Exploring Distributional Learning of Synthetic Data Generators for Manifolds

Sonakshi Garg, Vicenc Torra

sgarg@cs.umu.se

Department of Computing Science

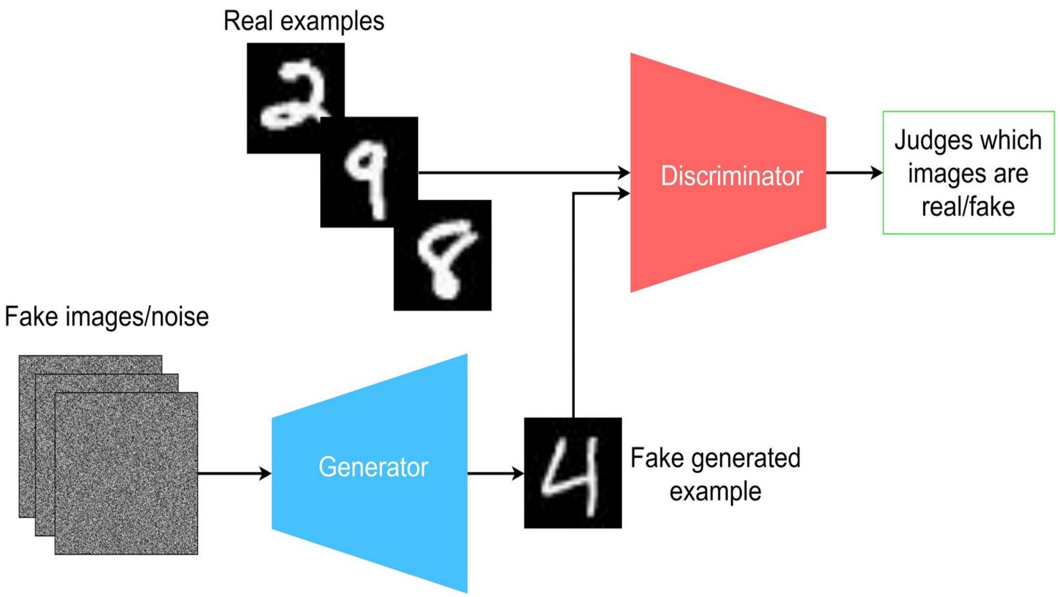
Umeå University, Sweden

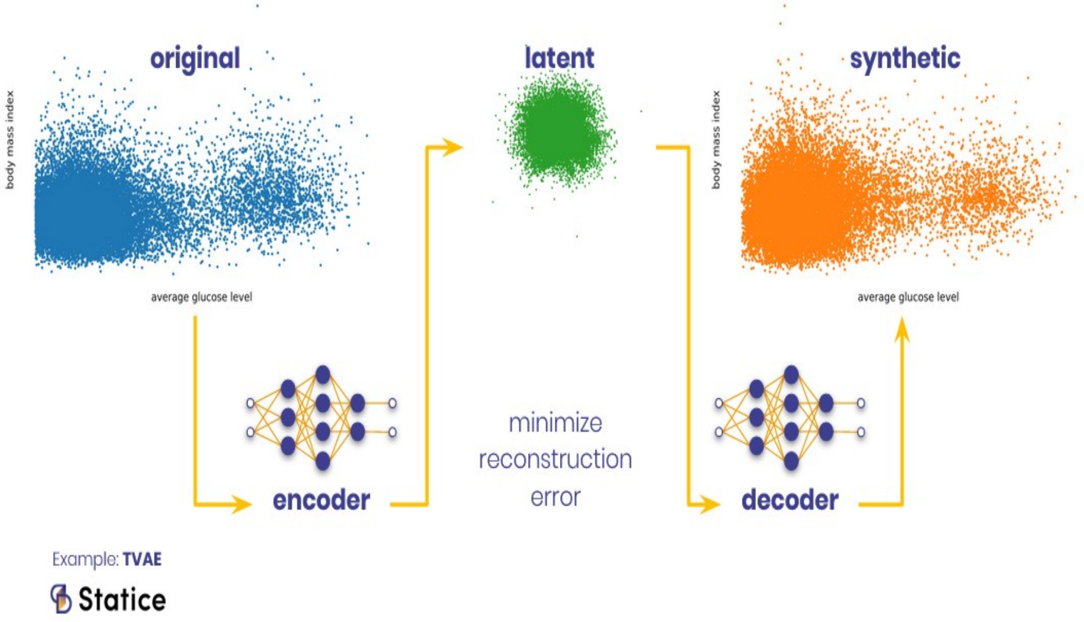
DPM 2024

Introduction

- Data may contain sensitive information that must be safeguarded from disclisure, following GDPR policies.
- Goal is to produce valid data mining results while protecting privacy.
- Privacy Models such as *k-Anonymity, Differential Privacy*
- Alternative: Synthetic Data Generation, preserves global properties without revealing individual identities.
- Mimics the properties of original data, substitutes sensitive data with synthetic data.
- Goal: Evaluate Distribution Learning Capabilities of Synthetic Data Generators

Generation of Synthetic Data- GAN & VAE





Generative Adversarial Network

Learn from random noise as input, and generate realistic copies of original data as the training progresses.

