Geometric Inductive Biases for Identifiable Unsupervised Learning of Disentangled Representations

Ziqi Pan, Li Niu*, Liqing Zhang*

MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{panziqi_ai, ustcnewly}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract

The model identifiability is a considerable issue in the unsupervised learning of disentangled representations. The PCA inductive biases revealed recently for unsupervised disentangling in VAE-based models are shown to improve local alignment of latent dimensions with principal components of the data. In this paper, in additional to the PCA inductive biases, we propose novel geometric inductive biases from the manifold perspective for unsupervised disentangling, which induce the model to capture the global geometric properties of the data manifold with guaranteed model identifiability. We also propose a *Geometric Disentangling Regularized AutoEncoder* (GDRAE) that combines the PCA and the proposed geometric inductive biases in one unified framework. The experimental results show the usefulness of the geometric inductive biases in unsupervised disentangling and the effectiveness of our GDRAE in capturing the geometric inductive biases.

1 Introduction

Learning disentangled representation has attracted considerable research interest in the field of representation learning. Disentangled representation is considered to be interpretable with each dimension relevant to one generative factor of the data yet irrelevant to other generative factors (Higgins et al. 2018; Liu et al. 2021). Disentangled representations are considered to be more generalizable (Van Steenkiste et al. 2019), semantically meaningful, and thus useful for various downstream tasks (Bengio, Courville, and Vincent 2013).

The unsupervised learning of disentangled representations is more preferable than a supervised manner since acquiring ground-truth labels of generative factors is expensive. Given an unlabelled dataset, unsupervised disentangling methods aim to learn a generative model that aligns latent dimensions with ground-truth generative factors (Higgins et al. 2017). Classical methods such as PCA (Wold, Esbensen, and Geladi 1987) and ICA (Comon 1994; Theis 2006) are based on algebraic and statistical approaches, and more recent methods are based on neural network architectures and deep learning approaches (Zietlow, Rolinek, and Martius 2021), such as *Generative Adversarial Nets* (Goodfellow et al. 2014) (GANs) (Chen et al. 2016) and *Variational Autoencoders* (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) (VAEs). Compared to other methods, VAE-based models dominate unsupervised disentangling, and various VAE variants (Higgins et al. 2017; Kim and Mnih 2018; Chen et al. 2018) are proposed. However, the identifiability issue of VAE-based models is pointed out by (Locatello et al. 2019a), namely unsupervised disentangling is fundamentally impossible without inductive biases both on the models and datasets. To tackle this issue, recent works turn to employing weak supervision (Shu et al. 2020; Bouchacourt, Tomioka, and Nowozin 2018; Gabbay and Hoshen 2019; Hosoya 2019; Chen and Batmanghelich 2020; Locatello et al. 2019b; Klys, Snell, and Zemel 2018; Paige et al. 2017), altering the variational family and prior distribution (Kumar and Poole 2020; Mathieu et al. 2019), and introducing auxiliary variables (Kim, Guerrero, and Pavlovic 2021; Khemakhem et al. 2020; Mita, Filippone, and Michiardi 2021; Hyvarinen, Sasaki, and Turner 2019; Hälvä et al. 2021). Another line of research focuses on revealing inner mechanisms and inductive biases (Ridgeway 2016) for VAE-based models, and the PCA inductive biases are demystified (Rolinek, Zietlow, and Martius 2019; Zietlow, Rolinek, and Martius 2021; Bao et al. 2020) recently. It has been analyzed that the PCA inductive biases induce VAEs to improve local alignment of latent dimensions with principal components of the data (Zietlow, Rolinek, and Martius 2021; Rakowski and Lippert 2021). In this paper, we propose novel geometric inductive biases for unsupervised disentangling that induce the model to capture the global geometric properties of the data manifold, and the model identifiability is proven to be rigorously guaranteed. We also propose an unsupervised disentangling framework named *Geometric Disentangling Regularized AutoEncoder* (GDRAE), which combines the PCA and the proposed geometric inductive biases in one unified framework.

In the remainder of this paper, we introduce our proposed geometric inductive biases and GDRAE for unsupervised disentangling in Sec. 2-3, and discuss their relations to existing works in Sec. 4. We show experimental results in Sec. 5, and conclude the paper in Sec. 6.

2 Geometric Inductive Biases

We first briefly introduce some fundamentals of Riemannian geometry. For a thorough reference, refer to (Petersen 2006).

*Corresponding author

Given a manifold $\mathcal{M}$ and an open set $U \subset \mathcal{M}$, if there exists an open set U' of K-dimensional Euclidean space $\mathbb{R}^K$ and a homeomorphism $\varphi : U \to U'$, the pair (U, φ) is called a *coordinate chart* of $\mathcal{M}$. $\forall p \in U$, the *local coordinate* is given by $\varphi(p) = \left(z^1(p), z^2(p), \cdots, z^K(p)\right) \in \mathbb{R}^K$, and the *tangent vector* of the coordinate curve $\frac{\partial}{\partial z^i} : \mathrm{C}^r \to \mathbb{R}$ is defined as $\frac{\partial}{\partial z^i}(f) \triangleq \frac{\partial f \circ \varphi^{-1}}{\partial z^i}, \forall f \in \mathrm{C}^r$, where $f : U \to \mathbb{R}$ is a C^r function if it has r order continuous derivatives. The *tangent space* $T_p\mathcal{M}$ is the set of all tangent vectors at p, and $T\mathcal{M} \triangleq \cup_{p \in \mathcal{M}} T_p\mathcal{M}$. The *tangent bundle* $\pi : T\mathcal{M} \to \mathcal{M}$ is a mapping that maps each tangent vector of $T_p\mathcal{M}$ to p, and a *section* $s : \mathcal{M} \to T\mathcal{M}$ is a mapping such that $\pi \circ s = \mathrm{id}$. A set of sections $\{s_i\}_{i=1}^K$ is a *frame field* if $\{s_i(p)\}_{i=1}^K$ is a basis of $T_p\mathcal{M}$ at $\forall p \in \mathcal{M}$. A *tangent vector field* on $\mathcal{M}$ is a mapping $X : \mathcal{M} \to T\mathcal{M}$ such that $X(p) \in T_p\mathcal{M}, \forall p \in \mathcal{M}$, and $\mathfrak{X}(\mathcal{M})$ denotes the set of all tangent vector fields on $\mathcal{M}$. The *connection* is a mapping $\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$ and $\nabla_X Y$ gives the *covariant derivative* of Y along X. Given $\gamma : I \to \mathcal{M}$ as a smooth curve where I is an open interval, $X \in \mathfrak{X}(\mathcal{M})$ along γ is *parallel* if $\nabla_{\gamma'(t)} X \equiv 0$.

