Supplementary for Isometric Manifold Learning using Hierarchical Flow

1. Proofs and Analysis

In this section, we provide detailed proofs and analysis for our theoretical claims for the properties of *distance preserving* and *rigorous projection* introduced in main text.

1.1. The *Distance Preserving* Property

We provide proofs for Prop. 1 (restated as Prop. 1) and Prop. 3 (restated as Prop. 2) in main text for satisfying the *distance preserving* property.

Proposition 1 (*Distance preserving* property)**.** *Assume that $\mathcal{U}^1(\mathcal{M})$ is a convex set, and $v^i(x) = 0$ for $\forall i \in [L], x \in \mathcal{M}$. The length of the geodesic between $g(u_1)$ and $g(u_2)$ on $\mathcal{M}$ for $\forall u_1, u_2 \in \mathcal{U}^1(\mathcal{M})$ equals to $\|u_1 - u_2\|_2$, if $J_g(u)$ is orthonormal for $\forall u \in \mathcal{U}^1(\mathcal{M})$, where $J_g(u)$ is the Jacobian matrix of the generator g at u.*

Proof. According to Riemannian geometry [12], given a Riemannian manifold $\mathcal{M}$, a smooth curve $\gamma : [a, b] \to \mathcal{M}$ on $\mathcal{M}$ is a geodesic of $\mathcal{M}$, if $\nabla_{\gamma'(t)}\gamma'(t) \equiv 0, \forall t \in [a, b]$. Since $v^i(x) = 0, \forall i \in [L], x \in \mathcal{M}$, we know that $g(u) \in \mathcal{M}, \forall u \in \mathcal{U}^1(\mathcal{M})$. Therefore, we can learn that $\gamma(t) \triangleq g(tu_2 + (1-t)u_1)$ is a curve on $\mathcal{M}$ for $t \in [0, 1]$. We firstly prove that such a curve $\gamma : [0, 1] \to \mathcal{M}$ is a geodesic connecting $g(u_1)$ and $g(u_2)$ on $\mathcal{M}$. Since $\gamma' = (u_2 - u_1)^i \frac{\partial}{\partial z^i}$[1], where $(u_2 - u_1)^i$ is the i-th element of vector $u_2 - u_1 \in \mathbb{R}^K$, we further have that

$$\nabla_{\gamma'}\gamma' = \nabla_{(u_2-u_1)^j \frac{\partial}{\partial z^j}} (u_2 - u_1)^i \frac{\partial}{\partial z^i} = \xi^j \xi^i \nabla_{\frac{\partial}{\partial z^j}} \frac{\partial}{\partial z^i}, \tag{1}$$

where we use $\xi \triangleq u_2 - u_1$. It is known that the coefficient Γ_{ij}^k of the connection ∇ is given by the *Christoffel symbol*

$$\Gamma_{ij}^k = \frac{1}{2}\mathfrak{g}^{kl}\left(\frac{\partial \mathfrak{g}_{il}}{\partial z^j} + \frac{\partial \mathfrak{g}_{jl}}{\partial z^i} - \frac{\partial \mathfrak{g}_{ij}}{\partial z^l}\right), \tag{2}$$

where $\mathfrak{g}_{ij}$ and $\mathfrak{g}^{ij}$ are the covariant and contravariant components of the Riemannian metric $\mathcal{G}$, respectively, and the matrix form of $\mathcal{G}$ is induced by $J_g^\top J_g$, *i.e.*, $\left(J_g^\top J_g\right)_{ij} = \mathfrak{g}_{ij}$, where $\left(J_g^\top J_g\right)_{ij}$ is the (i, j)-th element of $J_g^\top J_g$. Since J_g is orthonormal, we have that $\mathfrak{g}_{ij} \equiv 0$ for $i \neq j$ and $\mathfrak{g}_{ij} \equiv 1$ for $i = j$. Hence we further have that $\frac{\partial \mathfrak{g}_{ij}}{\partial z^l} \equiv 0$ and $\Gamma_{ij}^k \equiv 0$. So we have

$$\nabla_{\frac{\partial}{\partial z^j}} \frac{\partial}{\partial z^i} = \Gamma_{ij}^k \frac{\partial}{\partial z^k} \equiv 0, \tag{3}$$

and hence $\nabla_{\gamma'}\gamma' \equiv 0$, *i.e.*, γ is a geodesic connecting $g(u_1)$ and $g(u_2)$ on $\mathcal{M}$. To further prove that $d_{\mathcal{M}}(g(u_1), g(u_2)) = \|u_1 - u_2\|_2$, according to Riemannian geometry, we have

$$d_{\mathcal{M}}(g(u_1), g(u_2)) \triangleq \int_0^1 \sqrt{\sum_{ij} \mathfrak{g}_{ij} \frac{\mathrm{d}z^i}{\mathrm{d}t} \frac{\mathrm{d}z^j}{\mathrm{d}t}} \mathrm{d}t = \int_0^1 \sqrt{\sum_i \mathfrak{g}_{ii} (\xi^i)^2} \mathrm{d}t = \sqrt{\sum_i (\xi^i)^2} \int_0^1 \mathrm{d}t = \|u_1 - u_2\|_2, \tag{4}$$

where we use $\mathfrak{g}_{ii} \equiv 1$ due to the orthonormality of J_g. The convex property of $\mathcal{U}^1(\mathcal{M})$ is to ensure that $tu_2 + (1-t)u_1 \in \mathcal{U}^1(\mathcal{M}), \forall t \in [0, 1]$. Summarizing the above analysis completes the proof. □

Proposition 2. *$J_g(z)$ is orthonormal for $\forall z \in \mathcal{U}^1(\mathcal{M})$, if $J_{f_i}(u)$ is orthonormal for $\forall u \in \mathcal{A}^{i-1}(\mathcal{M}; r), \forall i \in [L]$.*

[1] The Einstein summation convention is used, namely $X^i \frac{\partial}{\partial z^i} \triangleq \sum_{i=1}^K X^i \frac{\partial}{\partial z^i}$, where z^i is the i-th coordinate.

Proof. Given $g = f_L \circ p_L \circ f_{L-1} \circ p_{L-1} \circ \cdots \circ f_1 \circ p_1$, we have that $\forall u^1 \in \mathcal{U}^1(\mathcal{M})$,

$$J_g\left(u^1\right) = J_{f_L}\left(u^L\right) I_{K_L \times K_{L-1}}\left(u^L\right) J_{f_{L-1}}\left(u^{L-1}\right) I_{K_{L-1} \times K_{L-2}}\left(u^{L-1}\right) \cdots J_{f_1}\left(u^1\right) I_{K_1 \times K_0}\left(u^1\right), \tag{5}$$

where we use $I_{K_i \times K_{i-1}} \in \mathbb{R}^{K_i \times K_{i-1}}$ to denote an identity matrix (*i.e.*, all diagonal elements are 1 meanwhile others are 0), and $u^{i+1} \triangleq f_i\left(u^i\right)$. Since $u^1 \in \mathcal{U}^1(\mathcal{M})$, we have $u^{i+1} \in \widetilde{\mathcal{A}}^i(\mathcal{M}; r) \subset \mathcal{A}^i(\mathcal{M}; r)$ according to the definition of $\widetilde{A}^i(\mathcal{M}; r)$ and $\mathcal{A}^i(\mathcal{M}; r)$ in main text, namely $u^i \in \mathcal{A}^{i-1}(\mathcal{M}; r)$. Hence we know that $J_{f_i}\left(u^i\right)$ is orthonormal. Therefore, we have

$$\begin{aligned}
&J_g^\top\left(u^1\right) J_g\left(u^1\right) = I_{K_1 \times K_0}^\top\left(u^1\right) J_{f_1}^\top\left(u^1\right) \cdots I_{K_{L-1} \times K_{L-2}}^\top\left(u^{L-1}\right) J_{f_{L-1}}^\top\left(u^{L-1}\right) I_{K_L \times K_{L-1}}^\top\left(u^L\right) J_{f_L}^\top\left(u^L\right) && (6)\\
&\quad J_{f_L}\left(u^L\right) I_{K_L \times K_{L-1}}\left(u^L\right) J_{f_{L-1}}\left(u^{L-1}\right) I_{K_{L-1} \times K_{L-2}}\left(u^{L-1}\right) \cdots J_{f_1}\left(u^1\right) I_{K_1 \times K_0}\left(u^1\right) && (7)\\
&= I_{K_1 \times K_0}^\top\left(u^1\right) J_{f_1}^\top\left(u^1\right) \cdots I_{K_{L-1} \times K_{L-2}}^\top\left(u^{L-1}\right) J_{f_{L-1}}^\top\left(u^{L-1}\right) && (8)\\
&\quad J_{f_{L-1}}\left(u^{L-1}\right) I_{K_{L-1} \times K_{L-2}}\left(u^{L-1}\right) \cdots J_{f_1}\left(u^1\right) I_{K_1 \times K_0}\left(u^1\right) && (9)\\
&= I_{K_1 \times K_0}^\top\left(u^1\right) J_{f_1}^\top\left(u^1\right) \cdots J_{f_1}\left(u^1\right) I_{K_1 \times K_0}\left(u^1\right) && (10)\\
&= I_{K_0 \times K_0}, && (11)
\end{aligned}$$

namely $J_g\left(u^1\right)$ is orthonormal, where $I_{K_i \times K_{i-1}}^\top\left(u^i\right) I_{K_i \times K_{i-1}}\left(u^i\right) = I_{K_{i-1} \times K_{i-1}}$ and $J_{f_i}^\top\left(u^i\right) J_{f_i}\left(u^i\right) = I_{K_i \times K_i}$. □

1.2. The *Rigorous Projection* Property

We firstly provide proofs for Prop. 2 (restated as Prop. 3) and Prop. 4 (restated as Prop. 4) in main text for satisfying the *rigorous projection* property, and then provide intuitive understandings about the *rigorous projection* property. We finally prove Cor. 1 (restated as Cor. 1) that claims the *rigorous projection* property for an one-layer generator, and provide further analysis on Rem. 1 (restated as Rem. 1) that discusses the *rigorous projection* property for a multi-layer generator.

