A Appendix

A.1 Derivation of Local Evidence Lower Bound (Eq. 5)

We begin with taking the log of the random walk transition likelihood,

$$\log p_\theta(x'|x) = \log \int_z' p_\theta(x', z'|x) dz' \tag{A.1}$$

$$= \log \int_z p_\theta(x'|z', x) p(z'|x) \frac{q(z')}{q(z')} dz' \tag{A.2}$$

$$= \log \mathbb{E}_{z' \sim q(z')} \left[p_\theta(x'|z', x) \frac{p(z'|x)}{q(z')} \right] \tag{A.3}$$

$$\geq \mathbb{E}_{z' \sim q(z')} \left[\log p_\theta(x'|z', x) \right] + \mathbb{E}_{z' \sim q(z')} \left[\log \frac{p(z'|x)}{q(z')} \right] \tag{A.4}$$

$$\geq \mathbb{E}_{z' \sim q(z')} \left[\log p_\theta(x'|z', x) \right] + D_{KL}[q(z')||p(z'|x)] \tag{A.5}$$

where $q(z')$ is an arbitrary distribution. We let $q(z')$ to be the conditional distribution $q(z'|x)$. Furthermore, if we make the simplifying assumption that $p_\theta(x'|z', z) = p_\theta(x'|z')$, then we obtain Eq. 5

$$\log p_\theta(x'|x) \geq -D_{KL}(q_\phi(z'|x)||p_\theta(z'|x)) + \mathbb{E}_{z' \sim q_\phi(z'|x)} \log p_\theta(x'|z'). \tag{A.6}$$

A.2 Results in [17]

To state the result in [17], we need the following set-up:

(C1) $\mathcal{M}$ is a d-dimensional smooth compact manifold, possibly having boundary, equipped with a smooth (at least C^2) Riemannian metric g;

We denote the geodesic distance by $d_{\mathcal{M}}$, and the geodesic ball centering at x with radius r by $B_{\mathcal{M}}(x, r)$. Under (C1), for each point $x \in \mathcal{M}$, there exists $r_{\mathcal{M}}(x)$ which is the inradius, that is, r is the largest number s.t. $B_{\mathcal{M}}(x, r)$ is contained $\mathcal{M}$.

Let $\triangle_{\mathcal{M}}$ be the Laplacian-Beltrami operator on $\mathcal{M}$ with Neumann boundary condition, which is self-adjoint on $L^2(M, \mu)$, μ being the Riemannian volume given by g. Suppose that $\mathcal{M}$ is re-scaled to have volume 1. The next condition we need concerns the spectrum of the manifold Laplacian

(C2) $\triangle_{\mathcal{M}}$ has discrete spectrum, and the eigenvalues $\lambda_0 \leq \lambda_1 \leq \cdots$ satisfy the Weyl's estimate, i.e. exists constant C which only depends on $\mathcal{M}$ s.t.

$$|\{j : \lambda_j \leq T\}| \leq C T^{d/2}.$$

Let ψ_j be the eigenfunction associated with λ_j, $\{\psi_j\}_j$ form an orthonormal bases of $L^2(M, \mu)$. The last condition is

(C3) The heat kernel (defined by the heat equation on $\mathcal{M}$) has the spectral representation as

$$K_t(x, y) = \sum_{j=0}^{\infty} e^{-t\lambda_j} \psi_j(x) \psi_j(y).$$

Theorem 3 (Theorem 2 [17], simplified version) *Under the above setting and assume (C1)-(C2), then there are positive constants c_1, c_2, c_3 which only depend on $\mathcal{M}$ and g, s.t. for any $x \in \mathcal{M}$, $r_{\mathcal{M}}(x)$ being the inradius, there are d eigenfunctions of $\triangle_{\mathcal{M}}$, $\psi_{j_1}, \cdots, \psi_{j_d}$, which collectively give a mapping $\Psi : \mathcal{M} \to \mathbb{R}^d$ by*

$$\Psi_x(x) = (\psi_{j_1}(x), \cdots, \psi_{j_d}(x))$$

satisfying that $\forall y, y' \in B(x, c_1 r_{\mathcal{M}}(x))$,

$$c_2 r_{\mathcal{M}}(z)^{-1} d_{\mathcal{M}}(y, y') \leq \|\Psi_x(y) - \Psi_x(y')\| \leq c_3 r_{\mathcal{M}}(z)^{-1-d/2} d_{\mathcal{M}}(y, y').$$

That is, Ψ is bi-Lipschitz on the neighborhood $B(x, c_1 r_{\mathcal{M}}(x))$ with the Lipschitz constants indicated as above. The subscript x in Ψ_x emphasizes that the indices $j_1, \cdots, j_d$ may depend on x.