Variational AutoEncoder

Autoencoder is a neural network that converts high dimensional input into the latent vector and converts the latent vector back to the input with the highest possible quality.

Challenges

GANs and VAEs are black-box in nature due to complex learning mechanism

Visualization and understanding difficult for high-dimensional datasets

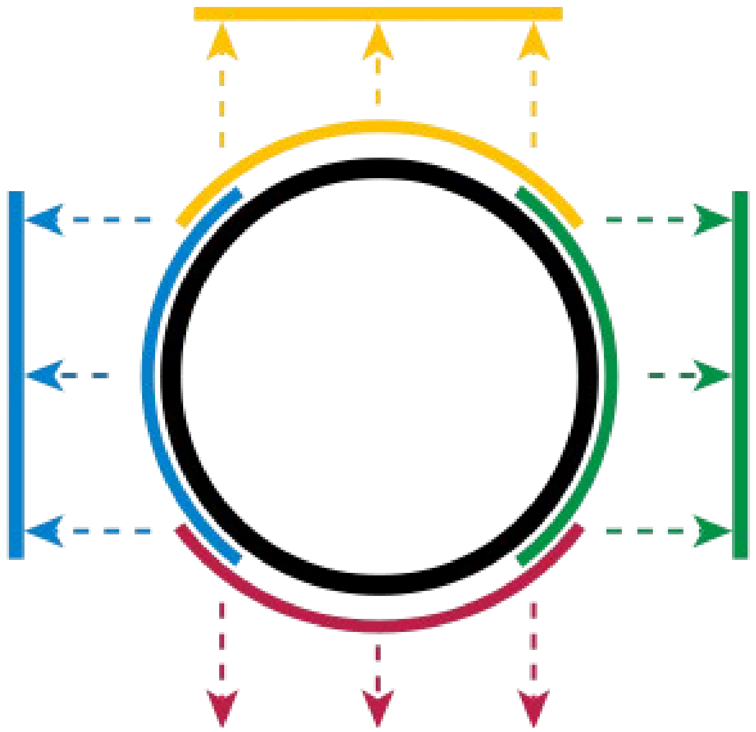
How well do GANs and VAEs capture complex data distributions?

Motivation

- Assess the effectiveness of GAN & VAE in learning data distributions
- Determine whether the manifolds generated using synthetic data generators converge to real data manifolds.
- GANs have demonstrated impressive results on certain datasets, but limitations on others, such as ImageNet . The intricate distribution of natural images poses challenges for GAN.
- **Datasets**: Artificially generated datasets (Swish Roll, S-Curve) and point datasets with discontinuities, MNIST dataset.

How to handle high-dimensional data: Manifold

- A topological space that locally resembles Euclidean space.
- Take a geometric object from $\mathbb{R}^k$ and try to fit it into $\mathbb{R}^n$, $n>k$.
- Eg. of a 1D manifold: Embed a line segment in 2D.



