Learning disentangled representations via product manifold projection

Marco Fumero [1]   **Luca Cosmo** [1 2]   **Simone Melzi** [1]   **Emanuele Rodolà** [1]

Abstract

We propose a novel approach to disentangle the generative factors of variation underlying a given set of observations. Our method builds upon the idea that the (unknown) low-dimensional manifold underlying the data space can be explicitly modeled as a product of submanifolds. This definition of disentanglement gives rise to a novel weakly-supervised algorithm for recovering the unknown explanatory factors behind the data. At training time, our algorithm only requires pairs of non i.i.d. data samples whose elements share at least one, possibly multidimensional, generative factor of variation. We require no knowledge on the nature of these transformations, and do not make any limiting assumption on the properties of each subspace. Our approach is easy to implement, and can be successfully applied to different kinds of data (from images to 3D surfaces) undergoing arbitrary transformations. In addition to standard synthetic benchmarks, we showcase our method in challenging real-world applications, where we compare favorably with the state of the art.

1. Introduction

Human intelligence is often understood as a process of turning experience into new behavior, knowledge, and skills (Locke, 1689). Humans achieve this by constructing small models of the world in their brains to explain the sensory experience and use them to infer new consequences. These models can be understood as turning experience into compact representations: functions of the experienced data which are useful for a given task. Representation learning (Bengio et al., 2014) aims to mimic this process with machines, via studying and formalizing what makes a good representation of high dimensional data, and how can we compute it in the form of an algorithm. A central topic in this research field is *disentangled* representations, which advocates that a representation of an entity should capture the different latent factors of variation in the world where that entity is observed. This abstract concept has led to several formalizations, but the community has not yet settled on a common definition.

In this work, we propose a new interpretation of disentanglement, based on the concept of projection over product spaces, and we provide an algorithm to compute disentangled representations by a weakly supervised approach. Since we do not require much knowledge of the data and their variations, our method fits different applications. Our main contributions can be summarized as follows:

- We reinterpret the notion of disentanglement in terms of geometric notions, and we show that our theoretical framework entails a generalization of current approaches.

- Relying on assumptions of the sample distribution, we provide a weakly supervised recipe that is simple to integrate into standard neural network models.

- We widely test our approach on synthetic datasets and more challenging real-world scenarios, outperforming the state of the art in several cases.

1.1. Related work

The task of identifying the underlying factors of variation in the data has been a primary goal in representation learning research (Bengio et al., 2014).

The problem of recovering independent components in a generative process relates to the independent component analysis (ICA). Given a multivariate signal, ICA aims to separate it into additive elements generated by independent non-Gaussian sources (Comon, 1994). While in the non-linear case, observing i.i.d. samples, ICA was proven to be unidentifiable (Hyvärinen & Pajunen, 1999), recent approaches (Hyvärinen et al., 2019; Hyvärinen & Morioka, 2016) have built on the assumption of having some degree of supervision to couple ICA and disentanglement.

A recent and promising line of research (Higgins et al., 2017; Kumar et al., 2018; Chen et al., 2018; Kim & Mnih, 2018) aimed to attain disentanglement in a fully unsupervised way,

[1]Sapienza, University of Rome, Rome, Italy [2]Università della Svizzera italiana, Lugano, Switzerland. Correspondence to: Marco Fumero <fumero@di.uniroma1.it>.

relying on a probabilistic interpretation according to which a representation is defined to be disentangled if the data distribution can be modeled as a nonlinear transformation of a product of independent probability distributions. Unfortunately, similarly to the ICA case, this leads to pessimistic results (Locatello et al., 2019) in terms of the identifiability of the factors without any assumptions on the model or the data (inductive bias). In this work we rephrase disentanglement in a new definition, entirely based on the geometric notion of metric and manifolds, and we rely on the assumption of observing non i.i.d. samples as data – thus escaping the non-identifiability result of (Locatello et al., 2019).