A.3 Proofs

Proof (of Theorem 1). The proof of Theorem 1 is actually a simple extension of the following theorem, Theorem 4, which needs to be proved for each individual extrinsic coordinate X_k, hence the additional factor of m coming from the $L2$ norm of m functions.

Theorem 4 *Let $\mathcal{M} \subset \mathbb{R}^m$ be a smooth d-dimensional manifold, $\psi(\mathcal{M}) \subset \mathbb{R}^D$ be the diffusion map for $D \geq d$ large enough to have a subset of coordinates that are locally bi-Lipschitz. Let one of the m extrinsic coordinates of the manifold be denoted $X(\psi(x))$ for $x \in \mathcal{M}$. Then there exists a sparsely-connected ReLU network f_N, with $4DC_{\mathcal{M}}$ nodes in the first layer, $8dN$ nodes in the second layer, and $2N$ nodes in the third layer, such that*

$$\|X - f_N\|_{L^2(\psi(\mathcal{M}))} \leq \frac{C_\psi}{\sqrt{N}} \tag{A.7}$$

where C_ψ depends on how sparsely $X(\psi(x))\big|_{U_i}$ can be represented in terms of the ReLU wavelet frame on each neighborhood U_i, and $C_{\mathcal{M}}$ on the curvature and dimension of the manifold $\mathcal{M}$.

Proof (of Theorem 4).

The proof borrows from the main theorem of [39]. We adopt this notation and summarize the changes in the proof here. For a full description of the theory and guarantees for neural networks on manifolds, see [39]. Let $C_{\mathcal{M}}$ be the number of neighborhoods $U_i = B(x_i, \delta) \cap \mathcal{M}$ needed to cover $\mathcal{M}$ such that $\forall x, y \in U_i$, $(1-\epsilon)\|x-y\| \leq d_{\mathcal{M}}(x, y) \leq (1+\epsilon)\|x-y\|$. Here, we choose $\delta = \min(\delta_{\mathcal{M}}, \kappa^{-1}\rho)$ where $\delta_{\mathcal{M}}$ is the largest δ that preserves locally Euclidean neighborhoods and $\kappa^{-1}\rho$ is the smallest value from [17] such that every neighborhood U_i has a bi-Lipschitz set of diffusion coordinates.

Because of the locally bi-Lipschitz guarantee from [17], we know for each U_i there exists an equivalent neighborhood $\widetilde{\psi}(U_i)$ in the diffusion map space, where

$\widetilde{\psi}(x) = \big[\psi_{i_1}(x), ..., \psi_{i_d}(x)\big]$. Note that the choice of these d coordinates depends on the neighborhood U_i. Moreover, we know the Euclidean distance on $\psi(U_i)$ is locally bi-Lipschitz w.r.t. $d_{\mathcal{M}}(\cdot,\cdot)$ on U_i.

First, we note that as in [39], the first layer of a neural network is capable of using $4D$ units to select the subset of d coordinates $\widetilde{\psi}(x)$ from $\psi(x)$ for $x \in U_i$ and zeroing out the other $D-d$ coordinates with ReLU bump functions. Then we can define $X(\widetilde{\psi}(x)) = X(\psi(x))$ on $x \in U_i$.

