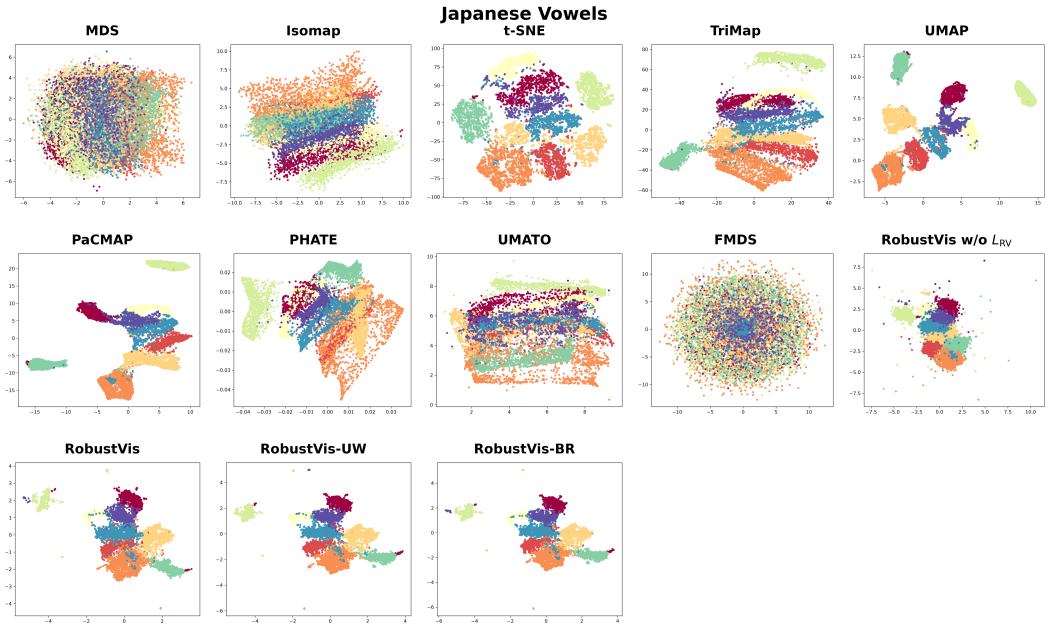


Figure 22: Two-dimensional embeddings of Japanese Vowels produced by the 13 methods under the default parameter setting. Colors indicate speaker identities. The neighborhood size is $K = 15$ for neighborhood-based methods, and the t-SNE perplexity is 30.

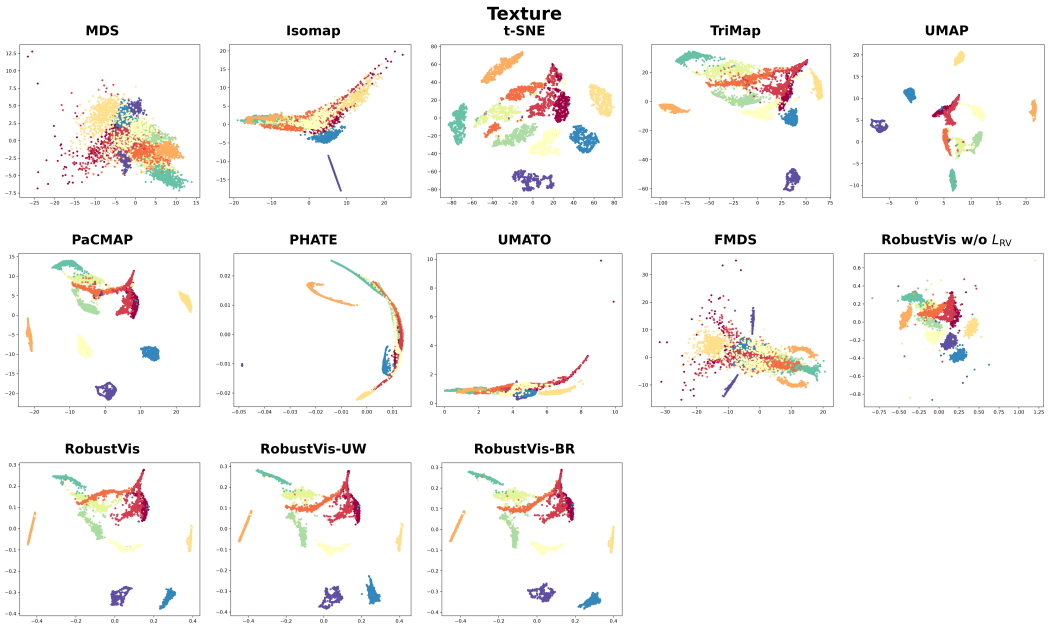


Figure 23: Two-dimensional embeddings of Texture produced by the 13 methods under the default parameter setting. Colors indicate texture-class labels. The neighborhood size is $K = 15$ for neighborhood-based methods, and the t-SNE perplexity is 30.