Proposition 3 (*Rigorous projection* property)**.** *Assume that $v^i(x) = 0$ for $\forall x \in \mathcal{M}$. Given $x \in \mathcal{A}^i(\mathcal{M}; r)$, by denoting $[\alpha; \beta] \triangleq f_i^{-1}(x)$ where $\alpha \in \mathbb{R}^{K_{i-1}}$ and $\beta \in \mathbb{R}^{K_i - K_{i-1}}$, the projection of x onto $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$ is $\widetilde{x} \triangleq f_i(\alpha)$, and the distance from x to $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$, $d_{\mathcal{E}}(x, \widetilde{x})$, equals to $\|\beta\|_2$, if $\forall u \in \mathcal{A}^{i-1}(\mathcal{M}; r), v \in \mathcal{V}^i(r)$,*

$$I([u; v]) \triangleq J([u; v])^\top J([u; v]) \in \mathbb{R}^{K_i \times K_i} \tag{12}$$

is a diagonal matrix, and $I_{jj}([u; v])$ equals to 1 for $K_{i-1} + 1 \leqslant j \leqslant K_i$, where $J([u; v]) \in \mathbb{R}^{K_i \times K_i}$ is the Jacobian of f_i at $[u; v]$, I_{jj} is the (j, j)-th element of I, and $d_{\mathcal{E}}(\cdot, \cdot)$ is the Euclidean distance in $\mathbb{R}^{K_i}$.

Proof. Given $\forall u \in \mathcal{A}^{i-1}(\mathcal{M}; r), \forall v \in \mathcal{V}^i(r)$, we firstly prove that the curve $\gamma(t) \triangleq f_i([u; tv])$ is a geodesic on $\mathcal{A}^i(\mathcal{M}; r)$. Because $\gamma' = v^m \frac{\partial}{\partial z^m}$ (note that $m \in \mathcal{I}_v$, where $\mathcal{I}_v$ is the set of indices corresponding to the v coordinates of f_i), we further have that

$$\nabla_{\gamma'} \gamma' = \nabla_{v^n \frac{\partial}{\partial z^n}} v^m \frac{\partial}{\partial z^m} = v^n v^m \nabla_{\frac{\partial}{\partial z^n}} \frac{\partial}{\partial z^m}. \tag{13}$$

Let $\mathfrak{g}_{mn}$ be the covariant components of the Riemannian metric $\mathcal{G}$, then from Eq. (12) we have $\mathfrak{g}_{mn} \equiv 0$ for $m \neq n \in [K_i]$ and $\mathfrak{g}_{mm} \equiv 1$ for $m \in \mathcal{I}_v$. Therefore, for $m, n \in \mathcal{I}_v$, we have $\frac{\partial \mathfrak{g}_{ml}}{\partial z^n} \equiv 0$, $\frac{\partial \mathfrak{g}_{nl}}{\partial z^m} \equiv 0$, $\frac{\partial \mathfrak{g}_{mn}}{\partial z^l} \equiv 0$. Hence from Eq. (2), we have $\Gamma_{mn}^k \equiv 0$, and $\nabla_{\frac{\partial}{\partial z^n}} \frac{\partial}{\partial z^m} = \Gamma_{mn}^k \frac{\partial}{\partial z^k} \equiv 0$, namely $\nabla_{\gamma'} \gamma' \equiv 0$. Hence γ is a geodesic on $\mathcal{A}^i(\mathcal{M}; r)$. Since $\mathcal{A}^i(\mathcal{M}; r)$ is essentially the ambient Euclidean space, we know that γ is a straight line, since the geodesics of an Euclidean space are straight lines. Due to the orthogonality of J, we know that $\gamma'(0) \perp T_{\gamma(0)} \widetilde{\mathcal{A}}^i(\mathcal{M}; r)$, where $T_{\gamma(0)} \widetilde{\mathcal{A}}^i(\mathcal{M}; r)$ is the tangent space of $\mathcal{M}$ at $\gamma(0)$. Therefore, we can learn that $\gamma \perp T_{\gamma(0)} \widetilde{\mathcal{A}}^i(\mathcal{M}; r)$, and the projection of any $x \in \gamma$ onto $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$ is $\gamma(0)$ according to the rigorous definition of projection, namely given $x = f_i([u; v])$, the projection of x onto $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$ is $\widetilde{x} \triangleq f_i(u)$. To further prove that $d_{\mathcal{E}}(x, \widetilde{x}) = \|v\|_2$ for $x = f_i([u; v])$, similar to Eq. (4), we have

$$d_{\mathcal{E}}(x, \widetilde{x}) = d_{\mathcal{A}^i(\mathcal{M}; r)}(x, \widetilde{x}) = \int_0^1 \sqrt{\sum_{m,n \in \mathcal{I}_v} \mathfrak{g}_{mn} \frac{\mathrm{d}z^m}{\mathrm{d}t} \frac{\mathrm{d}z^n}{\mathrm{d}t}} \mathrm{d}t = \sqrt{\sum_m \left(v^m\right)^2} \int_0^1 \mathrm{d}t = \|v\|_2, \tag{14}$$

where we use $\mathfrak{g}_{mm} \equiv 1$ for $i \in \mathcal{I}_v$. Summarizing the above analysis completes the proof. □

Proposition 4. *All singular values of* $J_{\phi_i^z}(\mathbb{1})$ *equal* $a > 0 \Leftrightarrow J_{f_i}(z)^\top J_{f_i}(z) = \operatorname{diag}\left(\frac{a^2}{\omega_1^2}, \cdots, \frac{a^2}{\omega_{K_i}^2}\right)$, *and* $\left\{\frac{a}{\omega_j}\right\}_{j=1}^{K_i}$ *are the singular values of* $J_{f_i}(z)$.

Proof. Note that $J_{\phi_i^z}(\mathbb{1}) = J_{f_i}(z)\operatorname{diag}(\omega_1, \omega_2, \cdots, \omega_{K_i})$ from Eq. 13 in main text. Since all singular values of $J_{\phi_i^z}(\mathbb{1})$ equal $a > 0$, we have that $J_{\phi_i^z}^\top(\mathbb{1}) J_{\phi_i^z}(\mathbb{1}) = a^2 I$ where I is an identity matrix, and therefore $S^\top J_{f_i}^\top(z) J_{f_i}(z) S = a^2 I$, where $S \triangleq \operatorname{diag}(\omega_1, \cdots, \omega_{K_i})$, namely

$$J_{f_i}^\top(z) J_{f_i}(z) = \operatorname{diag}\left(\frac{a^2}{\omega_1^2}, \cdots, \frac{a^2}{\omega_{K_i}^2}\right), \tag{15}$$

which means that J_{f_i} is orthogonal. Let the SVD of J_{f_i} be $J_{f_i} = U\Sigma V^\top$, where $\Sigma \triangleq \operatorname{diag}(\sigma_1, \cdots, \sigma_K)$, then we have that

$$J_{f_i}^\top J_{f_i} = V\Sigma^\top \Sigma V^\top = V\operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_K^2\right) V^\top = \operatorname{diag}\left(\frac{a^2}{\omega_1^2}, \cdots, \frac{a^2}{\omega_{K_i}^2}\right). \tag{16}$$

Note that Eq. (16) is a SVD of $\operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_K^2\right)$, hence up to permutation we have that $\sigma_j^2 = \frac{a^2}{\omega_j^2}$, namely $\left\{\frac{a}{\omega_j}\right\}_{j=1}^{K_i}$ are the singular values of $J_{f_i}(z)$. □

We then reveal that how the following objective

$$\begin{aligned} &\text{constrain all singular values of } J_{\phi_i^z}(\mathbb{1}) \text{ to be equal,} \\ &\text{s.t. } \omega_j = \sqrt[K_{i-1}]{\left(\prod_{l=1}^{K_{i-1}} \omega_l\right) \det J_{f_i}}, \quad K_{i-1} + 1 \leqslant j \leqslant K_i \end{aligned} \tag{17}$$

mentioned in main text leads to an f_i satisfying the conditions for the *rigorous projection* property in Prop. 3. Given that all singular values of $J_{\phi_i^z}(\mathbb{1})$ are equal, from Prop. 4 we have that $\det J_{f_i} = \prod_{l=1}^{K_i} \frac{a}{\omega_l}$, therefore from Eq. (17) we have that

$$\omega_j = \sqrt[K_{i-1}]{\left(\prod_{l=1}^{K_{i-1}} \omega_l\right)\left(\prod_{l=1}^{K_i} \frac{a}{\omega_l}\right)} = \sqrt[K_{i-1}]{a^{K_i} \prod_{l=K_{i-1}+1}^{K_i} \frac{1}{\omega_l}}, \quad K_{i-1} + 1 \leqslant j \leqslant K_i. \tag{18}$$

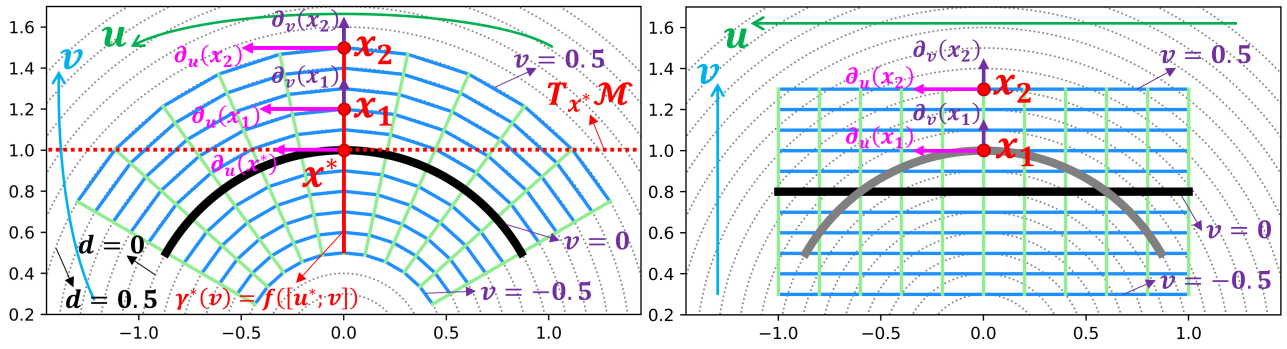
Since we know that $\omega_{K_{i-1}+1} = \omega_{K_{l-1}+2} = \cdots = \omega_{K_i}$ from Eq. (17), we denote that $\omega_{K_{i-1}+1} = \omega_{K_{l-1}+2} = \cdots = \omega_{K_i} = \xi$. From Eq. (18) we know that $\xi^{K_{i-1}} = a^{K_i} \xi^{K_{i-1}-K_i}$, namely $\frac{a^{K_i}}{\xi^{K_i}} = 1$ and hence $\frac{a}{\xi} = 1$. Since we know that $\sigma_j = \frac{a}{\omega_j}$ from Prop. 4, we know that $\sigma_j = \frac{a}{\xi} = 1$ for $K_{i-1} + 1 \leqslant j \leqslant K_i$, where σ_j is a singular value of J_{f_i}. Note that J_{f_i} is orthonormal up to a scalar, therefore $J_{f_i}^\top J_{f_i}$ is a diagonal matrix with the diagonal elements being the squared singular values of J_{f_i}. From the above analysis, we know that the conditions for the *rigorous projection* property in Prop. 3 are satisfied.