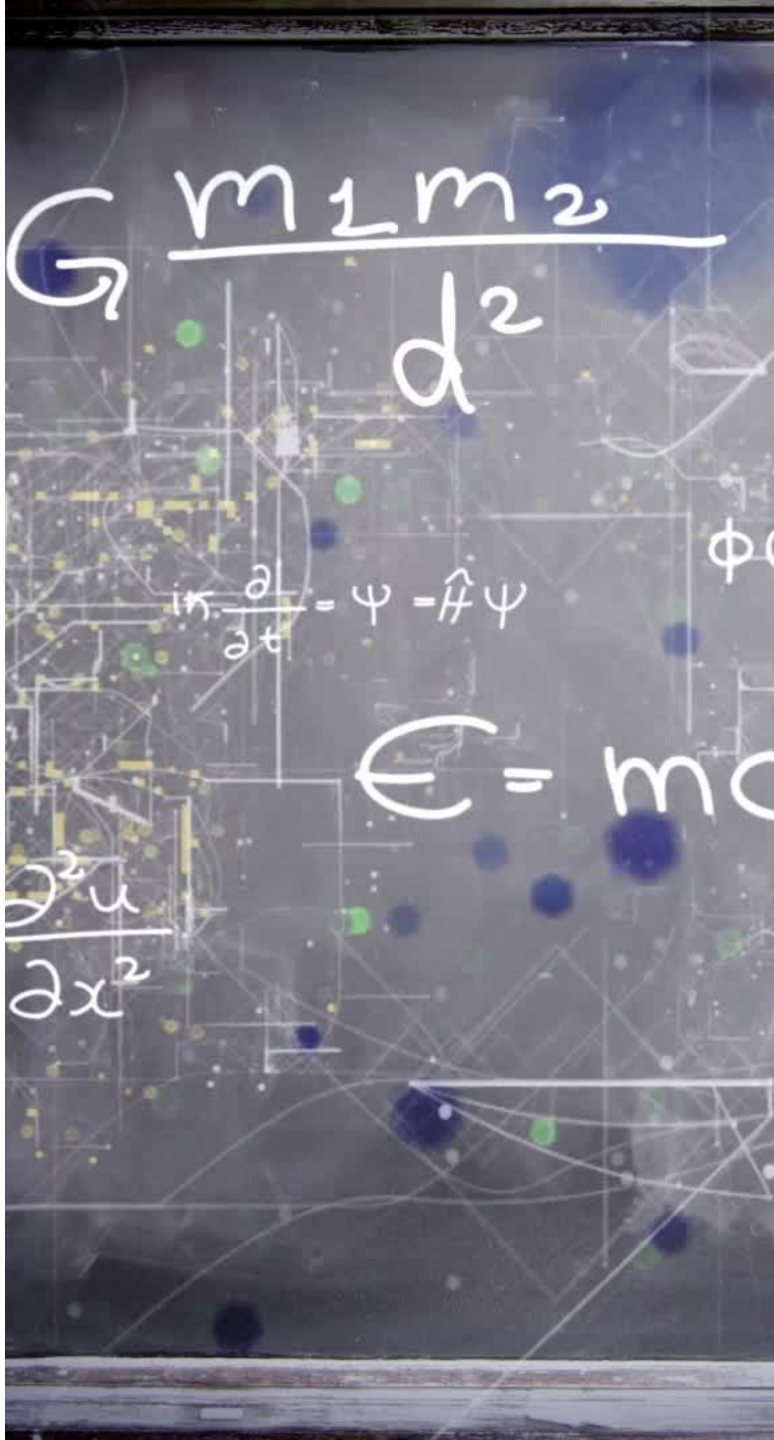
Manifold Hypothesis

- Any real-world high-dimensional data lie on low-dimensional manifolds embedded within the high-dimensional space.

- Manifold Learning techniques: UMAP, t-SNE, LLE etc

Why do we need a manifold?

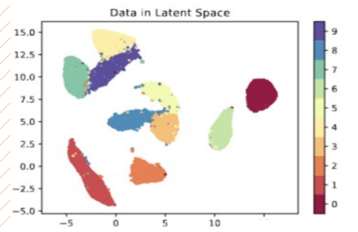
- More complicated structures are expressed and understood in simpler spaces.
- Additional structures are often defined on manifolds.
- Eg. Differentiable manifolds on which one can do calculus, Riemannian manifolds on which distances and angles can be defined, etc.