datasets is treated as an independent block, and the results under the default, $2\times$, and $4\times$ neighborhood settings are averaged within each dataset. We compare RobustVis with *RobustVis w/o* L_{RV}, which uses the same high-dimensional graph and experimental pipeline but replaces the proposed loss with the UMAP-style objective. Pairwise Wilcoxon signed-rank tests are applied to the dataset-

level metric values, followed by Holm correction over all method pairs. RobustVis significantly improves Trustworthiness, Continuity, NMI, and ARI, with adjusted $p = 0.019$ for all four metrics. The improvement in logistic-regression accuracy is favorable but does not remain significant after correction ($p = 0.073$). When logistic accuracy, NMI, and ARI are combined into the downstream criterion, RobustVis also significantly outperforms the UMAP-style loss ($p = 0.019$). RobustVis-BR shows the same pattern, with significant improvements in Trustworthiness ($p = 0.019$), Continuity ($p = 0.019$), NMI ($p = 0.032$), ARI ($p = 0.019$), and the combined downstream criterion ($p = 0.019$). Since the compared variants share the graph construction and differ primarily in their objectives, these results provide direct statistical evidence that the observed improvement is attributable to the proposed robust loss rather than to a change in the similarity graph.

Comparison with strong visualization baselines. We also compare RobustVis-BR with t-SNE, UMAP, and PaCMAP using the same dataset-level statistical protocol. RobustVis-BR obtains better average ranks than all three baselines on each downstream metric. Its mean ranks on logistic accuracy, NMI, and ARI are 3.77, 2.62, and 3.00, respectively, compared with 4.77, 6.31, and 6.23 for t-SNE; 4.31, 4.38, and 4.69 for UMAP; and 4.96, 4.15, and 4.46 for PaCMAP. Nevertheless, these pairwise downstream differences do not remain statistically significant after the conservative Holm correction over all method pairs. By contrast, t-SNE, UMAP, and PaCMAP generally achieve higher Trustworthiness, reflecting their stronger emphasis on preserving input-space neighborhoods. A paired t-test provides additional evidence for the advantage of RobustVis-BR over PaCMAP in ARI: the mean improvement is 0.030, with a two-sided $p = 0.010$ and a Holm-adjusted $p = 0.040$ when correcting over the five metrics for this comparison. Overall, RobustVis-BR exhibits the strongest average downstream performance, although its advantages over complete state-of-the-art pipelines are less conclusive than the statistically significant gains over the controlled UMAP-style loss ablation.

K Graph Visualization

We additionally evaluate the methods on five node-classification graphs: Cora, CiteSeer, and PubMed (Yang et al., 2016), Chameleon (Pei et al., 2020), and Amazon Photo (Shchur et al., 2018). Cora, CiteSeer, and PubMed are citation networks, Chameleon is a Wikipedia page network, and Amazon Photo is a product co-purchase network. We use the largest connected component of each graph, resulting in 2,485 nodes, 5,069 edges, and 7 classes for Cora; 2,120 nodes, 3,679 edges, and 6 classes for CiteSeer; 2,277 nodes, 31,371 edges, and 5 classes for Chameleon; 7,487 nodes, 119,043 edges, and 8 classes for Amazon Photo; and 19,717 nodes, 44,324 edges, and 3 classes for PubMed. All edge counts are undirected.