We now provide intuitive understandings about the *rigorous projection* property by using Fig. 1, including how the conditions stated in Prop. 3 leads to a *flow* module f_i that satisfies the *rigorous projection* property, and why we use Eq. 18 in main text instead of a simpler regularization that trivially constrains J_{f_i} to be orthonormal over $\mathcal{A}^{i-1}(\mathcal{M}; r) \times \mathcal{V}^i(r)$. For intuitive illustration, we use a 1-dimensional manifold $\mathcal{M} \triangleq \left\{(\cos\theta, \sin\theta) \,|\theta \in \left[\frac{\pi}{6}, \frac{5\pi}{6}\right]\right\}$ (*i.e.*, a curve) residing in the 2-dimensional Euclidean space as we introduced in main text.

Intuitively Understanding Prop. 3 We consider a flow module $f : \mathbb{R}^2 \to \mathbb{R}^2$ that satisfies the conditions in Prop. 3 and successfully recovers $\mathcal{M}$, namely there exists some $\mathcal{U} \subset \mathbb{R}^1$ such that $\mathcal{M} = \{f([u; 0]) \,|u \in \mathcal{U}\}$. The coordinate chart of f is plotted in Fig. 1(a). Consider the coordinate curve $\gamma^* \triangleq \{f([u^*; v]) \,|v \in [-0.5, 0.5]\}$ for $u^* \in \mathcal{U}$ as shown in Fig. 1(a). We know that $x^* \triangleq f([u^*; 0]) \in \gamma^*$ is on $\mathcal{M}$ and hence $\partial_u(x^*)$ is the basis of $T_{x^*}\mathcal{M}$, namely $T_{x^*}\mathcal{M} = \{a\partial_u(x^*) \,|a \in \mathbb{R}\}$ (see Fig. 1(a)). Since f satisfies the conditions in Prop. 3, we know that for $\forall x \in \gamma^*$,

$$\partial_u(x) \perp \partial_v(x), \tag{19}$$

$$\|\partial_v(x)\|_2 = 1, \tag{20}$$



(a) An intuitive understanding on the conditions stated in Prop. 3. (b) The trivial regularization restricts the flexibility of the *flow* module.

Figure 1. Intuitive understandings about the *rigorous projection* property by using the synthetic manifold $\mathcal{M}$ introduced in main text. In Fig. 1(a), we visualize the coordinate chart of a flow module $f : \mathbb{R}^2 \to \mathbb{R}^2$ that satisfies the conditions in Prop. 3, meanwhile in Fig. 1(b), we show the coordinate chart of a flow module $f' : \mathbb{R}^2 \to \mathbb{R}^2$ whose Jacobian is trivially constrained to be orthonormal over the ambient space. The ground-truth manifold $\mathcal{M}$ and the manifolds recovered by the flow modules are plotted by using the thick gray solid line and the thick black solid lines, respectively, and the contour lines with respect to distances to $\mathcal{M}$ are plotted by using the thin grey dotted lines, namely samples on the same contour line $d = \alpha$ have equal distances (*i.e.*, α) to $\mathcal{M}$. For the coordinate charts of the two flow modules, the two embedding dimensions u and v are represented in green and blue, respectively, namely samples on the same green (*resp.*, blue) line corresponds to the same u (*resp.*, v). For both flow modules, we use $\partial_u(x)$ (*resp.*, $\partial_v(x)$) to denote the tangent vector $\frac{\partial}{\partial u}\Big|_x$ (*resp.*, $\frac{\partial}{\partial v}\Big|_x$) of the coordinate curve u (*resp.*, v) at x. In Fig. 1(a), given $x^* \in \mathcal{M}$, we use $T_{x^*}\mathcal{M}$ to denote the tangent space of $\mathcal{M}$ at x^*.

which results in $\partial_u(x)$ being parallel along γ^*, namely $\partial_u(x_1) \parallel \partial_u(x_2)$ for $\forall x_1, x_2 \in \gamma^*$ (see Fig. 1(a)). Combined with Eq. (19), we can learn that γ^* is a straight line perpendicular to $\partial_u(x^*)$, and hence we have $\mathrm{vec}(x^*, x) \perp \partial_u(x^*)$, *i.e.*, $\mathrm{vec}(x^*, x) \perp T_{x^*}\mathcal{M}$. So according to the rigorous definition of projection, we know that the projection of $x = f([u^*; \beta]) \in \gamma^*$ onto $\mathcal{M}$ is $x^* = f([u^*; 0])$. Regarding the distance from x to x^*, given that γ^* is straight, from Eq. (20) we learn that

$$d_{\mathcal{E}}(x, x^*) = \int_0^{\beta} \|\partial_v \mathrm{d}v\|_2 = \int_0^{\beta} |\mathrm{d}v| = |\beta|, \tag{21}$$

which coincides with the results given in Prop. 3. The above analysis provides an intuitive understanding on how the conditions stated in Prop. 3 leads to a flow module f that satisfies the *rigorous projection* property.

Intuitively Understanding the Isometric Regularization on the Ambient Space We consider a flow module $f' : \mathbb{R}^2 \to \mathbb{R}^2$ that is constrained to be orthonormal over the ambient space $\mathcal{A} \triangleq \{f'([u; v]) \,|\, u \in \mathcal{U}, v \in [-0.5, 0.5]\}$ as in Fig. 1(b). Since $J_{f'}$ is orthonormal, we know that $\partial_u(x) \perp \partial_v(x)$ and $\|\partial_u(x)\|_2 = \|\partial_v(x)\|_2 = 1, \forall x \in \mathcal{A}$. Hence intuitively, all coordinate curves of u (*resp.*, v) are straight lines and are perpendicular to the coordinate curves of v (*resp.*, u), and all coordinate curves of u (*resp.*, v) are the same length, as we shown in Fig. 1(b). This leads to a trivial case where the generated manifold (plotted in thick black solid line in Fig. 1(b)) cannot be curved, and hence it is impossible to successfully recover the ground-truth manifold (plotted in thick gray solid line in Fig. 1(b)) by using f'. Therefore, trivially constraining $J_{f'}$ to be orthonormal over the ambient space can severely restrict the flexibility of f', and our proposed isometric regularization (Eq. 18 in main text) allows the flow module to be more flexible meanwhile satisfying the *rigorous projection* property.

1.2.1 Case of One-layer Generator

We prove Cor. 1 that claims the *rigorous projection* property for an one-layer generator.

Corollary 1. *Given $g = f_1 \circ p_1$, assume that $v^1(x) = 0$ for $\forall x \in \mathcal{M}$. Given $x \in \mathcal{A}^1(\mathcal{M}; r)$, the projection of x onto $\mathcal{M}$ is $\widetilde{x} \triangleq g\left(u^1(x)\right) = g \circ g^{\dagger}(x)$, and the distance from x to $\mathcal{M}$, $d_{\mathcal{E}}(x, \widetilde{x})$, is $\left\|v^1(x)\right\|_2$, if f_1 satisfies the conditions for the rigorous projection property in Prop. 3.*

Proof. Since $x = f_1\left(\left[u^1(x); v^1(x)\right]\right)$ and $\widetilde{\mathcal{A}}^i(\mathcal{M}; r) = \mathcal{M}$ in the case of one-layer generator, from Prop. 3, the projection of x onto $\mathcal{M}$ is $\widetilde{x} = f_1\left(u^1(x)\right) = g_1\left(u^1(x)\right) = g \circ g^{\dagger}(x)$, and distance from x to $\mathcal{M}$, $d_{\mathcal{E}}(x, \widetilde{x})$ is $\left\|v^1(x)\right\|_2$. □

1.2.2 Case of Multi-layer Generator

We provide further analysis on Rem. 1 that discusses the *rigorous projection* property for a multi-layer generator.