Notations We then introduce mathematical notations used throughout the paper. We consider decoders $f : \mathcal{Z} \to \mathbb{R}^D$, where the latent space $\mathcal{Z} \triangleq \mathcal{R}^K \subset \mathbb{R}^K$ with $\mathcal{R} \triangleq [0, r]$, $K \leqslant D$. Given $z \in \mathbb{R}^K$, we define $z_{\vee j} \in \mathbb{R}^K$ where $z_{\vee j}^j = z^j$ and $z_{\vee j}^{l \neq j} = 0$. Given $t \in \mathbb{R}$, we define $t_{\wedge j} \in \mathbb{R}^K$ where $t_{\wedge j}^j = t$ and $t_{\wedge j}^{l \neq j} = 0$. We use $[k] \triangleq \{1, 2, \cdots, k\}$.

Model Identifiability Definition Similar to (Khemakhem et al. 2020), we firstly provide the Def. 1 of an equivalence relation between decoders to depict the model identifiability in Def. 2. Intuitively speaking, given two decoders f, g in an identifiable set Θ, g can be obtained by rearranging latent dimensions (performed by P) of f and redefining how each latent dimension affects the respective generative factor (performed by φ) based on f. For example, consider f generates human face images, where two different latent dimensions z^i and z^j control the azimuth and illumination of faces, respectively, and the azimuth changes from -45° to 45° as z^i changes from 0 to r, while the illumination changes from dark to light as z^j changes from 0 to r. For g, it could be that z^i controls the illumination while z^j controls the azimuth (*i.e.*, the two latent dimensions are exchanged), and the azimuth changes from 45° to -45° as z^j changes from 0 to r while the illumination changes from light to dark as z^i changes from 0 to r (*i.e.*, the correspondence between values of latent dimension and factors are also inverted). The change between f and g does not involve latent space rotations, which is coincident with the *axes-preserving* property proposed in (Rolinek, Zietlow, and Martius 2019).

Definition 1. *Consider two decoders $f, g : \mathcal{Z} \to \mathbb{R}^D$, the equivalence relation $\sim$ between f and g is defined as*

$$f \sim g \iff \exists P, \varphi, \text{ s.t. } g = f \circ \varphi \circ P, \tag{1}$$

where $P \in \mathbb{R}^{K \times K}$ is a permutation matrix, and

$$\varphi(z) \triangleq \left(\varphi_1\left(z^1\right), \cdots, \varphi_K\left(z^K\right)\right), \quad \forall z \in \mathcal{Z} \tag{2}$$

with φ_j maintaining a bijection from $\mathcal{R}$ to $\mathcal{R}$.

Definition 2 (model identifiability)**.** *Consider decoders set $\Theta \triangleq \left\{f : \mathcal{Z} \to \mathbb{R}^D\right\}$. We say that Θ is identifiable $\iff$ $\forall f, g \in \Theta, f \sim g$.*

2.1 α-Structure

We provide the Def. 3 of α-structure and the Thm. 1 claiming the model identifiability induced by the α-structure inductive biases (see supplementary for proof). As indicated by (Locatello et al. 2019a), *the unsupervised learning of disentangled representation is fundamentally impossible without inductive biases both on the learning approaches and the datasets* (Locatello et al. 2019a), we demonstrate that α-structure specifies inductive biases both on the data manifold and the model, hence it allows identifiable unsupervised disentangled representation learning. In terms of inductive biases on data manifold $\mathcal{M}$, Eq. (3) essentially implies that there exists a global coordinate chart $\left(\mathcal{M}, f^{-1}\right)$ such that the tangent vector of the coordinate curve z^j (*i.e.*, $\frac{\partial}{\partial z^j}|_p$) at any $p \in \mathcal{M}$ is only determined by $z^j(p)$, which is equivalent to saying that $\mathcal{M}$ *is spanned by a set of curves*. An intuitive demonstration is present in Fig. 1(a). In terms of the inductive biases on the decoder, we show that α-structure implies a model prototype as stated in Prop. 1 (see supplementary for proof), where Eq. (4) indicates that α-structure decoders should be those *generative models that each latent dimension affects generated samples independently*, which constitutes the inductive biases on the model.

Definition 3 (α-structure)**.** *Given a C^2 bijective mapping $f : \mathcal{U} \to \mathbb{R}^D$ where $\mathcal{U} \supset \mathcal{Z}$ is an open set, its generative manifold restricted on $\mathcal{Z}$ is denoted as $\mathcal{M} \triangleq f(\mathcal{Z})$. We say that f and $\mathcal{M}$ satisfy* α-structure*, or f and $\mathcal{M}$ is* α-related*, if the frame field $\left\{\frac{\partial}{\partial z^j}\right\}_{j=1}^K$ satisfies*

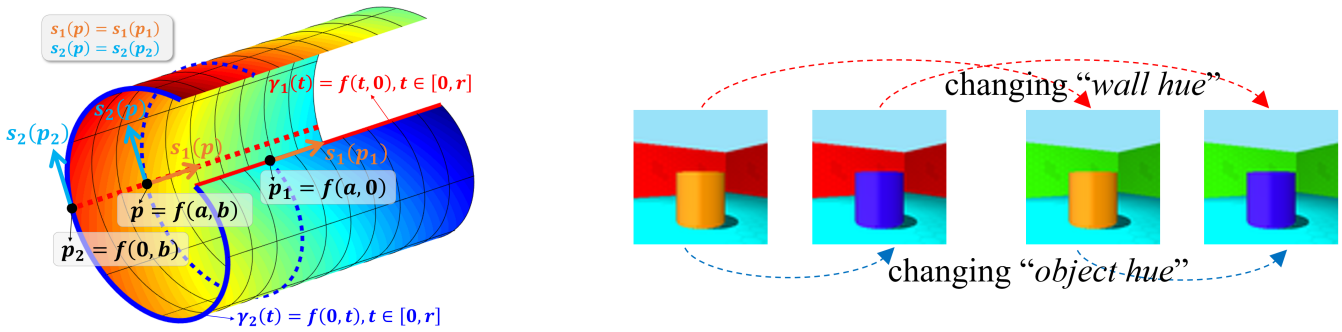
$$\left.\frac{\partial}{\partial z^j}\right|_p = \left.\frac{\partial}{\partial z^j}\right|_{f\left(z_{\wedge j}^j(p)\right)}, \quad \forall p \in \mathcal{M}, j \in [K]. \tag{3}$$