We compare RobustVis with DeepWalk (Perozzi et al., 2014), node2vec (Grover & Leskovec, 2016), a graph autoencoder (GAE) (Kipf & Welling, 2016), GraphTSNE (Leow et al., 2019), and GAE followed by either UMAP or PaCMAP. We also include RobustVis without $\mathcal{L}_{\mathrm{RV}}$ as an ablation to isolate the effect of the proposed robust loss. All methods use graph structure only. For methods requiring node attributes, the standardized node degree is used as the sole input feature; the original node features and labels are not used during embedding. DeepWalk and node2vec use 10 walks of length 40 per node and directly learn two-dimensional embeddings. GAE is trained for 300 epochs; for GAE+UMAP and GAE+PaCMAP, it first learns a 30-dimensional representation. GraphTSNE uses 1,500 landmarks for Amazon Photo and PubMed. RobustVis and its ablation use the same spectral initialization and 500 optimization epochs, with $\alpha = 0.01$ and $\beta = 2n/\bar{d}$, where $\bar{d}$ is the mean node degree.

Each embedding is evaluated using logistic-regression accuracy, NMI, ARI, and edge-reconstruction AUC. Logistic accuracy is averaged over five stratified 80/20 train–test splits. NMI and ARI are computed from K-means clustering with the number of clusters set to the number of classes. Edge AUC distinguishes observed edges from an equal number of randomly sampled non-edges using negative squared embedding distance. All methods are run with the same random seed, and class labels are used only for evaluation and coloring the visualizations.

Figure 24 compares representative two-dimensional embeddings on three graph datasets. Node2vec and GraphTSNE often compress the nodes into narrow curved structures, while GAE+PaCMAP preserves more visible graph variation but exhibits substantial class mixing. RobustVis without $\mathcal{L}_{\mathrm{RV}}$

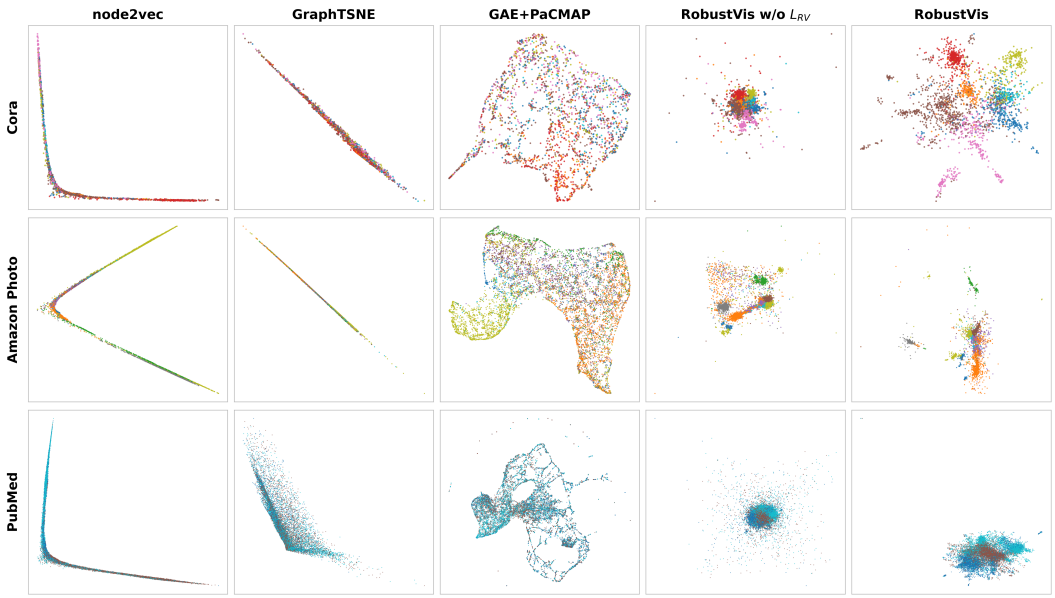


Figure 24: Two-dimensional node embeddings on Cora, Amazon Photo, and PubMed. Rows correspond to datasets and columns to methods. Nodes are colored by class labels, which are used only for visualization.

tends to concentrate most nodes near the center of the embedding. In contrast, RobustVis produces more informative spatial organization, with clearer class-associated regions on Cora and more visible substructures and peripheral groups on Amazon Photo and PubMed. These visualizations are consistent with the quantitative results and illustrate the benefit of the robust loss when the observed graph is used directly.

L Additional Limitations

The proposed objective limits the influence of individual erroneous relations, but it cannot identify the ideal neighborhood graph or guarantee recovery of a globally correct two-dimensional embedding. Its behavior still depends on the input distance, graph construction, initialization, and optimization. The parameters α and β control a robustness–fidelity trade-off and are not theoretically optimal. Moreover, all two-dimensional visualizations necessarily discard information when the intrinsic dimension exceeds two, so apparent cluster separation, density, and inter-cluster distances must be interpreted cautiously and checked against domain knowledge and quantitative diagnostics.