Remark 1. *Given g comprised of $L \geqslant 2$ layers, assume that $v^i(x) = 0$ for $\forall x \in \mathcal{M}, i \in [L]$. Given $x \in \mathcal{A}^L(\mathcal{M}; r)$, let $\widetilde{x}^i \triangleq f_i\left(u^i(x)\right), i \in [L]$, and $\widetilde{x} \triangleq g\left(u^1(x)\right) = g \circ g^\dagger(x)$. Assume that r is small such that $\mathcal{A}^L(\mathcal{M}; r)$ is an ambient space close to $\mathcal{M}$. Given f_i that satisfies the rigorous projection property for $i \in [L]$, we have*

- *The projection of $u^{i+1}(x)$ onto $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$ is $\widetilde{x}^i$, and the distance from $u^{i+1}(x)$ to $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$, $d_{\mathcal{E}}\left(u^{i+1}(x), \widetilde{x}^i\right)$, equals to $\left\|v^i(x)\right\|_2$, where $i \in [L]$.*
- *The projection of x onto $\mathcal{M}$ is $\widetilde{x}$, and the distance from x to $\mathcal{M}$, $d_{\mathcal{E}}(x, \widetilde{x})$, is approximately $\sqrt{\sum_{i=1}^{L} \left\|v^i(x)\right\|_2^2}$.*

Given $x \in \mathcal{A}^L(\mathcal{M}; r)$ and f_i that satisfies the *rigorous projection* property for $i \in [L]$, let $\widetilde{x}^i \triangleq f_i\left(u^i(x)\right)$, then we can directly know from Prop. 3 that the projection of $u^{i+1}(x)$ onto $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$ is $\widetilde{x}^i$, and the distance from $u^{i+1}(x)$ to $\widetilde{\mathcal{A}}^i(\mathcal{M}; r)$, $d_{\mathcal{E}}\left(u^{i+1}(x), \widetilde{x}^i\right)$, equals to $\left\|v^i(x)\right\|_2$. Given $\widetilde{x} \triangleq g\left(u^1(x)\right)$, assume that $\widetilde{x}$ is in a local region $\mathcal{O}^L \subset \widetilde{\mathcal{A}}^L(\mathcal{M}; r)$ around $\widetilde{x}^L$, *i.e.*, it can be regarded that $\mathcal{O}^L \approx T_{\widetilde{x}^L}\widetilde{\mathcal{A}}^L(\mathcal{M}; r)$. Since $\widetilde{x}^L$ is the projection of x onto $\widetilde{\mathcal{A}}^L(\mathcal{M}; r)$, according to the rigorous definition of projection, we have $\operatorname{vec}\left(x, \widetilde{x}^L\right) \perp T_{\widetilde{x}^L}\widetilde{\mathcal{A}}^L(\mathcal{M}; r)$, and hence $\operatorname{vec}\left(x, \widetilde{x}^L\right) \perp \mathcal{O}^L$ approximately holds. Because $\widetilde{x} \in \mathcal{O}^L$, we have that $\operatorname{vec}\left(x, \widetilde{x}^L\right) \perp \operatorname{vec}\left(\widetilde{x}^L, \widetilde{x}\right)$ approximately holds, and hence $d_{\mathcal{E}}^2(x, \widetilde{x}) \approx d_{\mathcal{E}}^2\left(x, \widetilde{x}^L\right) + d_{\mathcal{E}}^2\left(\widetilde{x}^L, \widetilde{x}\right)$. Note that $d_{\mathcal{E}}\left(x, \widetilde{x}^L\right) = \left\|v^L(x)\right\|_2^2$. For $d_{\mathcal{E}}\left(\widetilde{x}^L, \widetilde{x}\right)$, due to the locality of $\mathcal{O}^L$, we have that $d_{\mathcal{E}}\left(\widetilde{x}^L, \widetilde{x}\right) \approx d_{\widetilde{\mathcal{A}}^L(\mathcal{M}; r)}\left(\widetilde{x}^L, \widetilde{x}\right)$. Moreover, because $\widetilde{x} = f_L\left(u^L(\widetilde{x})\right)$ and f_L is an isometry, we have $d_{\widetilde{\mathcal{A}}^L(\mathcal{M}; r)}\left(\widetilde{x}^L, \widetilde{x}\right) = d_{\mathcal{E}}\left(u^L(x), u^L(\widetilde{x})\right)$. Therefore, summarizing the above analysis gives us that

$$d_{\mathcal{E}}^2(x, \widetilde{x}) \approx \left\|v^L(x)\right\|_2^2 + d_{\mathcal{E}}^2\left(u^L(x), u^L(\widetilde{x})\right). \tag{22}$$

The above analysis can be further generalized to the next layer. Since f_{L-1} satisfies the *rigorous projection* property, we can learn that the projection of $u^L(x)$ onto $\widetilde{\mathcal{A}}^{L-1}(\mathcal{M}; r)$ is $\widetilde{x}^{L-1}$, and $d_{\mathcal{E}}\left(u^L(x), \widetilde{x}^{L-1}\right) = \left\|v^{L-1}(x)\right\|_2$. Similarly, assume that $u^L(\widetilde{x})$ is in a local region $\mathcal{O}^{L-1} \subset \widetilde{\mathcal{A}}^{L-1}(\mathcal{M}; r)$ around $\widetilde{x}^{L-1}$, *i.e.*, it can be regarded that $\mathcal{O}^{L-1} \approx T_{\widetilde{x}^{L-1}}\widetilde{\mathcal{A}}^{L-1}(\mathcal{M}; r)$. According to the rigorous definition of projection, we have $\operatorname{vec}\left(u^L(x), \widetilde{x}^{L-1}\right) \perp T_{\widetilde{x}^{L-1}}\widetilde{\mathcal{A}}^{L-1}(\mathcal{M}; r)$, and hence we further have that $\operatorname{vec}\left(u^L(x), \widetilde{x}^{L-1}\right) \perp \mathcal{O}^{L-1}$ approximately holds. Because $u^L(\widetilde{x}) \in \mathcal{O}^{L-1}$, we have that $\operatorname{vec}\left(u^L(x), \widetilde{x}^{L-1}\right) \perp \operatorname{vec}\left(\widetilde{x}^{L-1}, u^L(\widetilde{x})\right)$ approximately holds. Hence we have $d_{\mathcal{E}}^2\left(u^L(x), u^L(\widetilde{x})\right) \approx d_{\mathcal{E}}^2\left(u^L(x), \widetilde{x}^{L-1}\right) + d_{\mathcal{E}}^2\left(\widetilde{x}^{L-1}, u^L(\widetilde{x})\right)$, where $d_{\mathcal{E}}\left(\widetilde{x}^{L-1}, u^L(\widetilde{x})\right) \approx d_{\widetilde{A}^{L-1}(\mathcal{M}; r)}\left(\widetilde{x}^{L-1}, u^L(\widetilde{x})\right)$ due to the locality of $\mathcal{O}^{L-1}$. Because $u^L(\widetilde{x}) = f_{L-1}\left(u^{L-1}(\widetilde{x})\right)$ and f_{L-1} is an isometry, we further have $d_{\widetilde{A}^{L-1}(\mathcal{M}; r)}\left(\widetilde{x}^{L-1}, u^L(\widetilde{x})\right) = d_{\mathcal{E}}\left(u^{L-1}(x), u^{L-1}(\widetilde{x})\right)$. Therefore, summarizing the above analysis gives us that

$$d_{\mathcal{E}}^2\left(u^L(x), u^L(\widetilde{x})\right) \approx \left\|v^{L-1}(x)\right\|_2^2 + d_{\mathcal{E}}^2\left(u^{L-1}(x), u^{L-1}(\widetilde{x})\right). \tag{23}$$

Therefore, combining Eq. (22) and Eq. (23), we have that

$$d_{\mathcal{E}}^2(x, \widetilde{x}) \approx \left\|v^L(x)\right\|_2^2 + \left\|v^{L-1}(x)\right\|_2^2 + d_{\mathcal{E}}^2\left(u^{L-1}(x), u^{L-1}(\widetilde{x})\right). \tag{24}$$

By further generalizing Eq. (24), we have that

$$d_{\mathcal{E}}^2(x, \widetilde{x}) \approx \left\|v^L(x)\right\|_2^2 + \left\|v^{L-1}(x)\right\|_2^2 + \cdots + \left\|v^1(x)\right\|_2^2 + d_{\mathcal{E}}^2\left(u^1(x), u^1(\widetilde{x})\right). \tag{25}$$

From $\widetilde{x} = g\left(u^1(x)\right)$, we know that $u^1(x) = u^1(\widetilde{x})$, namely $d_{\mathcal{E}}\left(u^1(x), u^1(\widetilde{x})\right) = 0$ and hence

$$d_{\mathcal{E}}(x, \widetilde{x}) \approx \sqrt{\sum_{j=1}^{L} \left\|v^j(x)\right\|_2^2}. \tag{26}$$

2. Supplemental Experiments

In this section, we provide supplemental experimental results on the synthetic manifold introduced in main text and natural image datasets for manifold learning including MNIST, CelebA, FFHQ and AFHQ Cat, *etc*. We evaluate models in terms of the *distance preserving* property, the *rigorous projection* property and the *anomaly detection* performance. We also analyze the time-efficiency for our proposed *two-stage dimensionality reduction* algorithm on natural image datasets. We compare different manifold learning models including *Generative Adversarial Net* [3] (GAN), *Variational Autoencoder* [8] (VAE), *Isometric Autoencoder* [4] (IAE), *Pseudo Invertible Encoders* [1] (PIE), $\mathcal{M}$*-flow* [2], and our *Hierarchical Flow* (HF).

Architectural Details Our experiments involve convolutional or fully-connected neural network models including GAN, VAE, IAE, and NF-based models including PIE, $\mathcal{M}$-flow and our proposed HF. For an introduction on NF modules such as *coupling* layers, see [10]. The architectural details of models involved in our experiments are shown in Tbl. 1-2.

Experimental Settings We introduce implementation details for conducting our experiments. In main text, we provide the training objectives for the manifold phase and the density phase of our HF. To see the detailed implementation of the LT technique mentioned in main text for estimating the spectral norm of Jacobians in a sample-efficient manner, see [10]. Based on the training objectives introduced in main text, we also provide the detailed overall training algorithm for our HF model as in Alg. 2. In terms of training settings, we train $200,000$ iterations with batch size 64, using an Adam optimizer [7] with fixed learning rate $1e^{-4}$. All models are train from scratch in an end-to-end manner by using PyTorch [11]. For settings of evaluation metrics corresponding to specific datasets, we introduce along with the respective experiments in the following Sections 2.1-2.2. All our experiments are done by using a single NVIDIA RTX-2080Ti GPU.

2.1. Results on the Synthetic Manifold

In main text, we provide demonstrative qualitative results showing that our proposed HF model can learn the manifold in a manner that satisfies the *distance preserving* property and the *rigorous projection* property while the other models such as $\mathcal{M}$-flow and IAE are not guaranteed to satisfy the two properties. In this section, we provide more quantitative results.