Now to apply the theorem from [39], we must establish that $X\big|_{U_i} : \widetilde{\psi}(U_i) \to \mathbb{R}$ can be written efficiently in terms of ReLU functions. Because of the manifold and diffusion metrics being bi-Lipschitz, we know at a minimum that $\widetilde{\psi}$ is invertible on $\widetilde{\psi}(U_i)$. Because of this invertibility, we will slightly abuse notation and refer to $X(\psi(x)) = X(x)$, where this is understood to be the extrinsic coordinate of the manifold at the point x that cooresponds to $\psi(x)$. we also know that $\forall x, y \in U_i$,

$$\begin{aligned}
|X(\widetilde{\psi}(x)) - X(\widetilde{\psi}(y))| &= |X(x) - X(y)| \\
&\leq \max_{z \in U_i} \|\nabla X(z)\| d(x,y) \\
&\leq \frac{\max_{z\in U_i} \|\nabla X(z)\|}{1-\epsilon} \|\widetilde{\psi}(x) - \widetilde{\psi}(y)\|,
\end{aligned}$$

where $\nabla X(z)$ is understood to be the gradient of $X(z)$ at the point $z \in \mathcal{M}$. This means $X(\widetilde{\psi}(x))$ is a Lipschitz function w.r.t. $\widetilde{\psi}(x)$. Because $X(\widetilde{\psi}(x))$ Lipschitz continuous, it can be approximated by step functions on a ball of radius $2^{-\ell}$ to an error that is at most $\frac{\max_{z\in U_i} \|\nabla X(z)\|}{1-\epsilon} 2^{-\ell}$. This means the maximum ReLU wavelet coefficient is less than $\frac{\max_{z\in U_i} \|\nabla X(z)\|}{1-\epsilon}(2^{-\ell} + 2^{-\ell+1})$. This fact, along with the fact that $\widetilde{\psi}(U_i)$ is compact, gives the fact that on $\widetilde{\psi}(U_i)$, set of ReLU wavelet coefficients is in ℓ^1. And from [39], if on a local patch the function is expressible in terms of ReLU wavelet coefficients in ℓ^1, then there is an approximation rate of $\frac{1}{\sqrt{N}}$ for N ReLU wavelet terms.

Proof (of Theorem 2). We borrow from [41] to prove the following result. Given that the bulk of the distribution q lies inside $\psi(U_{z_0})$, we can consider only the action of f_N on $\psi(U_{z_0})$ rather than on the whole space. Because the geodesic on U is bi-Lipschitz w.r.t. the Euclidean distance on the diffusion coordinates (the metric on the input space), we can use the results from [41] and say that on $\psi(U_{z_0})$ the output covariance matrix is characterized by the Jacobian of the function f_N mapping from Euclidean space (on the diffusion coordinates) to the output space, at the point z_0. So the covariance of the data lying insize $\psi(U_{z_0})$ is $J_{z_0} \Sigma J_{z_0}^T$, with an $O(\epsilon)$ perturbation for the fact that ϵ fraction of the data lies outside $\psi(U_{z_0})$.

The effective rank of C being at most d comes from the locally bi-Lipschitz property. We know $X(\psi(x))$ only depends on the d coordinates $\widetilde{\psi}(x)$ as in the proof of Theorem 1, which implies $f_N(\psi(x))$ satisfies a similarly property if f_N fully learned $X(\psi(x))$. Thus, while $J \in \mathbb{R}^{m\times D}$, it is at most rank d, which means $J\Sigma J^T$ is at most rank d as well.

A.4 Spectral Net

A.5 Additional Experimental Result

To evaluate the quality of the generated images in the Bulldog dataset, we use the Frechet inception distance (FID). We train the different generative models 5 times and compute the FID between source and generated images. In table A.1 we present the mean and standard deviations of the FID.

FID	GAN	VAE	SVAE	VDAE
Bulldog	264.4(18.4)	245.7(14.7)	400.6 (6.2)	144.3(12.6)

Table A.1: Frechet inception distance (FID) on the Bulldog dataset, mean and standard deviation.