Theorem 1 (α-identifiability)**.** *Given an α-structure manifold $\mathcal{M}$, we denote $\Theta_\alpha^{\mathcal{M}} \triangleq \{f | f$ is α-related to $\mathcal{M}\}$. Then $\Theta_\alpha^{\mathcal{M}}$ is identifiable.*

Proposition 1 (α-structure model prototype)**.** *$f \in \Theta_\alpha^{\mathcal{M}} \iff \exists p \in \mathcal{M}$, C^2 bijective mappings $\left\{f_j : \mathcal{U} \to \mathbb{R}^D\right\}_{j=1}^K$ with $\mathcal{U} \supset \mathcal{Z}$ being an open set, such that*

$$f(z) = p + \sum_{j=1}^K f_j\left(z_{\wedge j}^j\right), \quad \forall z \in \mathcal{Z}. \tag{4}$$

Though the identifiability of α-structure models is guaranteed, one may concern if the α-structure inductive biases can be applied to unsupervised disentangling in real scenarios, since not only the inductive biases on the data manifold may not be applicable to certain datasets (*i.e.*, it is hard to satisfy that the data manifold $\mathcal{M}$ is spanned by a set of curves), but also the inductive biases on the decoder (*i.e.*, Eq. (4)) can restrict the flexibility of the decoder. Though the α-structure inductive biases are not universally applicable to real scenarios, we observe some cases where the α-structure inductive biases can be applied. We consider images datasets $\mathcal{D} = \left\{x \in \mathbb{R}^{H \times W \times C}\right\}$. For certain generative



(a) α-structure manifold.

(b) Applicability in 3DShapes dataset.

Figure 1: We provide a demonstrative 2D α-structure manifold $\mathcal{M}$ as in Fig. 1(a). Let $\left(\mathcal{M}, f^{-1}\right)$ be a global coordinate chart of $\mathcal{M}$, then $\forall p \in \mathcal{M}, j \in [2]$, the tangent vector of the coordinate curve z^j (*i.e.*, $s_j(p) \triangleq \frac{\partial}{\partial z^j}|_p$) is identical to the corresponding tangent vector of the coordinate curve z^j at the "edge" of the manifold (*i.e.*, $s_j(p_j) = \frac{\partial}{\partial z^j}|_{f\left(z^j_{\wedge_j}(p)\right)}$), which means that $\mathcal{M}$ is spanned by two edge curves $\gamma_1(t)$ and $\gamma_2(t)$. In Fig. 1(b), we give a case of real data to which the α-structure is applicable by using the 3DShapes (Burgess and Kim 2019) dataset. For factors *object hue* and *wall hue*, they affect disjoint subspaces of an image, hence images varies only in these factors can be depicted by an α-structure data manifold.

factors of certain datasets, different factors only affect disjoint subspaces of $\mathbb{R}^{H\times W\times C}$, so in such a case Eq. (4) is permitted when different latent dimensions z^j align with different factors. For example, given a human face image from the CelebA (Liu et al. 2015) dataset, the factors of *smile degree* and *hair color* may affect disjoint subspaces of the image, since the spatial areas of *mouth* and *hair* are disjoint. For the 3DShapes (Burgess and Kim 2019) dataset, samples varying only in *floor hue*, *wall hue* and *object hue* can be depicted by an α-structure manifold, since spatial areas affected by these factors are also disjoint (see Fig. 1(b)). We provide experimental results demonstrating the applicability of the α-structure inductive biases in real scenarios in Sec. 5.

2.2 β-Structure

We provide the Def. 4 of β-structure and the Thm. 2 guaranteeing identifiability (see supplementary for proof). Like α-structure, the β-structure also specifies inductive biases both on the data manifold and the decoder. In terms of the inductive biases on the data manifold $\mathcal{M}$, Eq. (6) essentially implies that there exists a global coordinate chart $\left(\mathcal{M}, f^{-1}\right)$ such that starting from any $p \in \mathcal{M}$, the tangent vector of the coordinate curve z^j (*i.e.*, $\frac{\partial}{\partial z^j}$) along any smooth curve γ is *parallel*, where

$$\gamma : I \to \mathcal{M}, \quad \text{s.t. } z^j \circ \gamma(t) = z^j(p), \forall t \in I \subset \mathbb{R} \tag{5}$$

is a curve on the manifold such that z^j is equal for all curve points. We refer to this inductive biases on $\mathcal{M}$ as the *parallel transport* property of $\mathcal{M}$. We provide an intuitive demonstration in Fig. 2(a). To capture the parallel transport property of $\mathcal{M}$, the decoders should be those *generative models satisfying Eq.* (6), which is the inductive biases on decoders.

Definition 4 (β-structure)**.** *Given a smooth bijective mapping $f : \mathcal{U} \to \mathbb{R}^D$ where $\mathcal{U} \supset \mathcal{Z}$ is an open set, we denote $\mathcal{M} \triangleq f(\mathcal{Z})$ as its generative manifold restricted on $\mathcal{Z}$. Let ∇ be the Levi-Civita connection of $\mathcal{M}$. We say that f and $\mathcal{M}$ satisfy β-structure, or f and $\mathcal{M}$ is β-related, if*

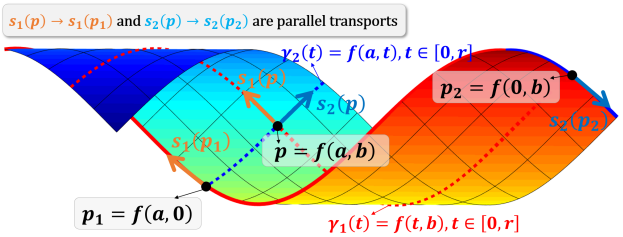
$$\nabla_{\frac{\partial}{\partial z^i}} \frac{\partial}{\partial z^j} \equiv 0, \quad \forall i, j \in [K] \wedge i \neq j. \tag{6}$$

Theorem 2 (β-identifiability)**.** *Given an β-structure manifold $\mathcal{M}$, we denote $\Theta^{\mathcal{M}}_{\beta} \triangleq \{f | f$ is β-related to $\mathcal{M}\}$. Then $\Theta^{\mathcal{M}}_{\beta}$ is identifiable.*