Experimental Settings We introduce the metrics we used to evaluate models. To evaluate the *distance preserving* property, we propose to compute the *Mean Absolute Error* (MAE) between $\frac{d_{\mathcal{M}}(x_1,x_2)}{\|z_1-z_2\|_2}$ and 1, because for an isometric g, the manifold distance $d_{\mathcal{M}}(x_1, x_2)$ equals to the corresponding embedding distance $\|z_1 - z_2\|_2$, where $x_1, x_2 \in \mathcal{M}$ and $x_1 = g(z_1), x_2 = g(z_2)$, and $d_{\mathcal{M}}$ is the manifold distance that can be analytically computed as the arc length between x_1 and x_2 considering that $\mathcal{M}$ is a part of a circle as we introduce in main text. For VAE, IAE, PIE, $\mathcal{M}$-flow and HF, the metric is computed as

$$\mathbb{E}_{x_1,x_2\sim\mathcal{M}}\left|\frac{d_{\mathcal{M}}(x_1,x_2)}{\|g^{-1}(x_1)-g^{-1}(x_2)\|_2}-1\right|, \tag{27}$$

where for VAE and IAE we choose g^{-1} to be the encoder. Because GAN is incapable of inferring the embedding of a given sample, the metric for GAN is computed differently as

$$\mathbb{E}_{z_1,z_2\sim\mathcal{Z}}\left|\frac{d_{\mathcal{M}}(g(z_1),g(z_2))}{\|z_1-z_2\|_2}-1\right|, \tag{28}$$

where $\mathcal{Z}$ is a latent space such that $g(z) \in \mathcal{M}, \forall z \in \mathcal{Z}$. In addition to the above metric, we also propose a different metric for evaluating the *distance preserving* property by measuring how far the Jacobian of g deviates from an orthonormal matrix. Specifically, the computation of the metric is given as

$$\mathbb{E}_{z\sim\mathcal{Z}}\left|J_g(z)^\top J_g(z)-I\right|, \tag{29}$$

where $\mathcal{Z}$ is the latent space for ancestral sampling, $J_g(z)$ is the Jacobian of g at z, and I is an identity matrix. Given a matrix $M \in \mathbb{R}^{K\times K}$, we denote $|M| \triangleq \frac{1}{K^2}\sum_{i,j}^{K}|M_{ij}|$, where $|M_{ij}|$ is the absolute value of the (i,j)-th element of M. To evaluate the *rigorous projection* property of the model, we evaluate how far the projection point $\widetilde{x}'$ given by the model g deviates from the ground-truth $\widetilde{x}$ for an off-manifold sample x, the metric is computed as

$$\mathbb{E}_{x\sim\mathcal{A}^1}\|\widetilde{x}'-\widetilde{x}\|_2, \tag{30}$$

where we choose the ambient space $\mathcal{A}^1 = \left\{(r\cos\theta, r\sin\theta)\,|r \in [-0.5, 0.5]\,, \theta \in \left[\frac{\pi}{6}, \frac{5\pi}{6}\right]\right\}$, and the ground-truth projection point $\widetilde{x}$ for x can be analytically given as introduce in Fig. 3 in main text. For $\mathcal{M}$-flow and our proposed HF, in addition to the above metric, we also propose a different metric for evaluating the *rigorous projection* property for the model by measuring how far the Jacobian of f_1 deviates from the conditions stated in Prop. 3, which is computed as

$$\underbrace{\mathbb{E}_{u\sim\mathcal{U}^1, v\sim\mathcal{V}^1} \frac{\sum_{i\neq j}^{K_1} \left|\left(S_1\left([u;v]\right)\right)_{ij}\right|}{(K_1-1)\sum_{i=1}^{K_1}\left|\left(S_1\left([u;v]\right)\right)_{ii}\right|}}_{①}, \quad \underbrace{\mathbb{E}_{u\sim\mathcal{U}^1, v\sim\mathcal{V}^1} \frac{1}{K_1-K_0} \sum_{i=K_0+1}^{K_1} \left|\left(S_1\left([u;v]\right)\right)_{ii} - 1\right|}_{②} \tag{31}$$

where f_1 is the *flow* module of the generator g, $S_1\left([u;v]\right) \triangleq J_{f_1}\left([u;v]\right)^\top J_{f_1}\left([u;v]\right)$, and we use M_{ij} to denote the (i,j)-th element of a matrix M. The metric ① is used to measure the orthogonality of $S_1\left([u;v]\right)$, which is computed as the ratio of the averaged absolute value of non-diagonal elements of $S_1\left([u;v]\right)$ to the averaged absolute value of the diagonal elements of $S_1\left([u;v]\right)$, hence the lower the metric ①, the more orthogonal the Jacobian $J_{f_1}\left([u;v]\right)$. The metric ② is used to measure how far the last $K_1 - K_0$ diagonal elements of $S_1\left([u;v]\right)$ deviates from 1. For the synthetic manifold we introduced in main text, we have $K_0 = K = 1, K_1 = D = 2$, and we use $\mathcal{V}^1 = [-0.5, 0.5]$. We also evaluate how far the projection distance $d_g\left(x, \widetilde{x}'\right)$ predicted by the model deviates from the ground-truth Euclidean distance $d_{\mathcal{E}}\left(x, \widetilde{x}'\right)$ for $\mathcal{M}$-flow and HF, which is computed as

$$\mathbb{E}_{x\sim\mathcal{A}} \left|d_g\left(x, \widetilde{x}'\right) - d_{\mathcal{E}}\left(x, \widetilde{x}'\right)\right|, \tag{32}$$

where $\widetilde{x}' = f_1\left(\left[u^1\left(x\right); 0\right]\right)$ and $d_g\left(x, \widetilde{x}'\right) = \left|v^1\left(x\right)\right|$. To evaluate the *anomaly detection* performance, we propose to adopt the *False Positive Rate at* 95% *True Positive Rate* following [9, 5]. Specifically, we detect the out-of-distribution samples based on the sample-to-manifold distance given by the model instead of using a classifier. For the ambient space $\mathcal{A}^1$ we used in Eq. (30), we choose the in-distribution samples set $\mathcal{A}^1_{\text{in}} \subset \mathcal{A}^1$ such that $d\left(x, \mathcal{M}\right) < 0.25, \forall x \in \mathcal{A}^1_{\text{in}}$, and choose the out-distribution samples set $\mathcal{A}^1_{\text{out}} \subset \mathcal{A}^1$ such that $d\left(x, \mathcal{M}\right) \geqslant 0.25, \forall x \in \mathcal{A}^1_{\text{out}}$, where $d\left(x, \mathcal{M}\right)$ is the ground-truth distance from x to the manifold $\mathcal{M}$. Given in-distribution samples $\mathcal{X}_{\text{in}} \triangleq \left\{x^{(i)} \in \mathcal{A}^1_{\text{in}}\right\}_{i=1}^N$ and out-distribution samples $\mathcal{X}_{\text{out}} \triangleq \left\{x^{(i)} \in \mathcal{A}^1_{\text{out}}\right\}_{i=1}^N$, the metric is computed as

$$\underbrace{\frac{\left|\left\{|v\left(x\right)| < v^* | x \in \mathcal{X}_{\text{out}}\right\}\right|}{N}}_{\text{False Positive Rate (FPR)}}, \quad \text{s.t.} \underbrace{\frac{\left|\left\{|v\left(x\right)| < v^* | x \in \mathcal{X}_{\text{in}}\right\}\right|}{N}}_{\text{True Positive Rate (TPR)}} = 0.95, \tag{33}$$

where we use $N = 1,000$ and $|\mathcal{X}|$ is the number of elements in the set $\mathcal{X}$, namely we consider x an in-distribution sample if the distance to the manifold (*i.e.*, $|v\left(x\right)|$) predicted by the model is smaller than v^*.

Quantitative Results We provide quantitative results in Tbl. 3. As we can see in the table, models including GAN, VAE, PIE and $\mathcal{M}$-flow achieve inferior performances in terms of both the global and local metric scores of the *distance preserving* property, meanwhile IAE and our HF achieve better performances. Because IAE can be regarded as a VAE augmented with an isometric regularization proposed in [4], the gain of the performance verifies the effectiveness of the regularization in terms of encouraging the decoder to satisfy the *distance preserving* property. In terms of the *rigorous projection* property, as we can observe, our HF achieves the best performance across all models. Therefore, we learn that apart from our proposed HF, none of other models are guaranteed to satisfy the property. We also observe that our HF achieves the best performance in terms of *anomaly detection*. From the computation method of the *anomaly detection* metric as in Eq. (33), we argue that the inaccuracy of the projection distance of $\mathcal{M}$-flow causes the *anomaly detection* performance to drop. To see this, for $\mathcal{M}$-flow, the predicted projection distance of an in-distribution sample may larger than that of an out-distribution sample, hence the distributions of the predicted projection distances of in-distribution and out-distribution samples may not be discriminative, which leads an inferior score of the *FPR at* 95% *TPR* metric for $\mathcal{M}$-flow. However, as our HF can always predict correct projection distance of an off-manifold sample, the distributions of the predicted projection distances of in-distribution and out-distribution samples are always discriminative, and hence the *FPR at* 95% *TPR* metric score of our HF is superior. Moreover, the inaccuracy of the predicted projection distance $\left\|v\left(x\right)\right\|_2$ for x also makes it inappropriate for $\mathcal{M}$-flow to implement the reconstruction loss by minimizing $\left\|v\left(x\right)\right\|_2$, as we done in our HF to speed up training.

2.2. Results on Natural Image Datasets

In this section, we provide qualitative and quantitative results on the aforementioned natural image datasets, which includes a quantitative demonstration on the time-efficiency for our proposed *two-stage dimensionality reduction* algorithm.