M Proofs for Propositions and Theorems

M.1 Proof for Theorem 3.1

Proof. For any symmetric similarity matrix $\mathbf{S}$, recall that

$$\Omega^+(\mathbf{S}) = \{(i,j) : i < j,\ S_{i,j} > 0\}, \qquad \Omega^-(\mathbf{S}) = \{(i,j) : i < j,\ S_{i,j} = 0\}.$$

Since every unordered pair (i,j) with $i < j$ belongs to exactly one of $\Omega^+(\mathbf{S})$ and $\Omega^-(\mathbf{S})$, we have

$$|\Omega^+(\mathbf{S})| + |\Omega^-(\mathbf{S})| = N, \qquad N = \frac{n(n-1)}{2}.$$

Thus the robust objective can be written as a sum of pairwise losses:

$$\mathcal{L}(\mathbf{Z};\mathbf{S}) = \frac{1}{N}\sum_{i<j} \psi_{S_{i,j}}(A_{i,j}),$$

where, for $s \in [0,1]$,

$$\psi_s(A) = \begin{cases} -s\log(A+\alpha) - (1-s)\log(1-A+\alpha), & s > 0, \\ -\log(1-A+\beta), & s = 0. \end{cases}$$

Here $A = A_{i,j}$ is determined only by the embedding $\mathbf{Z}$ and is independent of the similarity matrix $\mathbf{S}$.

We first prove the objective perturbation bound. Since $0 \le A_{i,j} \le 1$, we have

$$\alpha \le A_{i,j} + \alpha \le 1 + \alpha, \qquad \alpha \le 1 - A_{i,j} + \alpha \le 1 + \alpha.$$

Therefore, for any $s > 0$,

$$-\log(1+\alpha) \le \psi_s(A_{i,j}) \le -\log\alpha.$$

Similarly, since

$$\beta \le 1 - A_{i,j} + \beta \le 1 + \beta,$$

we have

$$-\log(1+\beta) \le \psi_0(A_{i,j}) \le -\log\beta.$$

Under the assumption $0 < \alpha \le \beta$, we have

$$-\log(1+\beta) \le -\log(1+\alpha), \qquad -\log\beta \le -\log\alpha.$$

Hence, for every $s \in [0,1]$,

$$-\log(1+\beta) \le \psi_s(A_{i,j}) \le -\log\alpha.$$

It follows that for any two values $s, t \in [0,1]$,

$$|\psi_s(A_{i,j}) - \psi_t(A_{i,j})| \le \log\left(\frac{1+\beta}{\alpha}\right).$$

Now define the error set

$$E = \{(i,j) : i < j,\ S_{i,j} \ne S^\star_{i,j}\}.$$

For every pair $(i,j) \notin E$, the two similarity values are identical, and hence

$$\psi_{S_{i,j}}(A_{i,j}) = \psi_{S^\star_{i,j}}(A_{i,j}).$$

Therefore,

$$\begin{aligned} |\mathcal{L}(\mathbf{Z};\mathbf{S}) - \mathcal{L}(\mathbf{Z};\mathbf{S}^\star)| &= \left| \frac{1}{N} \sum_{i<j} \left[\psi_{S_{i,j}}(A_{i,j}) - \psi_{S^\star_{i,j}}(A_{i,j}) \right] \right| \\ &= \left| \frac{1}{N} \sum_{(i,j)\in E} \left[\psi_{S_{i,j}}(A_{i,j}) - \psi_{S^\star_{i,j}}(A_{i,j}) \right] \right| \\ &\le \frac{1}{N} \sum_{(i,j)\in E} \left| \psi_{S_{i,j}}(A_{i,j}) - \psi_{S^\star_{i,j}}(A_{i,j}) \right| \\ &\le \frac{|E|}{N} \log\left(\frac{1+\beta}{\alpha}\right). \end{aligned}$$

This proves equation 11.