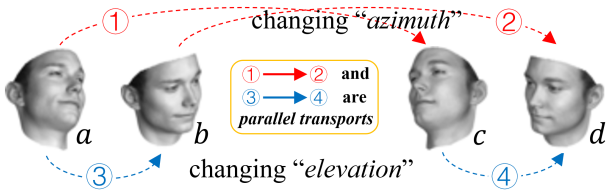
Starting from $p \in \mathcal{M}$ and along γ (Eq. (5)), the tangent vector $\frac{\partial}{\partial z^j}$ is required to be identical for the α-structure, while parallel transport is allowed for the β-structure. Hence the β-structure inductive biases offer more flexibility both on the data manifold shape and the decoder, and is more applicable in real scenarios. As shown by (Shao, Kumar, and Thomas Fletcher 2018), for images a, b and c from real data manifold such as CelebA (Liu et al. 2015) and SVHN (Netzer et al. 2011), parallel transport finds an image d that is related to c in the same semantic manner as a is related to b (Shao, Kumar, and Thomas Fletcher 2018), and such an analogy usually changes the same interpretable factors of the image. For example, let a, b and c be human face images with the factors *(blond hair, mouth closed)*, *(blond hair, mouth open)* and *(black hair, mouth closed)*, respectively, then d could be an image with factors *(black hair, mouth open)* (Shao, Kumar, and Thomas Fletcher 2018). Hence the parallel transport property (Eq. (6)) of the data manifold is reasonable and applicable in real scenarios, and a decoder capturing this property tends to align latent dimensions z^j with generative factors and thus disentangling. See Fig. 2(b) and experiments in Sec. 5 for applicability demonstration of the β-structure inductive biases.

3 The Model

In this section, we introduce our *Geometric Disentangling Regularized AutoEncoder* (GDRAE) that combines the PCA and the proposed geometric inductive biases in one unified framework. Based on the above introduction in Sec. 2, the



(a) β-structure manifold.



(b) Applicability in 3DFaces dataset.

Figure 2: We provide a demonstrative 2D β-structure manifold $\mathcal{M}$ as in Fig. 2(a). Let $\left(\mathcal{M}, f^{-1}\right)$ be a global coordinate chart of $\mathcal{M}$, then $\forall p \in \mathcal{M}$, the tangent vector of the coordinate curve z^1 (*i.e.*, $s_1(p)$) are parallel along the coordinate curve z^2 (*i.e.*, $\gamma_2(t)$), hence $s_1(p)$ and $s_1(p_1)$ are parallel. Similarly, the tangent vector of the coordinate curve z^2 (*i.e.*, $s_2(p)$) are parallel along the coordinate curve z^1 (*i.e.*, $\gamma_1(t)$), hence $s_2(p)$ and $s_2(p_2)$ are parallel as well. Fig. 2(b) demonstrates how parallel transport is applied in the 3DFaces (Paysan et al. 2009) dataset. Given images a and b such that b is obtained by changing *elevation* from a, parallelly transporting such a change to c gives an image d that is obtained by similarly changing *elevation* from c. The changing of *azimuth* can be parallelly transported as well.

α-structure inductive biases can be seen as a special case of the β-structure inductive biases, hence we consider the β-structure inductive biases when formalizing the training objectives of our GDRAE. To satisfy the β-structure (Eq. (6)) for the decoder, we provide the following Prop. 2 (see supplementary for the proof). To combine the PCA inductive biases, based on the analysis in (Zietlow, Rolinek, and Martius 2021) for VAE-based models, we found that given an orthogonal Jacobian J with K distinct nonzero singular values $\{\sigma_j\}_{j=1}^K$ as well as a bounded latent space $\mathcal{Z}$, minimizing $\sum_{j=1}^K \sigma_j^2$ captures the PCA inductive biases. See supplementary for a thorough analysis on how our GDRAE captures both the geometric and the PCA inductive biases.

Proposition 2. *Let $J(z) \in \mathbb{R}^{D\times K}$ be the Jacobian of the decoder g at $z \in \mathcal{Z}$, and $\{\sigma_j(z)\}_{j=1}^K$ be K singular values of $J(z)$. Eq. (6) is satisfied, if $\forall z \in \mathcal{Z}$, $J(z)$ is orthogonal and $\frac{\partial \sigma_j}{\partial z^i} = 0, \forall i, j \in [K] \wedge i \neq j$.*

We now introduce the proposed GDRAE and formalize the training objective. Our model involves an *encoder* $h : \mathcal{M} \to \mathcal{Z}$, a *decoder* $g : \mathcal{Z} \to \mathcal{M}$, and an auxiliary module named *singular value predictor* $s : \mathbb{Z} \to \mathbb{R}^K$. The *encoder* and *decoder* constitute an autoencoder, while the *singular value predictor* is a proxy module used for manipulating the singular values of the decoder Jacobian in an explicit manner to capture the PCA and the geometric inductive biases.

Autoencoding with bounded $\mathcal{Z}$ Given unlabelled dataset $\mathcal{X} = \left\{x^{(i)}\right\}$ sampled from an unknown data manifold $\mathcal{M}$, we constrain g to maintain a correspondence between $\mathcal{Z}$ and $\mathcal{M}$ by an autoencoding process with the following objective

$$\min_{g,h} \mathbb{E}_{x\in\mathcal{X}} \left(\|x - g \circ h(x)\|_2^2 + \underbrace{\|h(x)\| \cdot \mathbb{1}_{h(x)\notin\mathcal{Z}}}_{\mathcal{L}_{\text{bound}}(x)} \right), \tag{7}$$

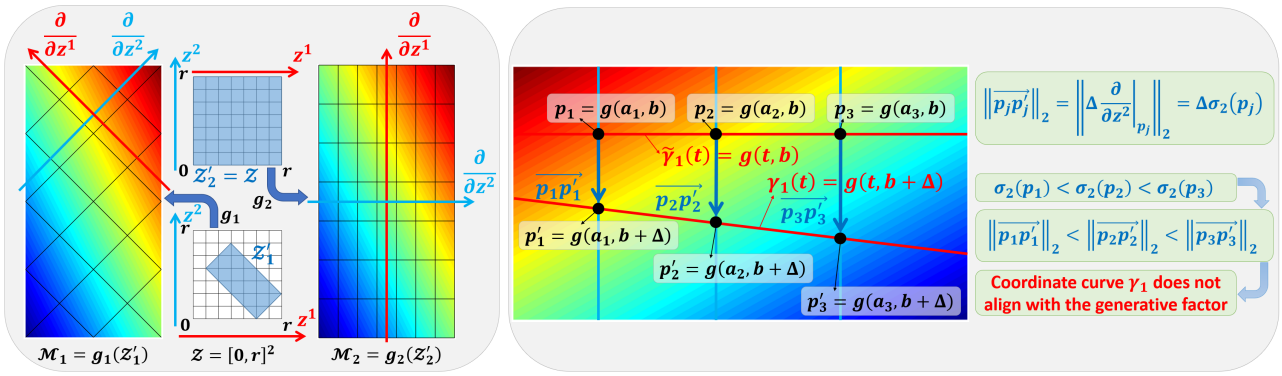
where we use $\mathcal{L}_{\text{bound}}(x)$ to constrain $h(x)$ to be within $\mathcal{Z}$, and $\mathbb{1}_{h(x)\notin\mathcal{Z}}$ is equal to 1 when $h(x) \notin \mathcal{Z}$, and 0 otherwise. It is notable that optimizing Eq. (7) does not guarantee that $h : \mathcal{M} \to \mathcal{Z}$ is a surjection, namely there could exist $z \in \mathcal{Z}$ such that $g(z) \notin \mathcal{M}$, which does not satisfy the prerequisite that g is a bijection between $\mathcal{Z}$ and $\mathcal{M}$ (see Fig. 3(a)). However, combined with the $\mathcal{L}_{\text{s\_norm}}$ regularization introduced below, minimizing Eq. (7) leads to a bijection g between $\mathcal{Z}$ and $\mathcal{M}$ (also see Fig. 3(a) for an intuitive understanding).