Methodology

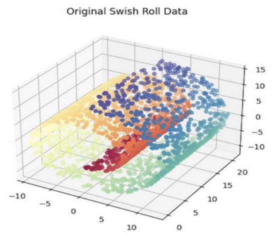
Dataset Selection

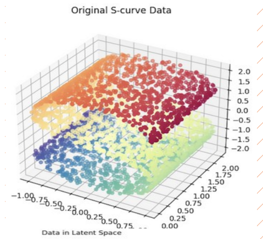
Train UMAP

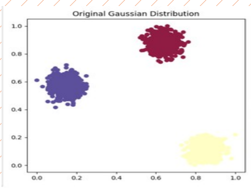


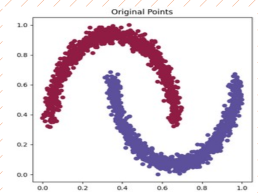
Reconstruct to Original Space

Generation of Artificial Data

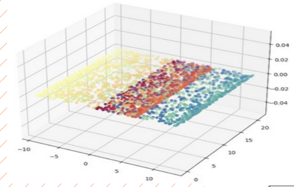




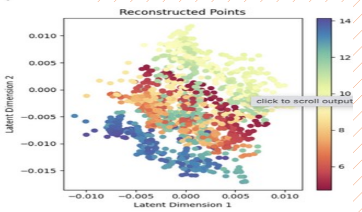