MMD	GAN	VAE	SVAE	VDAE
Circle	9.3(11.1)	8.3(4.4)	8.1 (4.2)	7.3(4.3)
Torus	12.3 (4.7)	63.3 (12.9)	84.5(11.7)	41.9 (4.1)
Bunny	175.6(68.6)	725.8(3.8)	601.7(41.1)	3.6(0.3)
Bulldog	741.8(88)	167.3(16.4)	213.7(13.1)	9.68(3.44)
Frey	34.9(5.1)	39.3(6.1)	29.4	47.0
MNIST	3.5(0.6)	27.9(1)	20.6(1.2)	5.79(0.3)
COIL-20	3.3(0.9)	39.2(9.6)	55.7(4.7)	7.4(1.07)

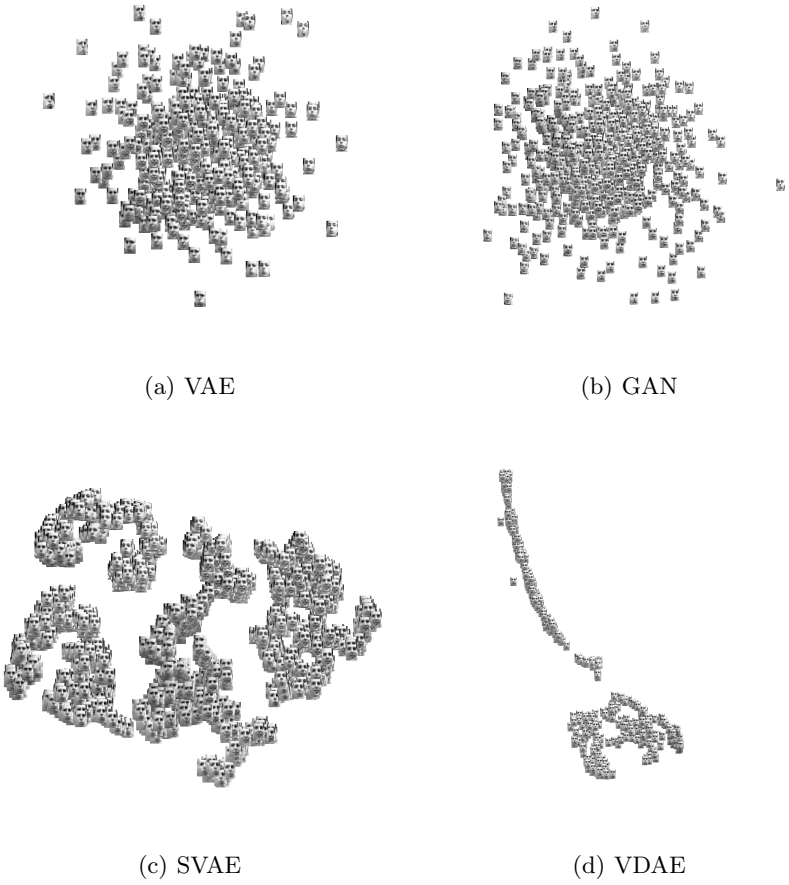
Table A.2: Measures of similarity between training data and generated data using Maximum Mean Discrepancy. Comparisons are across a variety of synthetic and real data sets

A.6 Experimental Architectures

For the circle, torus, Stanford bunny, Frey faces [8], and the 5x5 spherical density datasets, we used a single 500-unit hidden layer network for all models used in the paper (i.e. decoder, encoder, generator, discriminator, for the VAE, Wasserstein GAN, hyperspherical VAE, and our method).

As higher dimensional datasets, we used a slightly larger architecture for the MNIST, COIL-20, and rotating bulldog datasets: two hidden-layer decoder/generators of width 1024 and 2048, and two hidden-layer encoder/discriminators of width 2048 and 1024. All activations are still ReLU.