We next prove the gradient perturbation bound. For each pair (i,j), define

$$\gamma_s(A) = \frac{\partial \psi_s(A)}{\partial A} = \begin{cases} -\dfrac{s}{A+\alpha} + \dfrac{1-s}{1-A+\alpha}, & s > 0, \\ \dfrac{1}{1-A+\beta}, & s = 0. \end{cases}$$

By the chain rule,

$$\nabla_{\mathbf{Z}} \psi_s(A_{i,j}) = \gamma_s(A_{i,j}) \nabla_{\mathbf{Z}} A_{i,j}.$$

Thus

$$\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z};\mathbf{S}) = \frac{1}{N}\sum_{i<j}\gamma_{S_{i,j}}(A_{i,j})\nabla_{\mathbf{Z}}A_{i,j}.$$

Therefore,

$$\begin{aligned}&\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z};\mathbf{S}) - \nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z};\mathbf{S}^\star)\\ &= \frac{1}{N}\sum_{i<j}\left[\gamma_{S_{i,j}}(A_{i,j}) - \gamma_{S^\star_{i,j}}(A_{i,j})\right]\nabla_{\mathbf{Z}}A_{i,j}.\end{aligned}$$

For $(i,j)\notin E$, we have $S_{i,j} = S^\star_{i,j}$, so the corresponding term is zero. Hence

$$\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z};\mathbf{S}) - \nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z};\mathbf{S}^\star) = \frac{1}{N}\sum_{(i,j)\in E}\left[\gamma_{S_{i,j}}(A_{i,j}) - \gamma_{S^\star_{i,j}}(A_{i,j})\right]\nabla_{\mathbf{Z}}A_{i,j}.$$

We now bound the scalar coefficient

$$\left|\gamma_{S_{i,j}}(A_{i,j}) - \gamma_{S^\star_{i,j}}(A_{i,j})\right|.$$

There are two cases.

First, suppose both similarities are positive:

$$S_{i,j} > 0, \qquad S^\star_{i,j} > 0.$$

For any $s,t>0$, we have

$$\begin{aligned}\gamma_s(A) - \gamma_t(A) &= \left(-\frac{s}{A+\alpha} + \frac{1-s}{1-A+\alpha}\right) - \left(-\frac{t}{A+\alpha} + \frac{1-t}{1-A+\alpha}\right)\\ &= -(s-t)\left(\frac{1}{A+\alpha} + \frac{1}{1-A+\alpha}\right).\end{aligned}$$

Since $s,t\in[0,1]$, $|s-t|\le 1$. Therefore,

$$|\gamma_s(A) - \gamma_t(A)| \le \frac{1}{A+\alpha} + \frac{1}{1-A+\alpha}.$$

The function

$$f(A) = \frac{1}{A+\alpha} + \frac{1}{1-A+\alpha}, \qquad A\in[0,1],$$

is convex on $[0,1]$, so its maximum is attained at an endpoint. Hence

$$\max_{A\in[0,1]} f(A) = \frac{1}{\alpha} + \frac{1}{1+\alpha}.$$

Thus, when both similarities are positive,

$$|\gamma_s(A) - \gamma_t(A)| \le \frac{1}{\alpha} + \frac{1}{1+\alpha}.$$

Second, suppose one similarity is positive and the other is zero. Without loss of generality, let $s>0$ and $t=0$. Then

$$\gamma_s(A) = -\frac{s}{A+\alpha} + \frac{1-s}{1-A+\alpha}, \qquad \gamma_0(A) = \frac{1}{1-A+\beta}.$$

The value $\gamma_s(A)$ is a convex combination of $-\frac{1}{A+\alpha}$ and $\frac{1}{1-A+\alpha}$, because

$$\gamma_s(A) = s\left(-\frac{1}{A+\alpha}\right) + (1-s)\left(\frac{1}{1-A+\alpha}\right).$$

Therefore,

$$|\gamma_s(A)| \le \max\left\{\frac{1}{A+\alpha}, \frac{1}{1-A+\alpha}\right\} \le \frac{1}{\alpha}.$$

Moreover,

$$|\gamma_0(A)| = \frac{1}{1 - A + \beta} \leq \frac{1}{\beta}.$$

Hence, when one similarity is positive and the other is zero,

$$|\gamma_s(A) - \gamma_0(A)| \leq |\gamma_s(A)| + |\gamma_0(A)| \leq \frac{1}{\alpha} + \frac{1}{\beta}.$$

Combining the two cases, we obtain the uniform bound

$$|\gamma_s(A) - \gamma_t(A)| \leq C_{\text{grad}},$$

where

$$C_{\text{grad}} = \max\left\{\frac{1}{\alpha} + \frac{1}{1+\alpha}, \frac{1}{\alpha} + \frac{1}{\beta}\right\}.$$