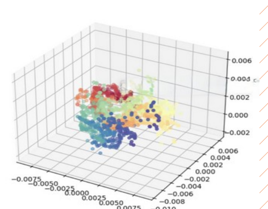
Test UMAP



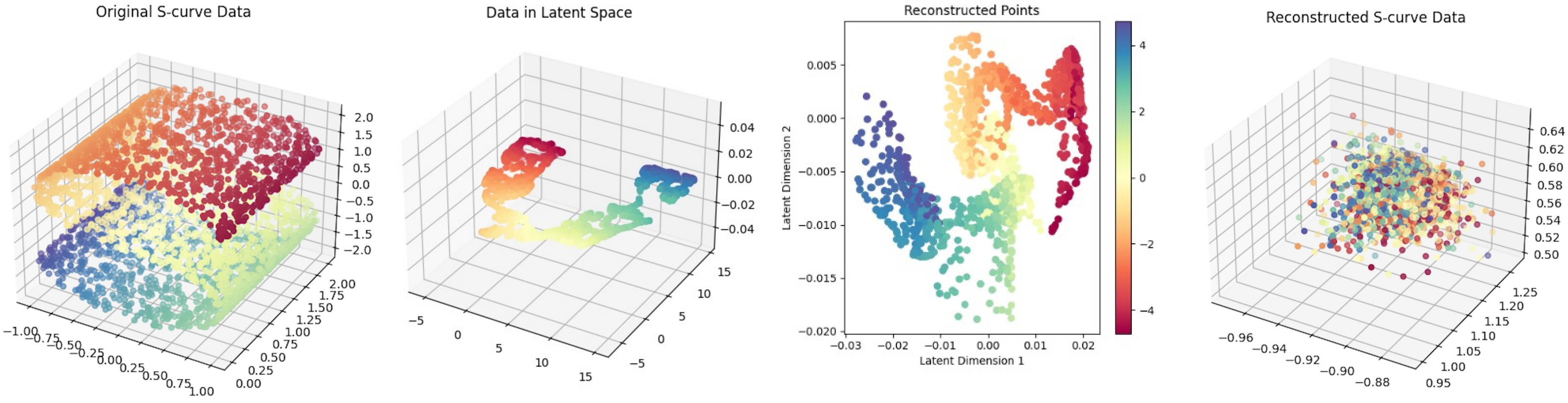
Synthetic Data Generation



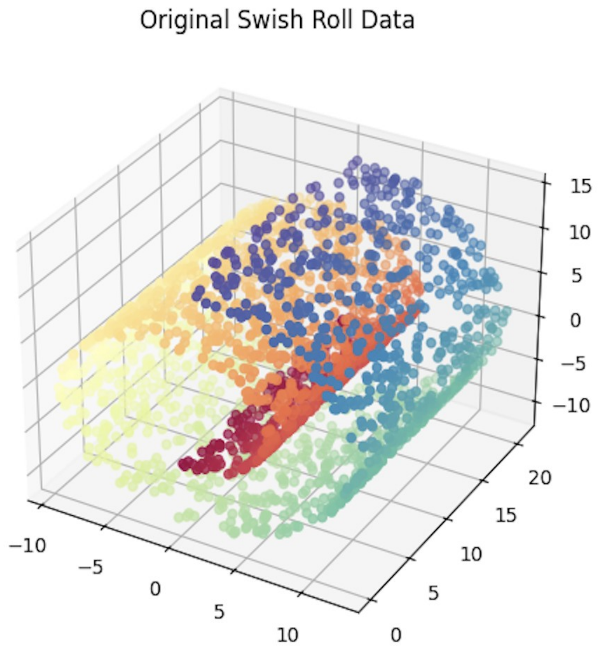
Reconstruct Synthetic Data

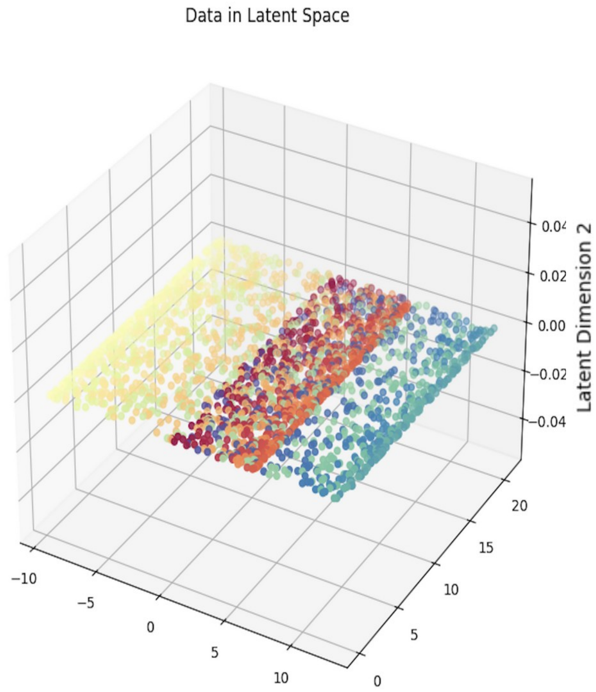


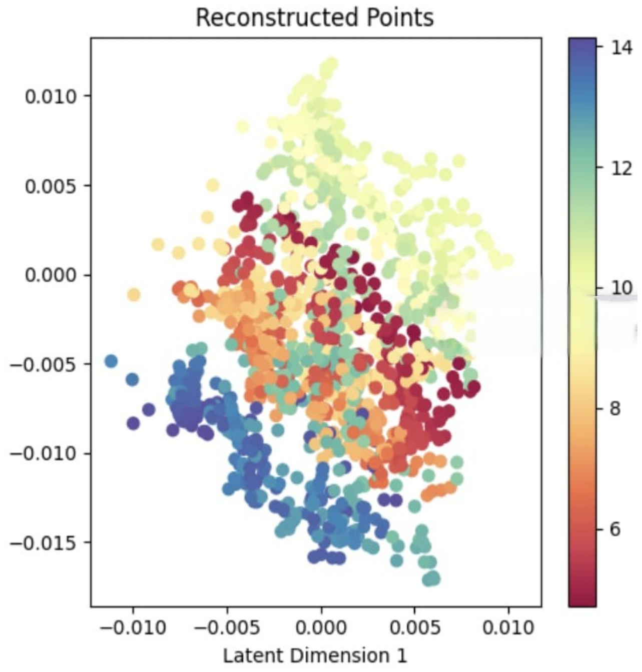
Visualizing Synthetic Generation from S-Curve Transformation

