Penn Medicine

Bi-Lipschitz Autoencoder With Injectivity Guarantee

Qipeng Zhan, Zhuoping Zhou, Zexuan Wang, Qi long, Li Shen
University of Pennsylvania
April 4, 2026

Motivation: Why Do Autoencoders Fail?

Standard Autoencoder Pipeline

Existing Regularization Approaches:

Gradient-based: Constrain Jacobian of encoder/decoder (CAE, IRAE, GeomAE)

Geometry-based: Align latent with k-NN graph distances (SPAE, TAE)

Embedding-based: Match embeddings learned by no-AE methods (DN, GRAE)

Critical Failures

Non-injective mappings

Distinct data points collapse to the same latent code

Rigid isometry

Requires $O(k^2)$ latent dims even when intrinsic dimension of the underlined manifold is only k

Distribution sensitivity

Fails under sparse sampling or distribution shifts

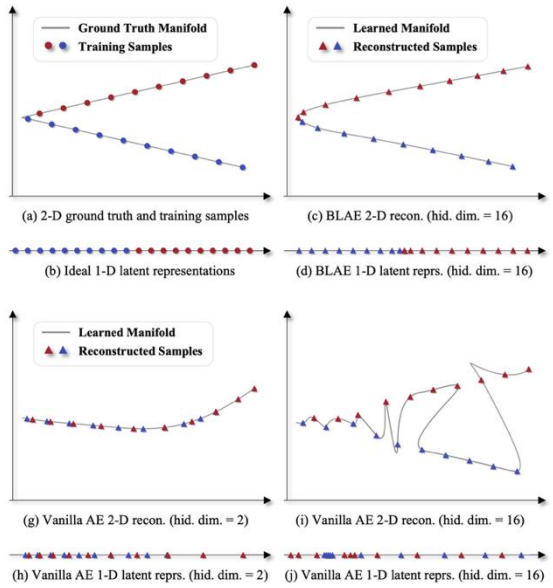
Local minima traps

Non-injectivity creates pathological local minima

Core Problem: The Non-Injectivity Bottleneck

Key Insight: When encoder is non-injective, disjoint manifold regions collapse into overlapping latent codes — creating inescapable local minima for gradient descent.

V-Shaped Manifold: Toy Example



Why This Matters

01 — Latent Space Collisions
Distant manifold regions map to identical or near-identical latent codes

02 — Decoder Instability
Sharp decoder variations needed to handle collapsed codes, scaling with data density

03 — Optimization Traps
Pathological local minima form — gradient descent cannot escape even with large capacity

04 — Local != Global
Jacobian constraints ensure local isometry but fail to guarantee global injectivity

BLAE Innovation 1: Injective Regularization

Def. (δ, ϵ)-Separation: $f: \mathrm{M} \to \mathrm{N}$ is (δ, ϵ)-separated if for all $x, y \in M$ with $d_M(x, y) \geq \delta$:

$$\frac{d_N(f(x), f(y))}{d_M(x, y)} > \epsilon$$

Theorem 1: f is injective <==> for any $\delta > 0$, there exists $\epsilon > 0$ such that f is (δ, ϵ)-separated.

Injective Loss

$$\mathcal{L}_{\text{inj}}(\delta, \epsilon) = \mathbb{E}_{x,y\sim\mathbb{P}}\left[\text{ReLU}\left(\log \frac{\epsilon d_{\mathcal{M}}(x, y)}{d_{\mathcal{N}}(\mathcal{E}_\theta(x), \mathcal{E}_\theta(y))}\right) \cdot \mathbf{1}_{d_{\mathcal{M}}(x,y)>\delta}\right]$$

Penalizes pairs violating the separation condition

Full Regularization

$$\mathcal{L}_{\text{reg}}(\delta, \epsilon) = \mathcal{L}_{\text{inj}}(\delta, \epsilon) + \alpha \cdot \mathbb{E}_{x,y\sim\mathbb{P}}\left[\text{ReLU}\left(\frac{d_{\mathcal{N}}(\mathcal{E}_\theta(x), \mathcal{E}_\theta(y))}{d_{\mathcal{M}}(x, y)} - 1\right) \cdot \mathbf{1}_{d_{\mathcal{M}}(x,y)>\delta}\right]$$

Non-expansive constraint prevents trivial scaling solutions

BLAE Innovation 2: Bi-Lipschitz Relaxation

Isometry Fails

Nash Embedding Theorem:
k-dim manifold needs $O(k\text{^}2)$ latent dims

Example: dim=10 => 50 latent dims needed!

Defeats dim. reduction; becomes non-admissible.

✓ Bi-Lipschitz: Principled Relaxation

κ-Bi-Lipschitz definition:

$$\frac{1}{\kappa} * d_M(x, y) \leq d_N(f(x), f(y)) \leq \kappa * d_M(x, y)$$

Bounded distortion — preserves geometry
Admissible with $O(m)$ latent dims (Thm 4)
Robust to distribution shifts across samplings

Bi-Lipschitz Regularization (via Decoder Jacobian J_D):

$$L_{bi-Lip}(\kappa) = E_{x\sim P}\left[ReLU\left(\frac{1}{\kappa} - \sigma_{\min}(x)\right)^2 + ReLU(\sigma_{\max}(x) - \kappa)^2\right]$$

$\sigma_{\min}(x)$, $\sigma_{\max}(x)$ = *minimum and maximum singular values of* $J_D(x)$. *Applied to decoder (not encoder) for dimensional compatibility.*

BLAE: $L_{BLAE} = L_{recon} + \lambda_{reg} * L_{reg} + \lambda_{bi-Lip} * L_{bi-Lip}$

Experiments: BLAE Outperforms 9 Baselines

Average Rank Across All Datasets (Lower = Better)