Regularizing Jacobian Our model involves the following Jacobian regularizations as aforementioned: 1) constraining $J(z)$ to be an orthogonal matrix with K distinct nonzero singular values $\{\sigma_j\}_{j=1}^K$, 2) constraining $\frac{\partial \sigma_j}{\partial z^i} = 0, \forall i, j \in [K] \wedge i \neq j$ (see Fig. 3(b) for an intuitive understanding), and 3) minimizing $\sum_{j=1}^K \sigma_j^2$. Our strategy is to employ an additional *singular value predictor* $s : \mathcal{Z} \to \mathbb{R}^K$ as a proxy module where $s(z)$ regresses to the ground-truth singular values $\{\sigma_j(z)\}_{j=1}^K$. In such a case, $\frac{\partial \sigma_j}{\partial z_i} = 0$ can be intrinsically satisfied by choosing proper architecture of s, and minimizing $\sum_{j=1}^K \sigma_j^2$ is also straightforward. Regarding constraining $\{\sigma_j\}_{j=1}^K$ to be distinct, we empirically found that in the case of autoencoding with bounded latent space $\mathcal{Z}$, minimizing $\sum_{j=1}^K \sigma_j^2$ commonly leads to $\{\sigma_j\}_{j=1}^K$ being distinct. Finally, we incorporate constraining $J(z)$ to be orthogonal and the regression of $s(z)$ into one objective as constraining

$$\widetilde{J}(z) \triangleq J(z)\,\text{diag}\left(\frac{1}{s_1(z)}, \cdots, \frac{1}{s_K(z)}\right) \tag{8}$$

to be orthogonal (*i.e.*, $\widetilde{J}^\top(z)\,\widetilde{J}(z) = I$, where I is an identity matrix) due to the following Prop. 3 (see supplementary for proof). In terms of constraining the Jacobian to be orthogonal, several approaches have been proposed by recent works (Qi et al. 2018; Karras et al. 2020). Given $z \in \mathcal{Z}$, we use $\mathcal{L}_{\text{ortho}}(z)$ to denote the regularization to constrain $\widetilde{J}(z)$ to be orthogonal. Due to space limitation, we provide the implementation details of $\mathcal{L}_{\text{ortho}}(z)$ in supplementary.

Proposition 3. *Given $z \in \mathcal{Z}$, $\widetilde{J}^\top(z)\,\widetilde{J}(z) = I \Rightarrow J(z)$ is orthogonal and $s_j(z) = \sigma_j, j \in [K]$.*



(a) Intuitively understanding $\mathcal{L}_{\text{bound}}$ and $\mathcal{L}_{\text{s\_norm}}$.

(b) Intuitively understanding $\frac{\partial \sigma_j}{\partial z^i} = 0, \forall i \neq j$.

Figure 3: We provide an intuitive understanding on the training objectives by considering 2D manifolds residing in 2D Euclidean space, namely $K = D = 2$. In Fig. 3(a), g_1 (*resp.*, g_2) maintains a bijection between $\mathcal{Z}'_1$ and $\mathcal{M}_1$ (*resp.*, $\mathcal{Z}'_2$ and $\mathcal{M}_2$), and the two latent directions z^1 and z^2 and their induced directions on the manifold are plotted using red arrow lines and blue arrow lines, respectively. By using $\mathcal{L}_{\text{bound}}$, we can constrain $\mathcal{Z}'_1$ and $\mathcal{Z}'_2$ to be subsets of $\mathcal{Z}$. We observe that for both g_1 and g_2, Eq. (7) is optimized (assume that $h = g^{-1}$). However, the latent dimensions of g_1 are not aligned with the directions of the generative factors of $\mathcal{M}_1$ (*i.e.*, the vertical and the horizontal direction). By further minimizing $\mathcal{L}_{\text{s\_norm}}$, the magnitude of $\sigma_j = \left\| \frac{\partial g_1}{\partial z^j} \right\|_2$ decreases, which makes $\mathcal{Z}'_1$ expand. In the case of $\mathcal{Z}'_1 = \mathcal{Z}$, the rotation asymmetry of $\mathcal{Z}$ guarantees the alignment between latent dimensions and the generative factors of the manifold given that the Jacobian is orthogonal, as shown by g_2. In Fig. 3(b), the coordinate curves of z^1 and z^2 are plotted by using red lines and blue lines, respectively, and the blue coordinate curves and $\widetilde{\gamma}_1(t)$ are in the vertical and horizontal direction, respectively. Assume that $\frac{\partial \sigma_2}{\partial z^1} > 0$, we have $\left\| \overrightarrow{p_1 p'_1} \right\|_2 < \left\| \overrightarrow{p_2 p'_2} \right\|_2 < \left\| \overrightarrow{p_3 p'_3} \right\|_2$, which results in a coordinate curve $\gamma_1(t)$ that does not align with the generative factor (*i.e.*, the horizontal direction).

The Overall Training Objective Summarizing the above gives the overall training objective of our GDRAE

$$\min_{g,h} \mathbb{E}_{x \in \mathcal{X}} \left(\underbrace{\|x - g \circ h(x)\|_2^2}_{\mathcal{L}_{\text{recon}}(x)} + \underbrace{\|h(x)\| \cdot \mathbb{1}_{h(x) \notin \mathcal{Z}}}_{\mathcal{L}_{\text{bound}}(x)} \right) + \min_{g,h,s} \mathbb{E}_{z \in \mathcal{Z}} \left(\underbrace{\|s(z)\|_2^2}_{\mathcal{L}_{\text{s\_norm}}(z)} + \mathcal{L}_{\text{ortho}}(z) \right). \tag{9}$$

The model can be trained in an end-to-end manner by using the above overall training objective.