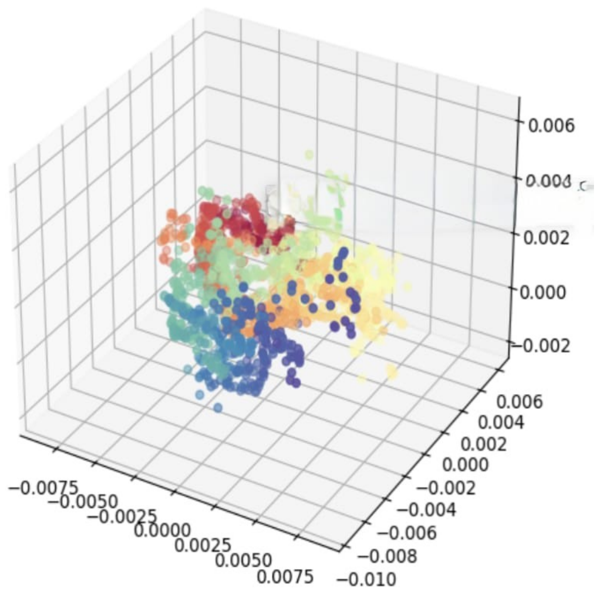


Unrolling the Swish Roll: Exploring Manifold Transformation

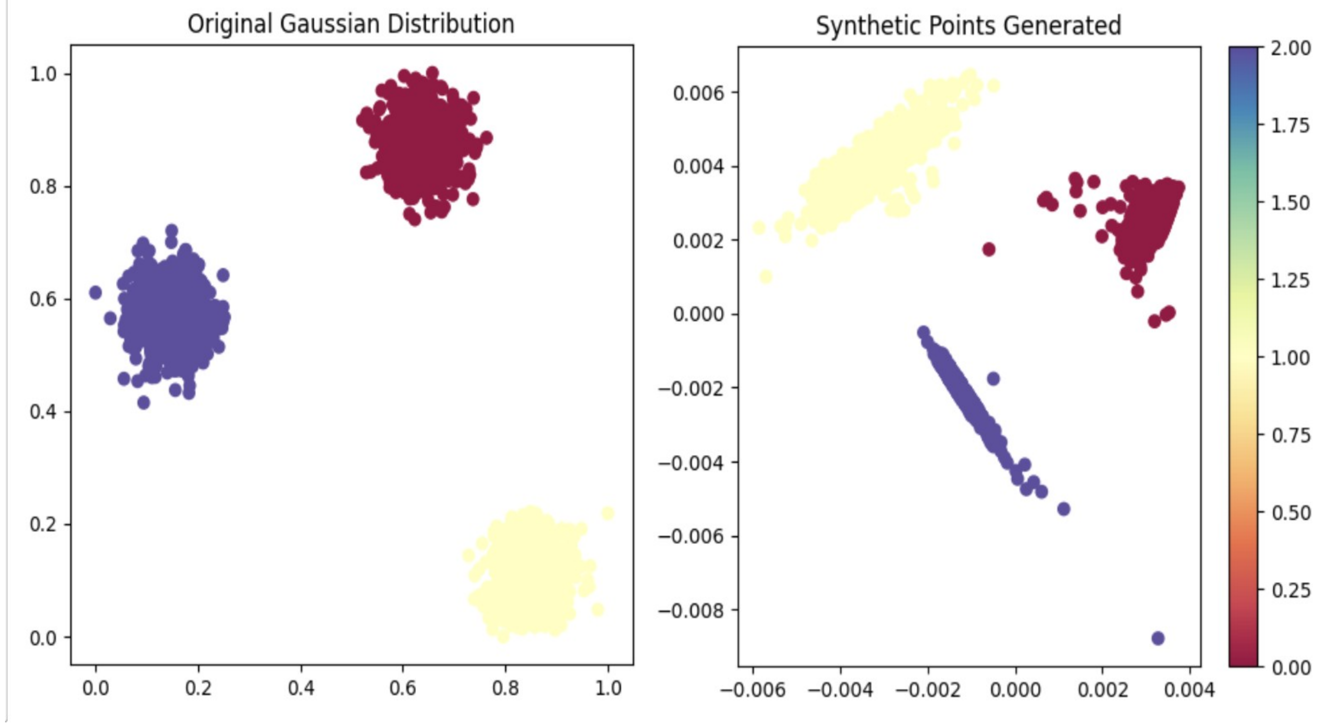




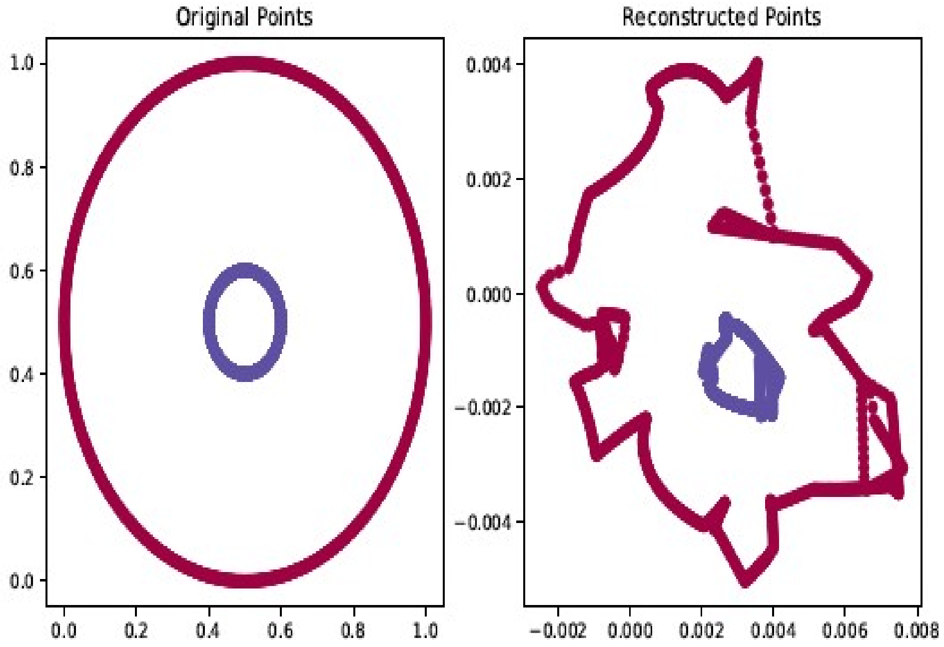




Understanding 2D Point Datasets

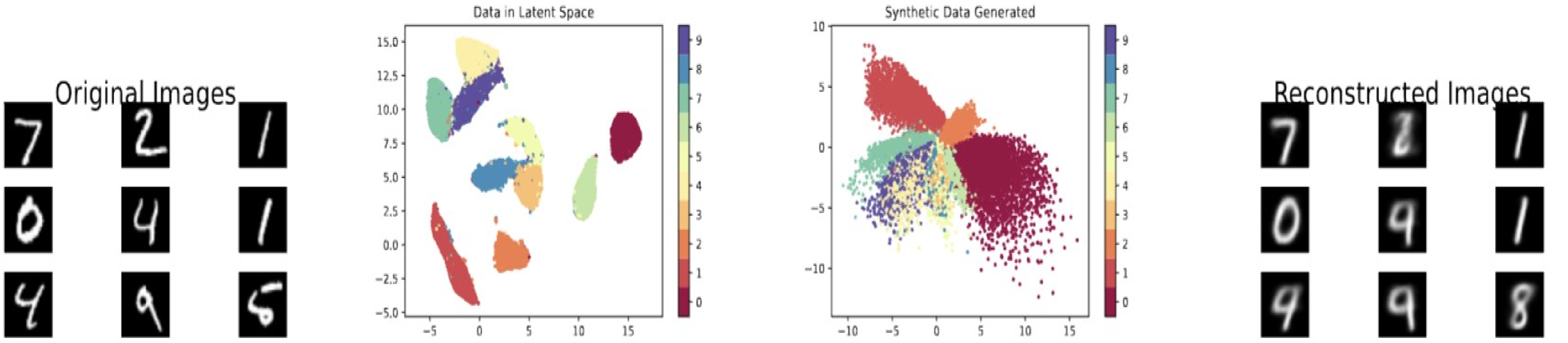


Concentric Circles