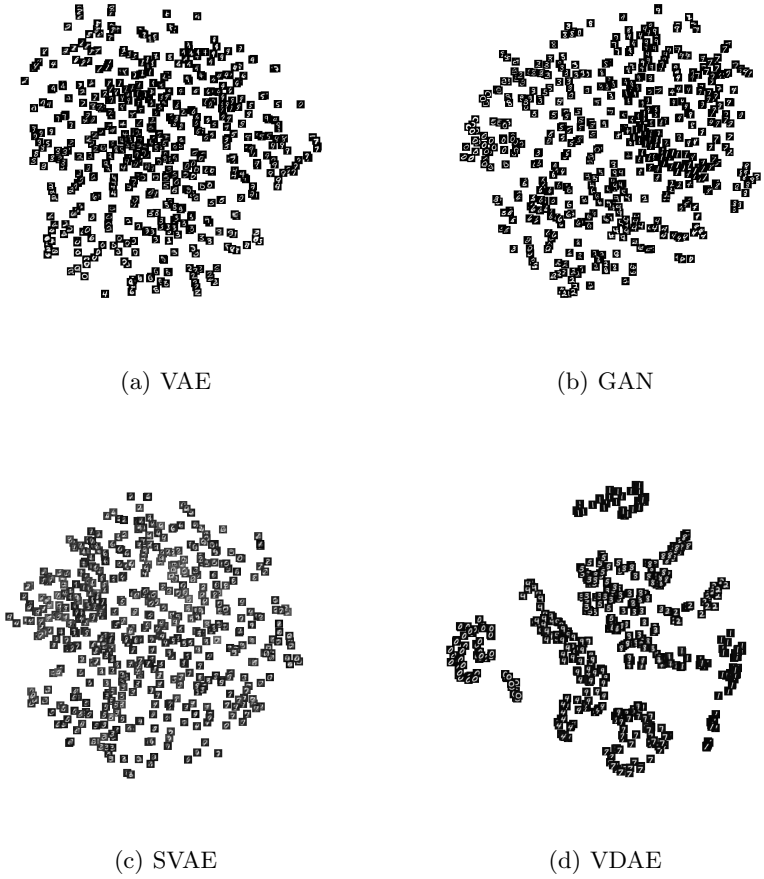
[8] https://cs.nyu.edu/ roweis/data.html



(a) VAE (b) GAN

(c) SVAE (d) VDAE

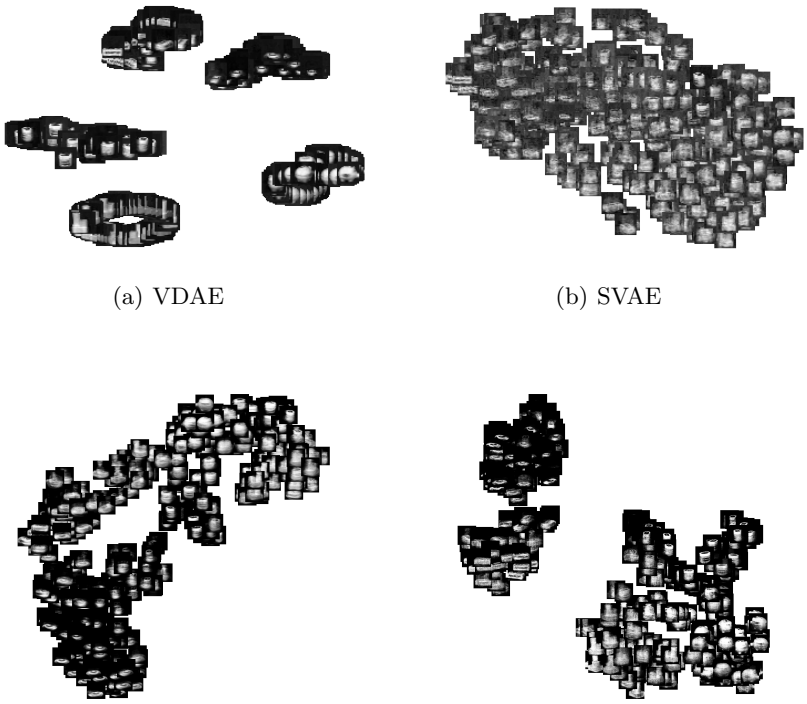
Fig. A.1: A tSNE plot of generated images from Frey data set. While the images from the VAE and GAN are compelling, they do not capture the geometric structure of the Frey faces dataset. This structure is much more apparent in the images generated by SVAE and VDAE. In particular, the VDAE has captured a linear structure in the data, which reflects the fact that the dataset was created from a video.



(a) VAE (b) GAN

(c) SVAE (d) VDAE

Fig. A.2: A tSNE plot of generated images from Frey data set. Like with Fig. A.1 (Frey faces), the images generated by VAE, GAN, and SVAE have a unimodal distribution that does not capture the clustered structure of the MNIST dataset. VDAE, on the other hand, organizes the digits into clear clusters, and does not generate from regions where there is low support in the training distribution.



(a) VDAE (b) SVAE

(c) β-VAE (d) WGAN

Fig. A.3: A tSNE embedding of 360 generated images from COIL-20 data set.