4 Related Work

We discuss how the geometric inductive biases proposed in Sec. 2 and the GDRAE for unsupervised disentangling proposed in Sec. 3 relate to existing works.

Connection to PCA Inductive Biases Recently, the PCA inductive biases for VAE-based architectures were demystified by (Zietlow, Rolinek, and Martius 2021; Rolinek, Zietlow, and Martius 2019). Specifically, driven by local noises, VAE-based models recover the local principal components of the data that are well aligned with the ground-truth generative factors for some classical datasets (Zietlow, Rolinek, and Martius 2021). Such an alignment is done according to the local structure of the data. Differing from this mechanism, our proposed geometric inductive biases is shown to focus on the global structure of the data, which is facilitated by the following two aspects: 1) the latent space $\mathcal{Z}$ is a Cartesian products of K closed intervals $\mathcal{R}$, which is proven to play a vital role in resisting latent rotations (see supplementary), and 2) the two geometric structures proposed in Sec. 2 designate global properties of the data manifold, namely how tangent vectors of coordinate curves at different points relate to each other. As indicated by (Zietlow, Rolinek, and Martius 2021), local structures can be perturbed by noise and a small change of the local data distribution can potentially lead to a disruptive change in the alignment of latent dimensions with generative factors, while global variances of data can still remain unchanged in such a case, see Fig. 4. Therefore, global structures could be more reliable than local structures in unsupervised disentangling.

Connection to RAEs Our proposed model is essentially a deterministic autoencoder (Ghosh et al. 2020). The *Regularized Autoencoders* (Ghosh et al. 2020) (RAEs) provide a deterministic autoencoding framework, which alters the encoding process $x \to z$ from a reparameterizing (Kingma and Welling 2013) trick (*i.e.*, $z \sim \mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are Gaussian parameters produced by the encoder h) to directly outputting $z = h(x)$. Accordingly, the KL divergence in the ELBO (Hoffman and Johnson 2016) of VAE is replaced by minimizing $\|h(x)\|_2^2$. However, two major problems could be encountered. First, the mechanism of noise injection is a key factor in regularizing the decoder (Ghosh et al. 2020), and for a deterministic decoder, a small variation in latent space could result in dramatic change of the reconstruction.

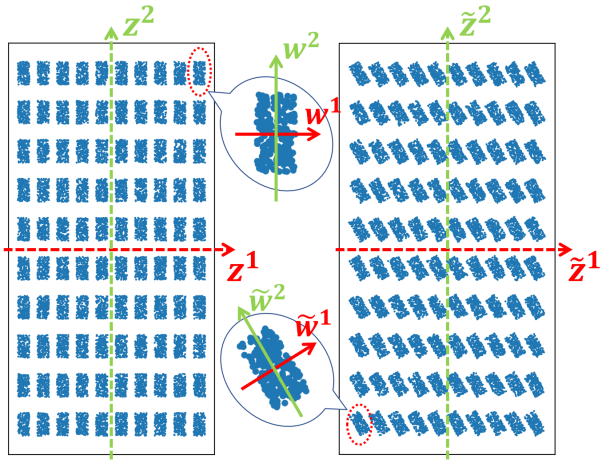


Figure 4: Intuitive illustration for the PCA inductive biases. For the data on the left (*resp.*, right), the global and local directions of generative factors are indicated by z (*resp.*, $\widetilde{z}$) and w (*resp.*, $\widetilde{w}$), respectively. The PCA inductive biases induce the model to align latent dimensions with local generative factors. Therefore, given that the local structure of the data are well aligned with the ground-truth generative factors (*i.e.*, the local and global directions of generative factors are well aligned) for most real datasets (Zietlow, Rolinek, and Martius 2021), the PCA inductive biases successfully disentangles on these datasets. However, as shown on the right, the PCA inductive biases fail to disentangle when the ground-truth generative factors (*i.e.*, $\widetilde{z}$) does not align with local directions of generative factors (*i.e.*, $\widetilde{w}$).

Second, due to the removal of the prior distribution, how to perform efficient sampling becomes implicit. To tackle the first problem, RAEs propose to smooth the latent space by imposing regularizations on the decoder, such as *spectral normalization* (Miyato et al. 2018), *gradient penalty* (Gulrajani et al. 2017), and *weight decay* (Krogh and Hertz 1992). In our model (Eq. (9)), we also regularize the decoder by minimizing its singular values, which is done by minimizing the $\mathcal{L}_{\text{s\_norm}}$ regularization. To tackle the second problem, RAEs perform density estimation on latent space ex-post to regain efficient sampling ability (Ghosh et al. 2020). In our model, we explicitly assign a bounded latent space $\mathcal{Z}$ and constrain $h(x)$ in $\mathcal{Z}$ instead of minimizing $\|h(x)\|_2^2$, therefore such an ex-post estimation process is unnecessary. Though it is not guaranteed that $\forall z \in \mathcal{Z}$, $g(z) \in \mathcal{M}$, where g and $\mathcal{M}$ are the decoder and real dataset respectively, we found that injecting Gaussian noise into z with fixed variance helps to alleviate this problem. Moreover, our bounded latent space structure plays a vital role in our identifiability analysis. In addition to the above, the *Gaussian RAE* (Kumar and Poole 2020) (GRAE) provides a deterministic approximations of β-VAE, which also involves encouraging orthogonal Jacobian as our model does. They approximate such a constrain by minimizing each Jacobian column norm (Kumar and Poole 2020). In our approach, we incorporate constraining the decoder Jacobian to be orthogonal and the regression of the predicted singular values of the decoder Jacobian to the ground-truth into one objective, which enables us to conveniently manipulate singular values of decoder Jacobian to capture the PCA and the geometric inductive biases.

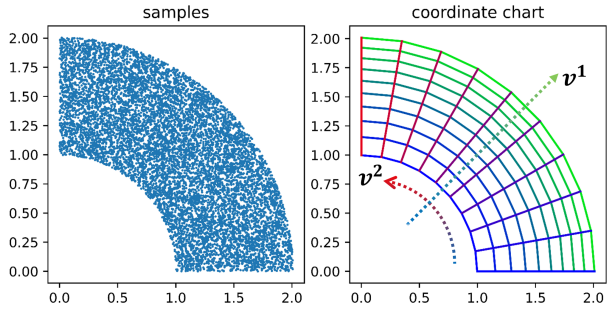


Figure 5: The 2D *sector* manifold.

5 Experiments

We provide experimental results showing that the proposed geometric inductive biases can be exploited for unsupervised disentangling, and that our GDRAE can capture the geometric inductive biases. Due to space limitation, we only provide demonstrative results in this section, while more experimental results can be found in supplementary.