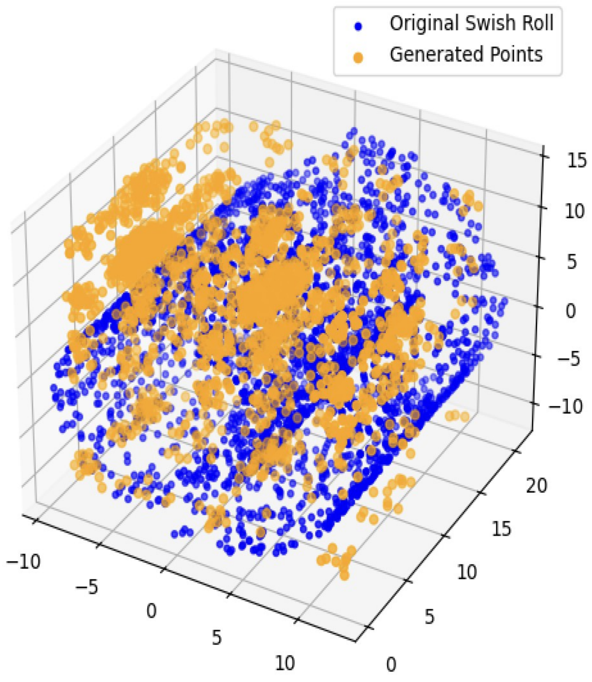
Two- Half Circles

Visualizing Real-World Data

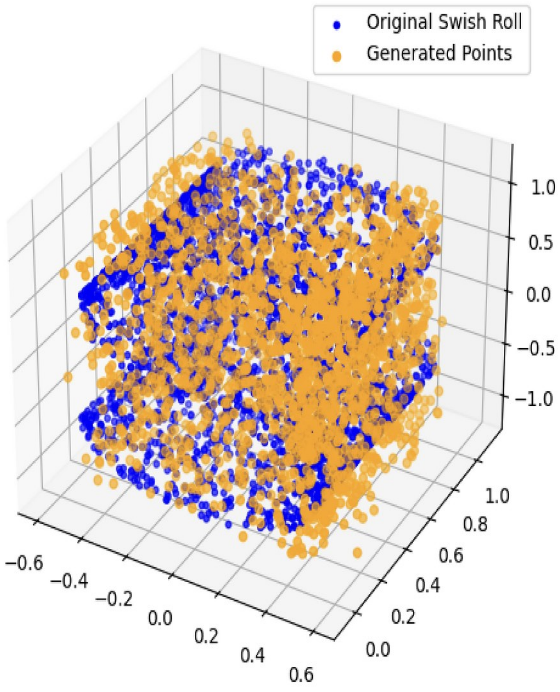


(a) Original Data (b) Data transformed (c) Synthetic Data (d) Reconstructed Data

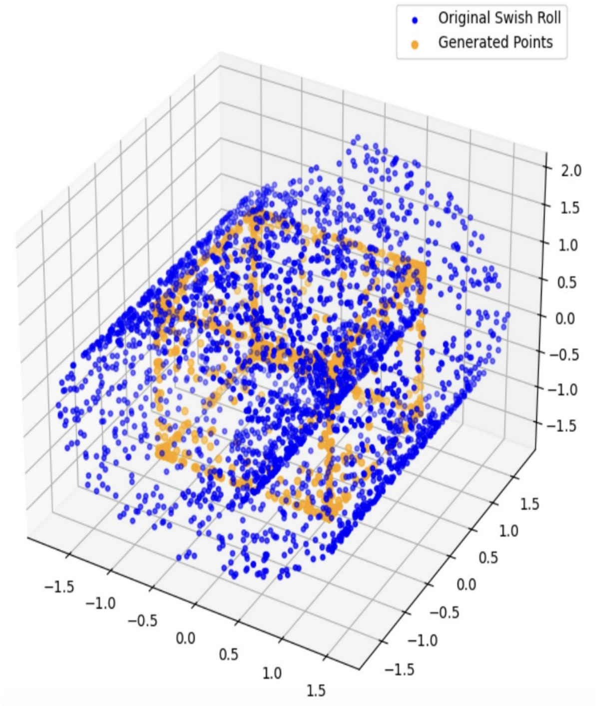
Visualization with Diverse GAN Architectures