Therefore,

$$\begin{aligned}
&\|\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}; \mathbf{S}) - \nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}; \mathbf{S}^\star)\|_F \\
&\leq \frac{1}{N}\sum_{(i,j)\in E}\left|\gamma_{S_{i,j}}(A_{i,j}) - \gamma_{S^\star_{i,j}}(A_{i,j})\right| \|\nabla_{\mathbf{Z}} A_{i,j}\|_F \\
&\leq \frac{1}{N}\sum_{(i,j)\in E} C_{\text{grad}} G_{\max}(\mathbf{Z}) \\
&= \frac{|E|}{N} C_{\text{grad}} G_{\max}(\mathbf{Z}),
\end{aligned}$$

where

$$G_{\max}(\mathbf{Z}) = \max_{i<j} \|\nabla_{\mathbf{Z}} A_{i,j}\|_F.$$

This proves the gradient perturbation bound.

It remains to compute $\|\nabla_{\mathbf{Z}} A_{i,j}\|_F$ for the Cauchy affinity. Let

$$\mathbf{d}_{i,j} = \mathbf{z}_i - \mathbf{z}_j, \qquad B_{i,j} = 1 + \frac{\|\mathbf{d}_{i,j}\|^2}{\sigma^2}.$$

Then

$$A_{i,j} = B_{i,j}^{-(l+1)/2}.$$

Differentiating with respect to $\mathbf{z}_i$ gives

$$\begin{aligned}
\nabla_{\mathbf{z}_i} A_{i,j} &= -\frac{l+1}{2} B_{i,j}^{-(l+3)/2} \nabla_{\mathbf{z}_i}\left(1 + \frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\sigma^2}\right) \\
&= -\frac{l+1}{\sigma^2} B_{i,j}^{-(l+3)/2}(\mathbf{z}_i - \mathbf{z}_j).
\end{aligned}$$

Since

$$A_{i,j} = B_{i,j}^{-(l+1)/2},$$

we have

$$B_{i,j}^{-(l+3)/2} = A_{i,j}^{\frac{l+3}{l+1}}.$$

Therefore,

$$\nabla_{\mathbf{z}_i} A_{i,j} = -(l+1)\sigma^{-2} A_{i,j}^{\frac{l+3}{l+1}}(\mathbf{z}_i - \mathbf{z}_j).$$

Similarly,

$$\nabla_{\mathbf{z}_j} A_{i,j} = -\nabla_{\mathbf{z}_i} A_{i,j}.$$

The full gradient $\nabla_{\mathbf{Z}} A_{i,j}$ has nonzero blocks only at $\mathbf{z}_i$ and $\mathbf{z}_j$. Hence

$$\begin{aligned}
\|\nabla_{\mathbf{Z}} A_{i,j}\|_F^2 &= \|\nabla_{\mathbf{z}_i} A_{i,j}\|_2^2 + \|\nabla_{\mathbf{z}_j} A_{i,j}\|_2^2 \\
&= 2\|\nabla_{\mathbf{z}_i} A_{i,j}\|_2^2.
\end{aligned}$$

Thus

$$\|\nabla_{\mathbf{Z}} A_{i,j}\|_F = \sqrt{2}(l+1)\sigma^{-2} A_{i,j}^{\frac{l+3}{l+1}} \|\mathbf{z}_i - \mathbf{z}_j\|_2.$$

To obtain a bound independent of the particular embedding, let $t = \|\mathbf{z}_i - \mathbf{z}_j\|_2/\sigma$. Then

$$\|\nabla_{\mathbf{z}} A_{i,j}\|_F = \frac{\sqrt{2}(l+1)}{\sigma} \frac{t}{(1+t^2)^{(l+3)/2}}.$$

The function $t/(1+t^2)^{(l+3)/2}$ on $[0,\infty)$ is maximized at $t = 1/\sqrt{l+2}$, where its value is

$$\frac{(l+2)^{(l+2)/2}}{(l+3)^{(l+3)/2}}.$$

Hence

$$\|\nabla_{\mathbf{z}} A_{i,j}\|_F \le \frac{\sqrt{2}(l+1)}{\sigma} \frac{(l+2)^{(l+2)/2}}{(l+3)^{(l+3)/2}},$$

which also gives the stated bound on $G_{\max}(\mathbf{Z})$. This completes the proof. □