FC Neural Network Models		NF-based Model
Encoder	**Decoder**	**Generator**
Input 2-dimensional sample	Input 1-dimensional embedding	Input 1-dimensional embedding
FC. 1024, ReLU	FC. 1024, ReLU	Pad. 2,
FC. 1024, ReLU	FC. 1024, ReLU	RT. 2, CP. 128, RT. 2
FC. 1	FC. 1024, ReLU	CP. 128, RT. 2
	FC. 2	CP. 128, RT. 2
		CP. 128, RT. 2
		CP. 128, RT. 2
		CP. 128, RT. 2

Table 1. The architectures we use for the synthetic manifold $\mathcal{M}$ as we introduced in main text. For VAE and IAE, the encoder $h : \mathbb{R}^2 \to \mathbb{R}^1$ and the decoder $g : \mathbb{R}^1 \to \mathbb{R}^2$ are fully-connected neural network models, where we use "FC. n" to represent a fully-connected layer with n output channels. For GAN, the generator and the discriminator use the architecture of the decoder and the encoder, respectively. For NF-based models, the generator $g = f_1 \circ p_1 : \mathbb{R}^1 \to \mathbb{R}^2$ consists of a *padding* module $p_1 : \mathbb{R}^1 \to \mathbb{R}^2$ and a *flow* module $f_1 : \mathbb{R}^2 \to \mathbb{R}^2$, where we use "Pad. n" to denote a padding module with the output dimensionality being n, use "RT. n" to denote a 1×1 convolutional layer with n channels, and use "CP. n" to denote an *affine coupling* layer with its nonlinear module being a convolutional architecture of "conv3. n, ReLU, conv3. n, ReLU, conv3. l", where we use "conv3. n" to denote a 3×3 convolution layer with output channel n, stride 1 and padding 1, and l is the number of channels of input to the "CP. n" layer.

Conv Neural Network Models		NF-based Model
Encoder	**Decoder**	**Generator**
Input $1 \times 28 \times 28$ image	Input $\in \mathbb{R}^K$	Input $\in \mathbb{R}^K$
$4 \times 4@(2,1)$ conv. 32, ReLU	$1 \times 1@(1,0)$ conv. 256, ReLU	Pad. $4 \times 7 \times 7$
$4 \times 4@(2,1)$ conv. 64, ReLU	$7 \times 7@(1,0)$ upconv. 64, ReLU	RT. 196, Flow@512@6
$4 \times 4@(1,0)$ conv. 128, ReLU	$4 \times 4@(2,1)$ upconv. 32, ReLU	Pad. $2 \times 14 \times 14$
$4 \times 4@(1,0)$ conv. 256, ReLU	$4 \times 4@(2,1)$ upconv. 3	RT. 8, Flow@256@6
$1 \times 1@(1,0)$ conv. K		Pad. $1 \times 28 \times 28$
		RT. 4, Flow@128@6

Table 2. The architectures we use for MNIST image datasets. For VAE and IAE, the encoder $h : \mathbb{R}^D \to \mathbb{R}^K$ and the decoder $g : \mathbb{R}^K \to \mathbb{R}^D$ are convolutional neural network work models, and we use "$k \times k@(s,p)$ conv. n" to represent a convolution layer with output channel n, kernel size k, stride s and padding p. The representation of a transposed convolution layer (denoted as upconv.) is similar. For GAN, the generator and the discriminator adopt the architecture of the decoder and encoder, respectively. For NF-based models, we use "Pad. $c \times h \times w$" to denote a padding module with output shape being $c \times h \times w$, and use "Flow@n@r" to denote a flow layer composed by r 1×1 convolutional layers and r *affine coupling* layers, where for each *affine coupling* layer, the nonlinear module is a convolutional architecture of "conv3. n, ReLU, conv3. n, ReLU, conv3. l" with l being the number of channels of the input to the *affine coupling* layer. We use "RT. n" to denote a 1×1 convolutional layer with n channels, and use $K = 10$.

Experimental Settings We introduce how we evaluate different models on the natural image datasets. In order to evaluate the *distance preserving* property, since the ground-truth geodesic length $d_{\mathcal{M}}(x_1, x_2)$ between two manifold samples x_1, x_2 is unknown, we only compute the metric score of Eq. (29) for all models. Similarly, since the ground-truth projection point $\widetilde{x}$ of an off-manifold x is unknown, the computation of Eq. (30) is no longer applicable for evaluating the *rigorous projection* property, hence we only compute the metric score of Eq. (31) and Eq. (32) for $\mathcal{M}$-flow and our proposed HF. For our HF with a multi-layer generator, the metric score of Eq. (31) is the average of the computed values of Eq. (31) for all *flow* modules f_i of the generator. To evaluate the *anomaly detection* performance, the computation of Eq. (33) is fundamentally applicable but the choices of $\mathcal{X}_{\text{in}}$ and $\mathcal{X}_{\text{out}}$ are different. Specifically, given a training dataset $\mathcal{X}$, we use $\mathcal{X}$ itself as the in-distribution dataset $\mathcal{X}_{\text{in}}$, and use $\mathcal{X}_{\text{out}} \triangleq \{x + \varepsilon | x \in \mathcal{X}\}$, where ε has the same dimensionality as x and each pixel of ε is a noise randomly sampled from $[-0.1, 0.1]$. We also provide qualitative results for intuitive comparison between different models.

Qualitative Results We provide qualitative results to visualize the quality and diversity of samples generated by using our HF model. In addition to train our HF model by using the objectives proposed in main text, we also adapt our HF model to an adversarial setting by replacing the reconstruction loss (*i.e.*, $\mathcal{L}_{\text{v}}$ in main text) with an adversarial loss, where a StyleGAN2 [6]

Algorithm 1 Manifold Training

Require: Training dataset $\mathcal{X} = \left\{x^{(i)}\right\}_{i=1}^{N} \subset \mathbb{R}^D$, embedding dimensionality K, generator $g : \mathbb{R}^K \to \mathbb{R}^D$ of L layers, singular values predictor $\left\{s_i : \mathbb{R}^{K_i} \to \mathbb{R}^{K_i}\right\}_{i=1}^{L}$, ambient radius r, hyper-parameters λ_{dist}, λ_{proj} and λ_{v}, all modules parameters Θ, optimizer ρ

Initialize running statistics $\mu_j \leftarrow 0, \sigma_j^2 \leftarrow 1, j \in [K]$
while not converge **do**
  Randomly select a batch of training samples $\mathcal{B} \subset \mathcal{X}$
  Obtain $u(x)$ and $\left\{v^i(x)\right\}_{i=1}^{L}$ by running inference for $x \in \mathcal{B}$ using $g^\dagger$. Compute $\mathcal{L}_{\text{v}}$ for $\mathcal{B}$
  Randomly sample a batch $\mathcal{S}^1 = \left\{u \sim \mathcal{N}\left(0, \sigma^2\right)\right\}$ and a batch $\mathcal{T}^i = \left\{v \in \mathcal{V}^i(r)\right\}$, then obtain

$$\mathcal{R}^i = \left\{f_{i-1}\left([u; v]\right) | u \in \mathcal{R}^{i-1}, v \in \mathcal{T}^i\right\} \tag{34}$$

  for $2 \leqslant i \leqslant L$ based on $\mathcal{R}^0 \triangleq \mathcal{S}^1$. Obtain $\mathcal{Z}^i \triangleq \mathcal{R}^{i-1} \times \mathcal{T}^i$. Compute $\mathcal{L}_{\text{dist}}$ and $\mathcal{L}_{\text{proj}}$ based on $\mathcal{Z}^i$, $i \in [L]$
  Compute $\mathcal{L} \leftarrow \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{v}}\mathcal{L}_{\text{v}}$
  Update $\Theta \leftarrow \rho\nabla\mathcal{L}$
  Update $\mu_j \leftarrow 0.99\mu_j + 0.01\text{avg}\left\{u_j(x) | x \in \mathcal{B}\right\}$
  Update $\sigma_j^2 \leftarrow 0.99\sigma_j^2 + 0.01\text{var}\left\{u_j(x) | x \in \mathcal{B}\right\}$
end while

Algorithm 2 The Overall Training Algorithm

Require: Training dataset $\mathcal{X} = \left\{x^{(i)}\right\}_{i=1}^{N} \subset \mathbb{R}^D$, dimensionality K of the first stage, threshold t
/* Manifold Training Phase */
Perform the *two-stage dimensionality reduction* algorithm as we introduced in Alg. 1 in main text, obtaining the detected manifold dimensionality $\widetilde{K}$ and the generator $\widetilde{g} : \mathbb{R}^{\widetilde{K}} \to \mathbb{R}^D$. For each stage, the generator is trained by using Alg. 1
Obtain the embeddings $\mathcal{U} = \left\{\widetilde{g}^\dagger\left(x^{(i)}\right)\right\}_{i=1}^{N}$ using $\widetilde{g}$
/* Density Training Phase */
Initialize a density estimator $h : \mathbb{R}^{\widetilde{K}} \to \mathbb{R}^D$, obtaining the parameters Θ of h. Then, setup an optimizer ρ for Θ
while not converge **do**
  Randomly select a batch of training samples $\mathcal{B} \subset \mathcal{U}$
  Compute $\mathcal{L}_{\text{d}}$ for $\mathcal{B}$
  Update $\Theta \leftarrow \rho\nabla\mathcal{L}_{\text{d}}$
end while

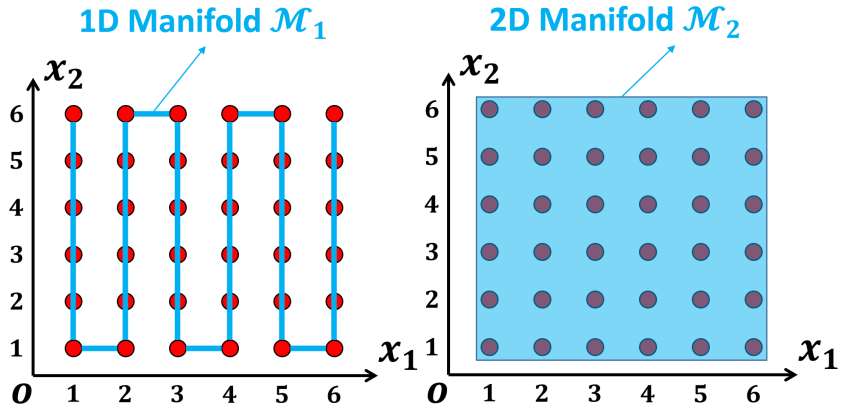
discriminator is involved. Our HF model in the adversarial setting is essentially a GAN framework augmented with the two isometric regularizations (*i.e.*, Eq. 16 and Eq. 17) we proposed in main text. As we can observe from Fig. 3, the adversarial setting results in more realistic, sharp and diverse samples, meanwhile the samples generated in the normal setting tends to be blurry. In Fig. 4, we also intuitively compare $\mathcal{M}$-flow with our HF in terms of the *rigorous projection* property by visualizing how the two models performs regarding generating samples extending to the ambient space. Specifically, for $\mathcal{M}$-flow, for a constant magnitude of change in v, the generated samples $f_1([u; v])$ should also change constantly if the model satisfies the *rigorous projection* property, as we analyzed in Sec. 2.1. We observe from Fig. 4 that our HF results in more constant change in generated samples in ambient space compared with $\mathcal{M}$-flow, and also generates samples with higher quality than $\mathcal{M}$-flow, which intuitively shows the effect of our isometric regularizations. For both models in the adversarial setting, we only show qualitative results, and for all quantitative evaluations below, we use the normal setting for both models.