We firstly conduce experiments on a synthetic manifold, showing that our proposed model is able to capture the geometric inductive biases of the manifold while β-VAE cannot. We use a 2D *sector* manifold $\mathcal{M}$ as in Fig. 5, where

$$\mathcal{M} \triangleq \left\{(x, y) \,|1 \leqslant \sqrt{x^2 + y^2} \leqslant 2\right\}. \tag{10}$$

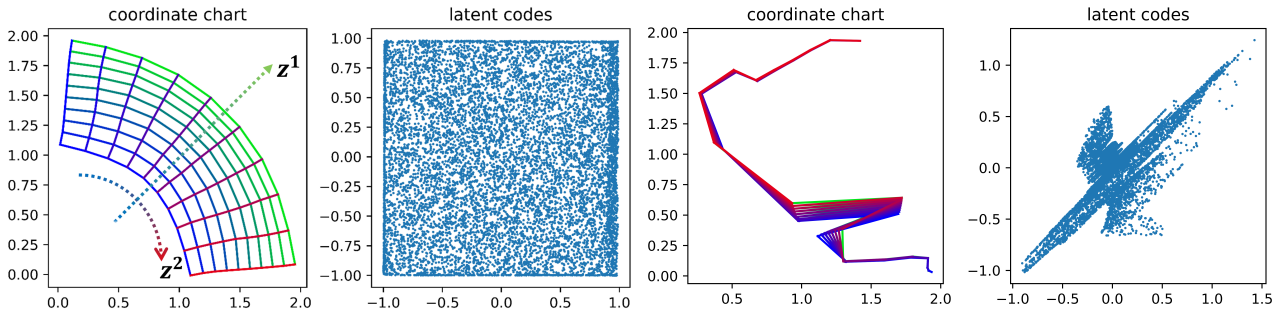
The global coordinate chart $(\mathcal{M}, \varphi)$ is induced by the polar coordinate transformation, namely

$$x = \left(\varphi^{-1}\right)^1 \left(v^1, v^2\right) = v^1 \cos v^2, \tag{11}$$

$$y = \left(\varphi^{-1}\right)^2 \left(v^1, v^2\right) = v^1 \sin v^2, \tag{12}$$

where $\left(v^1, v^2\right) \in [1, 2] \times \left[0, \frac{\pi}{2}\right]$. For $\mathcal{M}$, the primary and the secondary directions are along the angular direction (*i.e.*, v^2) and the radial direction (*i.e.*, v^1) respectively, hence v^1 and v^2 are two generative factors. In terms of inductive biases of $\mathcal{M}$, though global variances along v^1 and v^2 are distinct, the microscopic structure around $\forall p \in \mathcal{M}$ does not induce local PCA inductive biases. Hence for models successfully disentangling v^1 and v^2, inductive biases other than local PCA should be exploited. We also see that under $(\mathcal{M}, \varphi)$, $\mathcal{M}$ is a β-structure manifold under mild distortion (see supplementary), hence a model is unable to capture the geometric inductive biases of $\mathcal{M}$ if it fails to disentangle v^1 and v^2.

We present results on $\mathcal{M}$ obtained by using our GDRAE and a β-VAE (Higgins et al. 2017) model in Fig. 6. For both models, we set $K = 2$. For our model, we use a closed latent space $\mathcal{Z} \triangleq [-1, 1]^2$, while for β-VAE, the standard settings are chosen. In terms of disentangling, we visualize that while our model aligns latent dimensions z^1 and z^2 with the radial direction (*i.e.*, v^1) and the angular direction



(a) Recovered manifold using our proposed model. (b) Recovered manifold using β-VAE.

Figure 6: Results on the 2D *sector* manifold obtained by our proposed model and β-VAE. In Fig. 6(a), we visualize the global coordinate chart $\left(\widetilde{\mathcal{M}}, f^{-1}\right)$ of the manifold $\widetilde{\mathcal{M}}$ recovered by our proposed model f. For better visualizing the correspondence between coordinates of $\widetilde{\mathcal{M}}$ and latent space $\mathcal{Z}$, the color of the coordinate curve of $\widetilde{\mathcal{M}}$ along latent dimension z^2 changes from blue to green as z^1 changes from -1 to 1, and the color of the coordinate curve of $\widetilde{\mathcal{M}}$ along z^1 changes from blue to red as z^2 changes from -1 to 1. In Fig. 6(b), the same visualizing method is used for manifold recovered by β-VAE restricted on the latent region $[-0.5, 0.5]^2$. Additionally, we visualize latent codes of samples from $\mathcal{M}$ produced by the encoder. For β-VAE, the Gaussian means of posterior latent distribution are plotted.

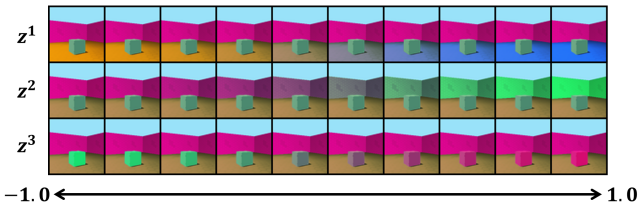


Figure 7: Visualization of latent traversal of an α-structure model trained on a subset of the 3DShapes datset with three generative factors: *object hue*, *wall hue* and *floor hue*. The latent traversal is done by first randomly sample a latent code $z \in \mathcal{Z}$, then varying each latent dimension z^i from -1 to 1 while keeping other latent dimension $z^{j \neq i}$ fixed, and plot the corresponding reconstruction.

(*i.e.*, v^2) respectively, β-VAE fails to do so. From Fig. 6(b), we learn that the manifold recovered by β-VAE is irregular, and no alignment between latent dimensions and generative factors is observed, though we have intensively tuned hyper-parameters for training β-VAE. Hence we learn that β-VAE is incapable of capturing the geometric inductive biases of $\mathcal{M}$, since it fails to disentangle. Meanwhile, further analysis in supplementary shows that our model exploits the geometric inductive biases to disentangle. We also observe that β-VAE is not able to perform efficient sampling on $\mathcal{M}$. Specifically, the decoded sample from a random latent code sampled from the prior distribution can not be guaranteed to locate on $\mathcal{M}$, because from Fig. 6(b) we observe that latent codes encoded from real samples distribute irregularly in latent space, which is partially related to the *posterior collapse* (Lucas et al. 2019a,b) problem of VAE. Meanwhile, our model tackles such a problem by explicitly specifying a bounded latent space $\mathcal{Z} = [-1, 1]^2$, therefore, it can always perform efficient sampling (see Fig. 6(a)).