M.2 Proof for Theorem 3.2

Proof. For the positive-pair loss,

$$\frac{\partial \ell_w^+(A)}{\partial A} = -\frac{w}{A+\alpha} + \frac{1-w}{1-A+\alpha},$$

and

$$\frac{\partial^2 \ell_w^+(A)}{\partial A^2} = \frac{w}{(A+\alpha)^2} + \frac{1-w}{(1-A+\alpha)^2} > 0.$$

Thus, ℓ_w^+ is strictly convex on $[0,1]$. The unconstrained stationary point satisfies

$$-\frac{w}{A+\alpha} + \frac{1-w}{1-A+\alpha} = 0,$$

or equivalently,

$$(1-w)(A+\alpha) = w(1-A+\alpha).$$

Solving for A gives

$$A = w + (2w-1)\alpha.$$

Therefore, after imposing $A \in [0,1]$, the unique minimizer is

$$A_w^\star = \Pi_{[0,1]}(w + (2w-1)\alpha)\,.$$

Since $w \in [0,1]$, we have $\Pi_{[0,1]}(w) = w$. By the nonexpansiveness of projection onto $[0,1]$,

$$|A_w^\star - w| = \left|\Pi_{[0,1]}(w + (2w-1)\alpha) - \Pi_{[0,1]}(w)\right| \tag{33}$$
$$\le |(2w-1)\alpha| \le \alpha, \tag{34}$$

where the last inequality follows from $|2w-1| \le 1$. For a binary positive relation, $w = 1$, so the unconstrained minimizer is $1+\alpha > 1$, and hence

$$A_1^\star = 1.$$

For a negative pair,

$$\ell^-(A) = -\log(1-A+\beta),$$

and

$$\frac{\partial \ell^-(A)}{\partial A} = \frac{1}{1-A+\beta} > 0$$

for all $A \in [0,1]$. Thus ℓ^- is strictly increasing and is uniquely minimized at

$$A_-^\star = 0.$$

Therefore, the robustification preserves the ideal affinities of binary positive and negative relations exactly, while shifting a weighted positive target by at most α. □

M.3 Proof for Proposition 3.3

Proof. For each negative pair $e = (i, j) \in \Omega^-$, define the inclusion indicator

$$I_e = \mathbb{1}\{e \in \Omega_s^-\}.$$

Because Ω_s^- is sampled uniformly without replacement from Ω^- with fixed size b, every negative pair has the same inclusion probability

$$\mathbb{P}(e \in \Omega_s^-) = \mathbb{E}[I_e] = \frac{b}{|\Omega^-|}. \tag{35}$$

Recall that the sampled loss can be written as

$$\widehat{\mathcal{L}}_{\mathrm{RV}} = \frac{2}{n(n-1)} \left[\sum_{e \in \Omega^+} \ell_e^+ + \frac{|\Omega^-|}{b} \sum_{e \in \Omega_s^-} \ell_e^- \right].$$

Using the inclusion indicators,

$$\sum_{e \in \Omega_s^-} \ell_e^- = \sum_{e \in \Omega^-} I_e \ell_e^-.$$

Therefore,

$$\mathbb{E}\left[\frac{|\Omega^-|}{b} \sum_{e \in \Omega_s^-} \ell_e^- \right] = \frac{|\Omega^-|}{b} \mathbb{E}\left[\sum_{e \in \Omega^-} I_e \ell_e^- \right] \tag{36}$$

$$= \frac{|\Omega^-|}{b} \sum_{e \in \Omega^-} \mathbb{E}[I_e] \ell_e^- \tag{37}$$

$$= \frac{|\Omega^-|}{b} \sum_{e \in \Omega^-} \frac{b}{|\Omega^-|} \ell_e^- \tag{38}$$

$$= \sum_{e \in \Omega^-} \ell_e^-. \tag{39}$$

The positive-pair term is evaluated exactly and does not depend on the sampled negative set. Hence,

$$\mathbb{E}\left[\widehat{\mathcal{L}}_{\mathrm{RV}} \right] = \frac{2}{n(n-1)} \left[\sum_{e \in \Omega^+} \ell_e^+ + \sum_{e \in \Omega^-} \ell_e^- \right] \tag{40}$$

$$= \mathcal{L}_{\mathrm{RV}}, \tag{41}$$

which proves unbiasedness of the sampled objective.