Quantitative Results We provide quantitative results in Tbl. 3 to compare different models in terms of a series of metrics. We observe that the behaviors of different models are similar to those in the case of the aforementioned synthetic manifold. Specifically, we see that for natural image datasets, our proposed isometric regularizations Eq. 16 and Eq. 17 in main text are still effective in constraining the model to satisfy the properties of *distance preserving* and *rigorous projection*, respectively, and the performance of our HF on *anomaly detection* is superior due to the satisfaction of the *rigorous projection* property.

Model	Synthetic Manifold $\mathcal{M}$						CelebA					
	GAN	VAE	IAE	PIE	$\mathcal{M}$-flow	HF	GAN	VAE	IAE	PIE	$\mathcal{M}$-flow	HF
Distance Pres.↓	0.76	0.58	0.14	0.34	0.31	**0.09**	N/A	N/A	N/A	N/A	N/A	N/A
	1.38	0.76	0.17	0.52	0.48	**0.14**	10.23	9.56	3.28	9.45	4.56	**0.96**
Rigorous Proj.↓	N/A	0.45	0.38	0.42	0.46	**0.15**	N/A	N/A	N/A	N/A	N/A	N/A
	N/A	N/A	N/A	N/A	1.29	**0.12**	N/A	N/A	N/A	N/A	4.79	**1.03**
	N/A	N/A	N/A	N/A	0.36	**0.10**	N/A	N/A	N/A	N/A	0.56	**0.11**
					1.35	**0.13**					1.58	**0.15**
Anomaly Dete.↓	N/A	0.57	0.33	0.35	0.38	**0.11**	N/A	0.51	0.47	0.42	0.38	**0.15**

Model	FFHQ						AFHQ Cat					
	GAN	VAE	IAE	PIE	$\mathcal{M}$-flow	HF	GAN	VAE	IAE	PIE	$\mathcal{M}$-flow	HF
Distance Pres.↓	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	9.85	10.52	4.73	9.86	5.37	**1.02**	8.42	8.69	4.61	7.52	6.24	**0.94**
Rigorous Proj.↓	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	N/A	N/A	N/A	N/A	5.31	**1.02**	N/A	N/A	N/A	N/A	4.69	**1.04**
	N/A	N/A	N/A	N/A	0.52	**0.13**	N/A	N/A	N/A	N/A	0.49	**0.10**
					1.69	**0.14**					1.57	**0.14**
Anomaly Dete.↓	N/A	0.54	0.48	0.43	0.41	**0.14**	N/A	0.49	0.51	0.41	0.42	**0.12**

Table 3. Quantitative comparison between different models on different datasets in terms of various metrics. In terms of evaluating the *distance preserving* property, the first rows and the second rows are the global and local metric scores, respectively, where the global metric scores are computed using Eq. (27) or Eq. (28), and the local metric scores are computed using Eq. (29). In terms of evaluating the *rigorous projection* property, the metric scores of the first and second rows are computed using Eq. (30) and Eq. (32), respectively, and the third and forth rows are computed using ① and ② of Eq. (31), respectively. In terms of *anomaly detection*, the metric scores are computed using Eq. (33). For each metric, ↓ indicates that lower values are better, and the best results across all models are shown in bold.



(a) Recover $\mathcal{X}$ using a 1D manifold. (b) Recover $\mathcal{X}$ using a 2D manifold.

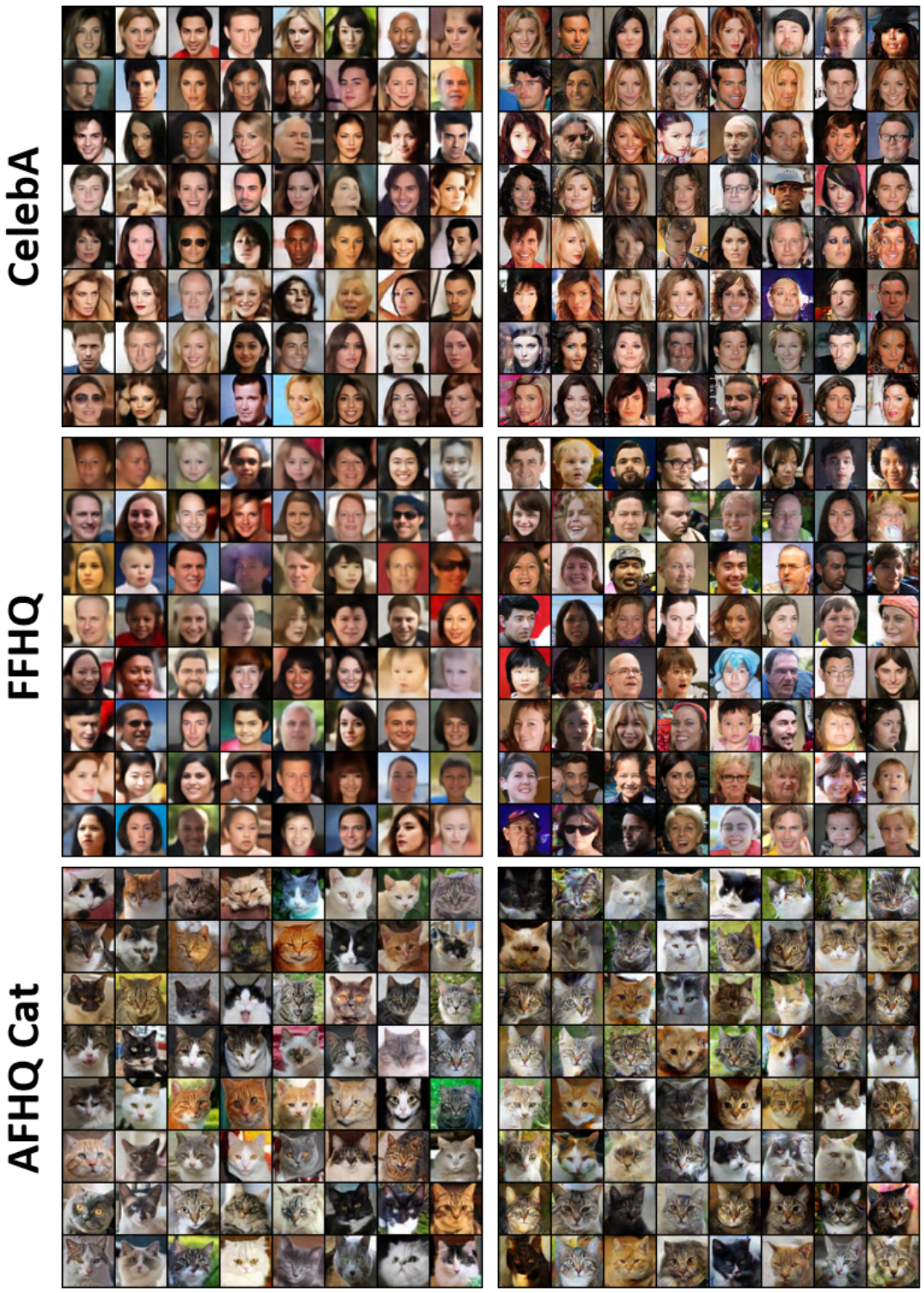
Figure 2. Given samples $\mathcal{X} \triangleq \{(i,j)\,|i,j \in [6]\}$, we demonstrate that $\mathcal{X}$ can be recovered by manifolds of arbitrary dimensions. We plot samples $\mathcal{X} \triangleq \{(i,j)\,|i,j \in [6]\}$ by using red points, and the recovered manifolds are shown in blue.

Dimensionality Reduction We demonstrate that our *two-stage dimensionality reduction* algorithm is an efficient and effective manner for detecting the ground-truth dimensionality of the data manifold. We firstly provide an intuitive understanding on how our proposed *distance preserving* property facilitates detecting the manifold dimensionality by using Fig. 2. Given training samples $\mathcal{X} \triangleq \{(i,j)\,|i,j \in [6]\}$ (see Fig. 2), we can recover $\mathcal{X}$ by using manifolds of arbitrary dimensions, as long as the model expressive power is unlimited. For example, we can recover $\mathcal{X}$ by using a 1D manifold $\mathcal{M}_1$ as shown in Fig. 2(a), and we can also recover $\mathcal{X}$ by using a 2D manifold $\mathcal{M}_2$ as we show in Fig. 2(b). Therefore, the criterion for dimensionality reduction based solely on the reconstruction error could be meaningless if the model expressive power is strong enough. We

Dataset	Search Process of $\widetilde{K}$	Time Cost
CelebA	$128 \rightarrow 127 \rightarrow \cdots \rightarrow 7 \rightarrow \mathbf{6} \rightarrow 5$	$124 \times 4.5\text{h} = 558\text{h}$
	$128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 4 \rightarrow \mathbf{6} \rightarrow 5$	$8 \times 4.5\text{h} = 36\text{h}$
	$128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 4 \rightarrow \mathbf{6} \rightarrow 5$	$4.5\text{h} + 7 \times 0.5\text{h} = 8\text{h}$
FFHQ	$128 \rightarrow 127 \rightarrow \cdots \rightarrow 9 \rightarrow \mathbf{8} \rightarrow 7$	$122 \times 4.5\text{h} = 549\text{h}$
	$128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow \mathbf{8} \rightarrow 4 \rightarrow 6 \rightarrow 7$	$8 \times 4.5\text{h} = 36\text{h}$
	$128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow \mathbf{8} \rightarrow 4 \rightarrow 6 \rightarrow 7$	$4.5\text{h} + 7 \times 0.5\text{h} = 8\text{h}$
AFHQ Cat	$128 \rightarrow 127 \rightarrow \cdots \rightarrow 5 \rightarrow \mathbf{4} \rightarrow 3$	$126 \times 4.5\text{h} = 567\text{h}$
	$128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow \mathbf{4} \rightarrow 2 \rightarrow 3$	$8 \times 4.5\text{h} = 36\text{h}$
	$128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow \mathbf{4} \rightarrow 2 \rightarrow 3$	$4.5\text{h} + 7 \times 0.5\text{h} = 8\text{h}$