We provide experimental results showing that the applicability of α-structure in real datasets, and that α-structure models can disentangle on α-structure manifold by exploiting the α-structure inductive biases. We provide results of an α-structure model (see Eq. (4)) trained on a subset of the 3DShapes (Burgess and Kim 2019) with generative factors: *object hue*, *wall hue* and *floor hue*, see Fig. 7. The model is constituted as in Eq. (4) (see supplementary for the detailed process of model construction). From Fig. 7, we see that the model aligns latent dimensions z^1, z^2 and z^3 with factors *floor hue*, *wall hue* and *object hue* respectively. Further analysis in supplementary shows that the independence of subdecoders (*i.e.*, f_j in Eq. (4)) is a key factor in encouraging alignment, which coincides the α-structure geometric property as in Eq. (3). The above results show that the given dataset can be depicted by an α-structure model, which verifies the applicability of α-structure in real data. It is also shown that the α-structure inductive biases can be exploited by an α-structure model for unsupervised disentangling.

6 Conclusion

In this paper, we propose two geometric inductive biases for unsupervised disentangling with guaranteed model identifiability from the manifold perspective. Our proposed geometric inductive biases induce the model to capture the global geometric properties of the manifold, namely how tangent vectors of coordinate curves are transported, which is different from the PCA inductive biases that improve local alignment of latent dimensions with nonlinear principal components based on the local structure of the data. We also propose the GDRAE model that combines the PCA and geometric inductive biases in one unified framework. Empirical results show the existence of the geometric inductive biases in real data and verify the effectiveness of our model in capturing the geometric inductive biases.

Acknowledgements

The work was supported by the National Science Foundation of China (62076162), and the Shanghai Municipal Science and Technology Major/Key Project, China (2021SHZDZX0102, 20511100300).

References

Bao, X.; Lucas, J.; Sachdeva, S.; and Grosse, R. B. 2020. Regularized linear autoencoders recover the principal components, eventually. In *NeurIPS*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bouchacourt, D.; Tomioka, R.; and Nowozin, S. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*.

Burgess, C.; and Kim, H. 2019. 3D Shapes Dataset. https://github.com/deepmind/3d-shapes/. Accessed: 2019-03-18.

Chen, J.; and Batmanghelich, K. 2020. Weakly supervised disentanglement by pairwise similarities. In *AAAI*.

Chen, R. T.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*.

Comon, P. 1994. Independent component analysis, a new concept? *Signal Processing*.

Gabbay, A.; and Hoshen, Y. 2019. Latent optimization for non-adversarial representation disentanglement. *arXiv preprint arXiv:1906.11796*.

Ghosh, P.; Sajjadi, M. S.; Vergari, A.; Black, M.; and Schölkopf, B. 2020. From variational to deterministic autoencoders. In *ICLR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *NeurIPS*.

Hälvä, H.; Le Corff, S.; Lehéricy, L.; So, J.; Zhu, Y.; Gassiat, E.; and Hyvarinen, A. 2021. Disentangling identifiable features from noisy data with structured nonlinear ica. In *NeurIPS*.

Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; and Lerchner, A. 2018. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Hoffman, M. D.; and Johnson, M. J. 2016. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *NeurIPS Workshop*.

Hosoya, H. 2019. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*.

Hyvarinen, A.; Sasaki, H.; and Turner, R. 2019. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *ICAIS*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*.

Khemakhem, I.; Kingma, D.; Monti, R.; and Hyvarinen, A. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *ICML*.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *ICML*.

Kim, M.; Guerrero, R.; and Pavlovic, V. 2021. Learning Disentangled Factors from Paired Data in Cross-Modal Retrieval: An Implicit Identifiable VAE Approach. In *ACM MM*.

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

Klys, J.; Snell, J.; and Zemel, R. 2018. Learning latent subspaces in variational autoencoders. In *NeurIPS*.

Krogh, A.; and Hertz, J. A. 1992. A simple weight decay can improve generalization. In *NeurIPS*.

Kumar, A.; and Poole, B. 2020. On Implicit Regularization in β-VAEs. In *ICML*.

Liu, X.; Sanchez, P.; Thermos, S.; O'Neil, A. Q.; and Tsaftaris, S. A. 2021. A Tutorial on Learning Disentangled Representations in the Imaging Domain. *arXiv preprint arXiv:2108.12043*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019a. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*.

Locatello, F.; Tschannen, M.; Bauer, S.; Rätsch, G.; Schölkopf, B.; and Bachem, O. 2019b. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*.

Lucas, J.; Tucker, G.; Grosse, R.; and Norouzi, M. 2019a. Understanding posterior collapse in generative latent variable models. In *ICLR*.

Lucas, J.; Tucker, G.; Grosse, R. B.; and Norouzi, M. 2019b. Don't blame the Elbo! a linear Vae perspective on posterior collapse. In *NeurIPS*.

Mathieu, E.; Rainforth, T.; Siddharth, N.; and Teh, Y. W. 2019. Disentangling disentanglement in variational autoencoders. In *ICML*.

Mita, G.; Filippone, M.; and Michiardi, P. 2021. An identifiable double VAE for disentangled representations. In *ICML*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. In *ICLR*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*.

Paige, B.; van de Meent, J.-W.; Desmaison, A.; Goodman, N.; Kohli, P.; Wood, F.; Torr, P.; et al. 2017. Learning disentangled representations with semi-supervised deep generative models. In *NeurIPS*.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *AVSS*.

Petersen, P. 2006. *Riemannian geometry*. Springer.

Qi, G.-J.; Zhang, L.; Hu, H.; Edraki, M.; Wang, J.; and Hua, X.-S. 2018. Global versus localized generative adversarial nets. In *CVPR*.

Rakowski, A.; and Lippert, C. 2021. Disentanglement and Local Directions of Variance. In *ECMLKDD*.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

Ridgeway, K. 2016. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*.

Rolinek, M.; Zietlow, D.; and Martius, G. 2019. Variational autoencoders pursue pca directions (by accident). In *CVPR*.

Shao, H.; Kumar, A.; and Thomas Fletcher, P. 2018. The riemannian geometry of deep generative models. In *CVPR Workshop*.

Shu, R.; Chen, Y.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Weakly supervised disentanglement with guarantees. In *ICLR*.

Theis, F. 2006. Towards a general independent subspace analysis. In *NeurIPS*.

Van Steenkiste, S.; Locatello, F.; Schmidhuber, J.; and Bachem, O. 2019. Are disentangled representations helpful for abstract visual reasoning? In *NeurIPS*.

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*.

Zietlow, D.; Rolinek, M.; and Martius, G. 2021. Demystifying Inductive Biases for (Beta-) VAE Based Architectures. In *ICML*.