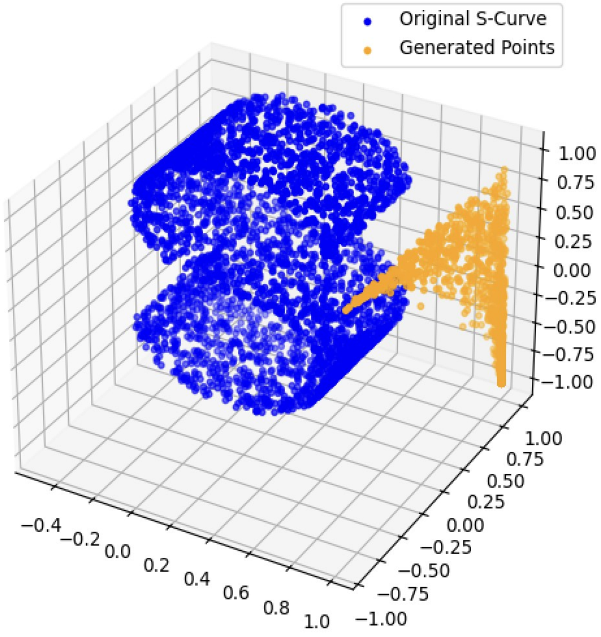
DPGAN



CTGAN

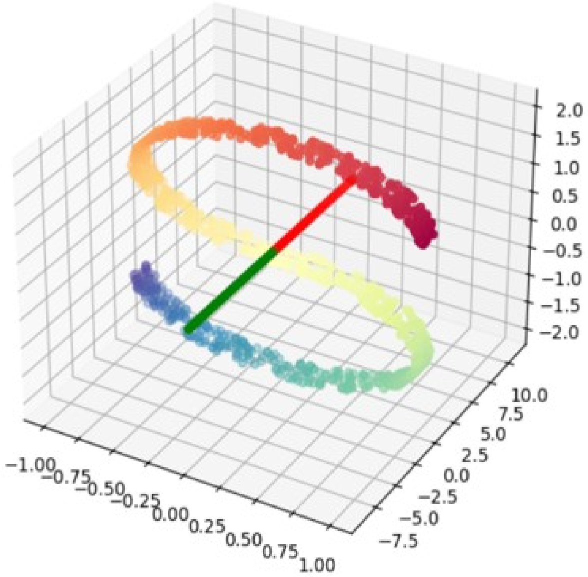


DCGAN

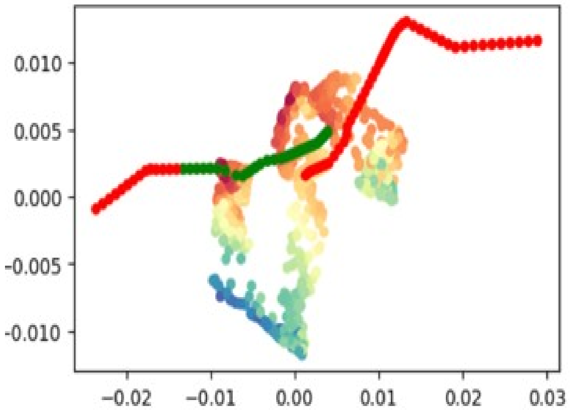


VGAN

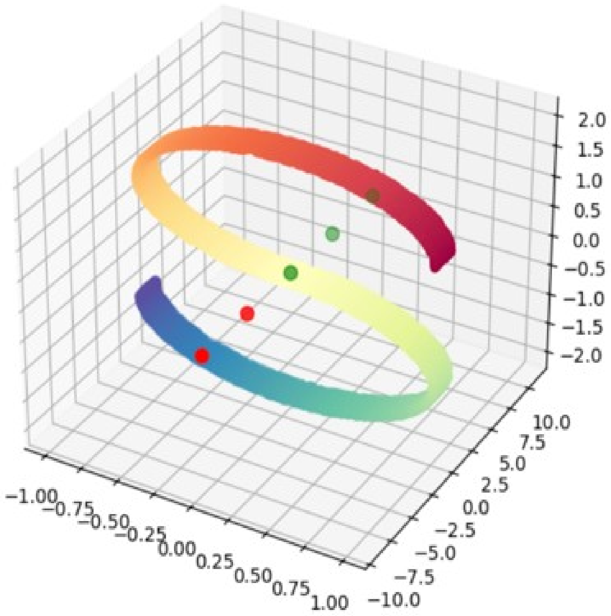
Privacy Risk Assessment



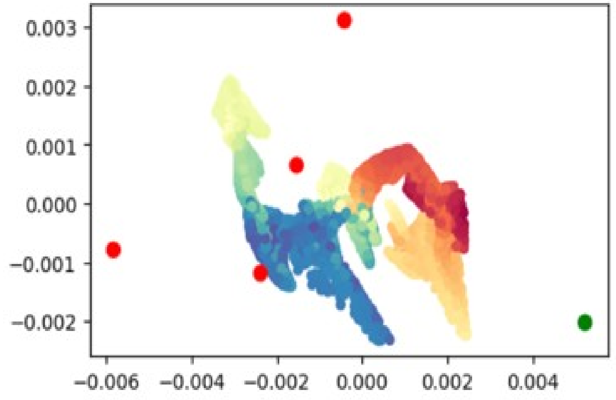
(a) 10% additional points arranged along a straight line



(b) VAE accurately predicting those points



(c) Only 0.01% of points are newly added and strategically placed



(d) VAE did not memorize the specific data samples but learned general patterns instead.

Summary

- GANs: High instability, requires complex optimization, struggle with certain distributions
- VAEs: Better performance in capturing data distribution and structure
- VAE demonstrate a superior ability to understand and learn the intrinsic structure of our artificial point dataset compared to GAN.
- Enhanced understanding of privacy-preserving methods for data generation
- Future Work: Improve GAN training and inverse-transformation of manifold techniques

Thank You!