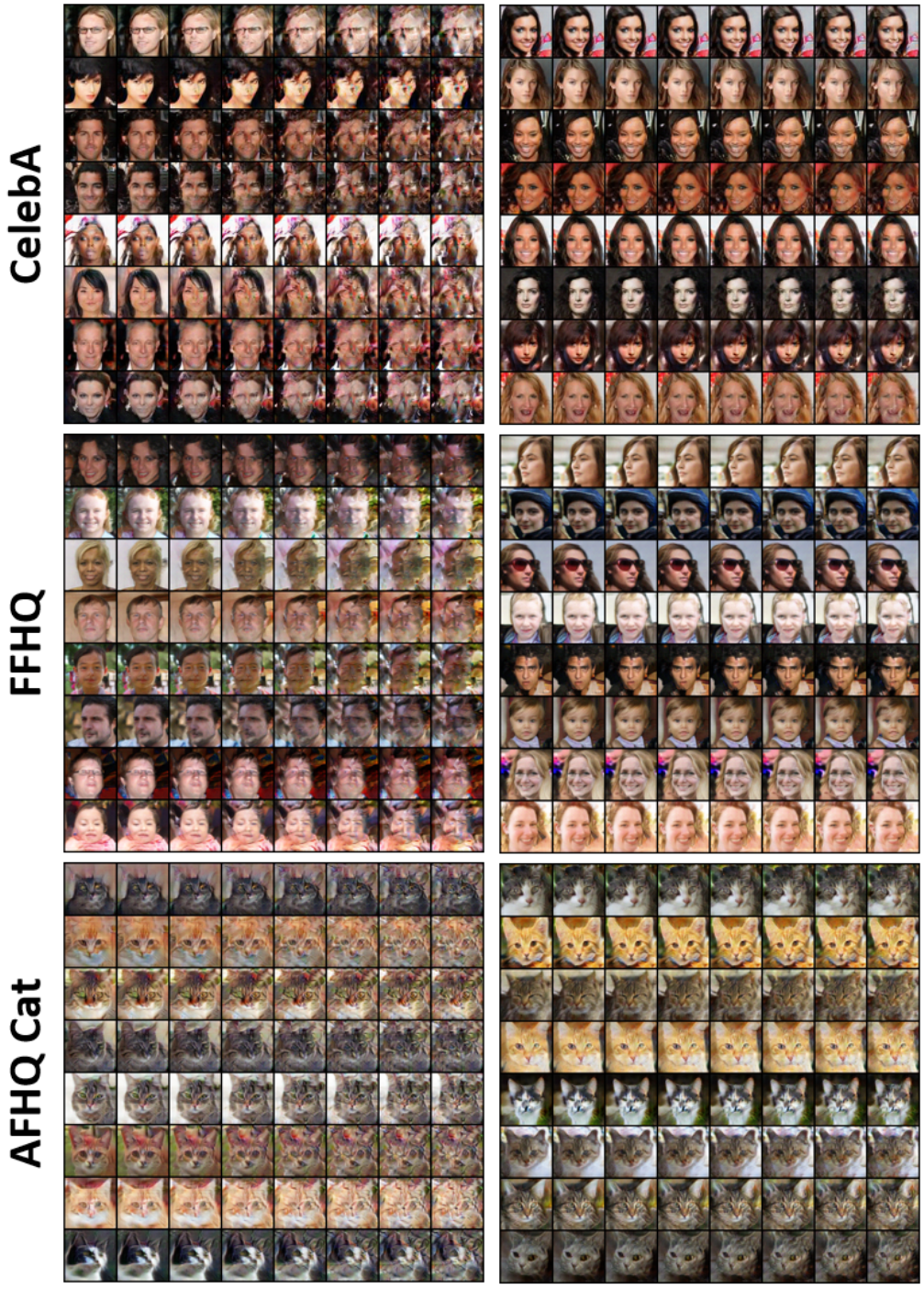
Table 4. Comparison between different strategies for searching the minimum dimensionality $\widetilde{K}$ using different models. The criterion used for determining the ground-truth dimensionality is not only based on a given reconstruction threshold t as we introduce in Alg. 1 in main text, but also based on the ratio $r \triangleq \frac{d_{\text{u}}\left(\widetilde{K}_t\right)}{d_{\text{u}}\left(\widetilde{K}_{t-1}\right)}$ with $d_{\text{u}}\left(\widetilde{K}_t\right)$ and $d_{\text{u}}\left(\widetilde{K}_{t-1}\right)$ corresponding to the latent dimensionalities of the current step t and the previous step $t-1$, respectively. Assume that the ground-truth manifold dimensionalities of CelebA, FFHQ and AFHQ Cat are 6, 8 and 4, respectively, and let the first-stage dimensionality K be 128. For each dataset, the first and second rows correspond to the search process using the brute-force algorithm and the binary search algorithm based on a $\mathcal{M}$-flow model, respectively, and the third row corresponds to the search process using the *two-stage dimensionality reduction* algorithm based on our HF model.

now show intuitively how our proposed *distance preserving* property helps alleviate this problem by using Fig. 2. We know that the ground-truth dimensionality of the manifold corresponding to $\mathcal{X}$ is 2. Let K be the dimensionality of the manifold to recover $\mathcal{X}$ induced by our model. For $K = 2$, assume that the model satisfies the *distance preserving* property, we know that $\max\left\{\left\|u\left(x_1\right) - u\left(x_2\right)\right\|_2 | x_1, x_2 \in \mathcal{X}\right\} = \sqrt{6^2 + 6^2} = 6\sqrt{2}$. However, for $K = 1$, given that the model satisfies the *distance preserving* property, we know that $\max\left\{\left\|u\left(x_1\right) - u\left(x_2\right)\right\|_2 | x_1, x_2 \in \mathcal{X}\right\} = 6 \times 7 = 42$, which is greatly larger than that of the case $K = 2$. Hence the metric $d_{\text{u}} \triangleq \max\left\{\left\|u\left(x_1\right) - u\left(x_2\right)\right\|_2 | x_1, x_2 \in \mathcal{X}\right\}$ increases dramatically if K is smaller than the ground-truth. However, for model such as $\mathcal{M}$-flow that is not guaranteed to satisfy the *distance preserving* property, the above analysis may not hold. Therefore, our proposed *distance preserving* property facilitates detecting the ground-truth dimensionality of the manifold. In terms of the time-efficiency of our proposed *two-stage dimensionality reduction* algorithm, in Tbl. 4, we compare different dimensionality reduction strategies, including the brute-force and binary search algorithms using $\mathcal{M}$-flow, and our proposed *two-stage dimensionality reduction* algorithm using our HF. We can learn that the time costs of the brute-force algorithm using $\mathcal{M}$-flow mentioned in [2] are totally unacceptable compared to the other two algorithms. Although $\mathcal{M}$-flow can speed up the search process by replacing the brute-force algorithm with the binary search algorithm, the one-layer design of $\mathcal{M}$-flow leads to training the generator $g : \mathbb{R}^{\widetilde{K}} \rightarrow \mathbb{R}^D$ involving a high-dimensionality D repeatedly, which still results in a higher cost of search time compared with our *two-stage dimensionality reduction* algorithm. Thanks to the hierarchical architecture design of our HF, the generator g can be composed as $g = g_1 \circ g_2$, where $g_1 : \mathbb{R}^K \rightarrow \mathbb{R}^D$ and $g_2 : \mathbb{R}^{\widetilde{K}} \rightarrow \mathbb{R}^K$ are the first-stage and second-stage generator, respectively. Though g_1 involves a high-dimensionality D, it is trained only once. Though g_2 is repeatedly trained during the search process, the dimensionalities $\widetilde{K}$ and K are small and hence the time costs are affordable. Therefore, combining the training of the first and second stage leads to a time-efficient algorithm for dimensionality reduction, as we demonstrated in Tbl. 4.



(a) Normal setting. (b) Adversarial setting.

Figure 3. Visualization of samples generated by our HF model, where the model generating samples on the right is trained by using the adversarial setting, and the model generating samples on the left is trained by using the normal setting as introduced in main text.



(a) $\mathcal{M}$-flow (b) HF

Figure 4. Qualitative comparison between $\mathcal{M}$-flow and our HF in terms of generating samples extending to the ambient space, where for each dataset, the samples on the left and right are generated using $\mathcal{M}$-flow and our HF, respectively, and we adopt the adversarial setting for both models. For each row of each model, the samples are generated as $f_1\left([u; v\varepsilon]\right)$ (*resp.*, u^{L+1} with $u^{i+1} \triangleq f_i\left(\left[u^i; v\varepsilon\right]\right)$ and $u^1 \triangleq u$) for $\mathcal{M}$-flow (*resp.*, our HF with a L-layer generator), where u is shared across the row, ε is a random unit vector, $v \in [0, 0.5]$, and the leftmost and rightmost samples correspond to $v = 0$ and $v = 0.5$, respectively. For intuitive observation, the magnitude of the change in v between any two adjacent samples is constant.

References

[1] J. J. Beitler, I. Sosnovik, and A. Smeulders. Pie: Pseudo-invertible encoder. *arXiv preprint arXiv:2111.00619*, 2021.

[2] J. Brehmer and K. Cranmer. Flows for simultaneous manifold learning and density estimation. *Advances in Neural Information Processing Systems*, 33:442–453, 2020.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[4] A. Gropp, M. Atzmon, and Y. Lipman. Isometric autoencoders. *arXiv preprint arXiv:2006.09289*, 2020.

[5] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[10] Z. Pan, L. Niu, and L. Zhang. Unigan: reducing mode collapse using a uniform generator. In *NeurIPS*, 2022.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[12] P. Petersen. *Riemannian geometry*, volume 171. Springer, 2006.