We next consider its gradient. For a fixed embedding $\mathbf{Z}$, define

$$\mathbf{g}_e^- = \nabla_{\mathbf{Z}} \ell_e^-, \qquad e \in \Omega^-.$$

Since the sampled set enters only through the indicators I_e,

$$\nabla_{\mathbf{Z}} \widehat{\mathcal{L}}_{\mathrm{RV}} = \frac{2}{n(n-1)} \left[\sum_{e \in \Omega^+} \nabla_{\mathbf{Z}} \ell_e^+ + \frac{|\Omega^-|}{b} \sum_{e \in \Omega^-} I_e \mathbf{g}_e^- \right]. \tag{42}$$

Taking expectation and again using equation 35,

$$\mathbb{E}\left[\nabla_{\mathbf{Z}} \widehat{\mathcal{L}}_{\mathrm{RV}} \right] = \frac{2}{n(n-1)} \left[\sum_{e \in \Omega^+} \nabla_{\mathbf{Z}} \ell_e^+ + \frac{|\Omega^-|}{b} \sum_{e \in \Omega^-} \mathbb{E}[I_e] \mathbf{g}_e^- \right] \tag{43}$$

$$= \frac{2}{n(n-1)} \left[\sum_{e \in \Omega^+} \nabla_{\mathbf{Z}} \ell_e^+ + \sum_{e \in \Omega^-} \nabla_{\mathbf{Z}} \ell_e^- \right] \tag{44}$$

$$= \nabla_{\mathbf{Z}} \mathcal{L}_{\mathrm{RV}}. \tag{45}$$

Thus, the sampled gradient is also unbiased.

We finally establish the finite-sample approximation bound. Define the mean negative-pair loss over all negative pairs by

$$\overline{\ell}^{-} = \frac{1}{|\Omega^{-}|} \sum_{e \in \Omega^{-}} \ell_e^{-},$$

and its sample counterpart by

$$\widehat{\ell}^{-} = \frac{1}{b} \sum_{e \in \Omega_s^{-}} \ell_e^{-}.$$

The full and sampled objectives differ only in their negative-pair terms. Therefore,

$$\widehat{\mathcal{L}}_{\mathrm{RV}} - \mathcal{L}_{\mathrm{RV}} = \frac{2}{n(n-1)} \left[\frac{|\Omega^{-}|}{b} \sum_{e \in \Omega_s^{-}} \ell_e^{-} - \sum_{e \in \Omega^{-}} \ell_e^{-} \right] \tag{46}$$

$$= \frac{2|\Omega^{-}|}{n(n-1)} \left(\widehat{\ell}^{-} - \overline{\ell}^{-} \right). \tag{47}$$

From equation 9, every negative-pair loss satisfies

$$-\log(1+\beta) < \ell_e^{-} \leq -\log \beta.$$

Hence all values in the finite population $\{\ell_e^{-} : e \in \Omega^{-}\}$ lie in an interval of width

$$R_\beta = -\log \beta + \log(1+\beta) \tag{48}$$

$$= \log\left(\frac{1+\beta}{\beta}\right). \tag{49}$$

Since Ω_s^{-} is a uniform sample without replacement, Hoeffding's concentration inequality for finite-population sampling implies that, for every $t > 0$,

$$\Pr\left(\left| \widehat{\ell}^{-} - \overline{\ell}^{-} \right| \geq t \right) \leq 2 \exp\left(-\frac{2bt^2}{R_\beta^2} \right). \tag{50}$$

Setting the right-hand side equal to δ gives

$$t = R_\beta \sqrt{\frac{\log(2/\delta)}{2b}}.$$

Consequently, with probability at least $1 - \delta$,

$$\left| \widehat{\ell}^{-} - \overline{\ell}^{-} \right| \leq \log\left(\frac{1+\beta}{\beta}\right) \sqrt{\frac{\log(2/\delta)}{2b}}.$$

Combining this inequality with equation 47 yields

$$\left| \widehat{\mathcal{L}}_{\mathrm{RV}} - \mathcal{L}_{\mathrm{RV}} \right| \leq \frac{2|\Omega^{-}|}{n(n-1)} \log\left(\frac{1+\beta}{\beta}\right) \sqrt{\frac{\log(2/\delta)}{2b}},$$

which proves equation 16. □