Metric	BLAE (Ours)	SPAE	TAE	DN	GRAE	CAE	GGAE	IRAE	GAE	Vanilla AE
k-NN Recall	**1.8**	3.2	3.8	4.5	4.0	5.8	7.5	7.2	9.0	7.8
KL_0.01	**1.0**	3.0	2.5	6.0	4.5	7.0	8.2	6.8	6.5	9.2
KL_0.1	**1.0**	3.2	2.8	5.5	3.5	7.5	9.0	7.2	6.5	8.8
KL_1	**1.0**	4.2	3.2	5.2	4.2	7.0	8.0	7,2	6.0	8.2
MSE	**1.2**	4.8	4.0	5.2	4.5	5.5	7.5	7.0	8.0	7.2

Swiss Roll

2D manifold in 3D space with topology gap. BLAE correctly unrolls geometry; gradient-based baselines distort near the removed strip.

>> Only method preserving full topology

dSprites

64x64 binary images. Cross-shaped excluded region tests cluster topology. BLAE best preserves the parallel plane structure.

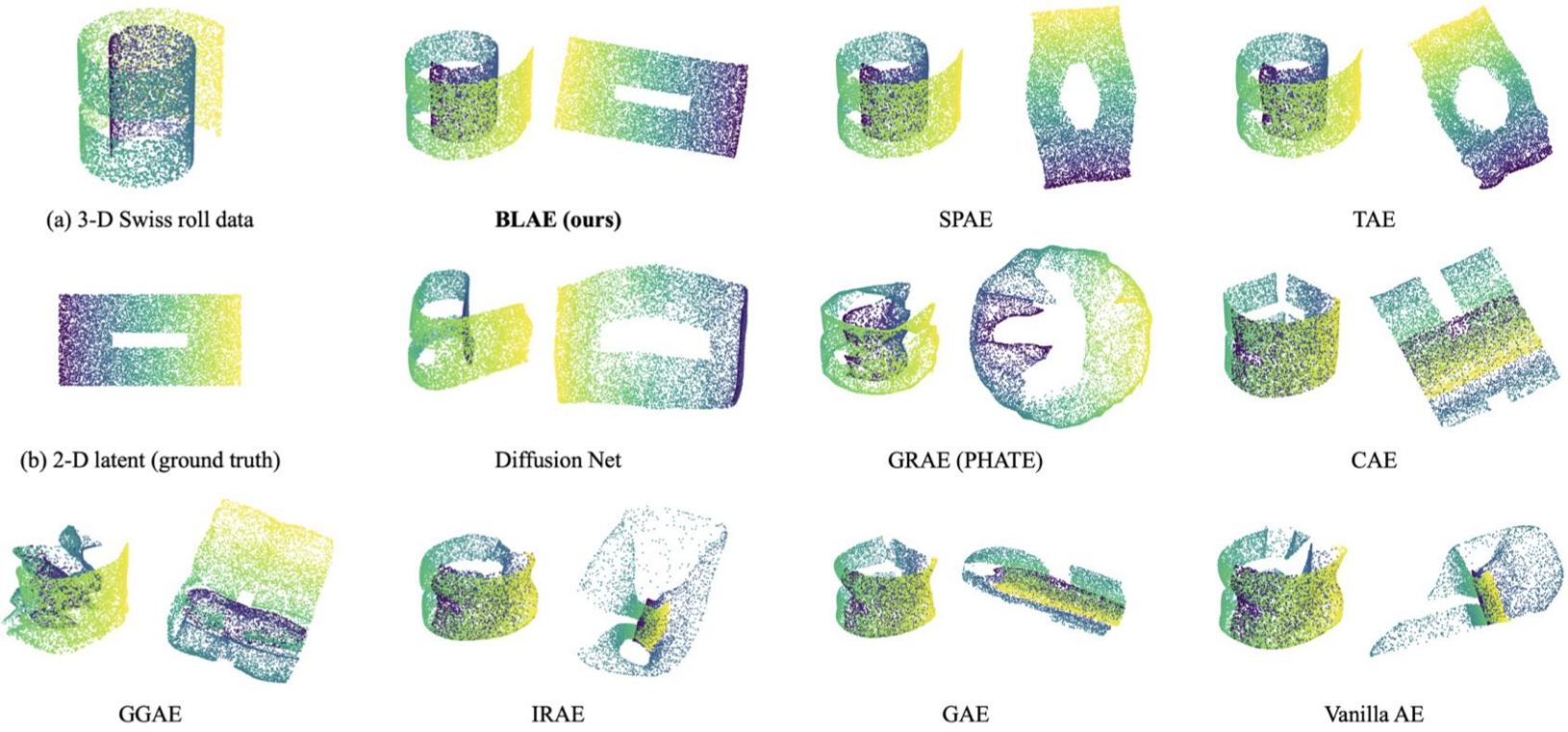
>> Least geometric distortion across clusters

MNIST (Rotated)

Distribution shift test: uniform vs. non-uniform rotation sampling. Ground truth: concentric circles in latent space.

>> Consistent rings across both distributions

Experiments: BLAE Outperforms 9 Baselines



Conclusion

1 **Diagnosed the Core Problem**

Non-injectivity of encoder creates pathological local minima — limiting convergence and representation quality

2 **Admissible Regularization Theory**

Formalized conditions for distribution-robust regularization: valid regardless of how data is sampled

3 **Two Novel Innovations**

(δ,ϵ)-separation injective loss + bi-Lipschitz relaxation for geometry-preserving, robust embeddings

4 **State-of-the-Art Results**

BLAE ranks #1 across all metrics on Swiss Roll, dSprites, and MNIST distribution-shift experiments