A more general definition of disentanglement in terms of group theory was given in (Higgins et al., 2018). Under their definition, a representation is disentangled if it decomposes in such a way that the action of a single subgroup leaves all factors of the representation invariant except for one. A first practical tentative in this direction was proposed by Pfau et al. (2020), which aims to discover a decomposition of a data manifold by investigating its holonomy group; but in order to work properly, the method requires to have full access to the metric of the data manifold. Differently, we provide a simple algorithm to obtain a disentangled representation in the form of a plug-in module for any autoencoder model.

By allowing supervision, one can escape to some extent the limitations underlined by Locatello et al. (2019), trading off identifiability of the model with scalability to realistic settings. The latter is due to the absence of labelled data in real scenarios. Supervision in disentanglement may come in different flavors, from weakly supervised settings (Locatello et al., 2020a; Shu et al., 2020), to semi-supervised (Locatello et al., 2020b) and fully supervised approaches (Kazemi et al., 2019; Kulkarni et al., 2015; Dubrovina et al., 2019; Cosmo et al., 2020). We position our work in the weakly supervised setting; we assume to observe non i.i.d. sample pairs which differ only by a projection on a single submanifold.

Closer to our framework is the line of research on manifold learning, a generalization of the classical field of linear dimensionality reduction to the nonlinear setting (Tenenbaum et al., 2000; van der Maaten & Hinton, 2008; Belkin & Niyogi, 2001; Roweis & Saul, 2000; McInnes et al., 2018). This line of works arises from the assumption that we observe the data in a high-dimensional Euclidean space $\mathcal{X} \subset \mathbb{R}^N$ sampled from a low-dimensional manifold $\mathcal{M}$, which can be approximately embedded into a low-dimensional Euclidean space $\mathcal{Z}$. We make this assumption stronger, by conjecturing that the data manifold is factorizable into a product of subspaces; each subspace models a generative factor of the data. A similar assumption was adopted in (Shukla et al., 2019), where the underlying manifold is assumed to be a product of orthogonal spheres.

Notably, in (Hu et al., 2018) it has been proposed to disentangle the factors of variation by mixing the latent vectors entries and imposing consistency constraints on the reconstructed images. In our work we adopt a similar constraint, but we require consistency in the latent space rather than on the reconstructions. Moreover, this does not represent the main ingredient to attain disentanglement in our framework. Finally, our work differs from all previous approaches as we consider factors of variation living on multi-dimensional submanifolds.

2. Theoretical justification

At the basis of our work is the assumption that high-dimensional data lie near a low-dimensional manifold, and that this can be factored into a product of submanifolds, each modelling one factor of variation. This factorization is not unique. We aim to find the one that best explains the data while being semantically valid, i.e. interpretable by humans. We denote by $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times ... \times \mathcal{M}_k$ the product manifold we want to approximate. In our context, we only have access to observations of $\mathcal{M} \subset \mathbb{R}^d$ in a high-dimensional space $\mathcal{X} = \mathbb{R}^N$, with $N \gg d$.

2.1. Background

Our definition of disentanglement is based upon the notion of metric. The manifold $\mathcal{M}$ is equipped with a metric $\mathfrak{g}_x : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$, where $\mathcal{T}_x\mathcal{M}$ is the tangent space at $x \in \mathcal{M}$; $\mathfrak{g}_x$ is a smoothly varying function that enables the computation of lengths, angles and areas on the manifold. In particular, the distance between two arbitrary points x, y on the manifold is defined in terms of the metric tensor: $d^2(x, y) = \int_0^1 \mathfrak{g}_{\gamma_t}(\dot{\gamma}_t, \dot{\gamma}_t)dt$ where $\gamma : [0, 1] \to \mathcal{M}$ is a geodesic parametrized curve with $\gamma(0) = x$ and $\gamma(1) = y$.

Since the metric allows us to compute distances, we can alternatively regard $\mathcal{M}$ as a metric space. To define distances on a *product* manifold, we recall the following definition for product metric spaces:

Definition 1 (Product metric space) *A product metric space is an ordered pair (M, d) where M is a product of sets $M_1 \times ... \times M_k$, each equipped with a function $d_i : M_i \times M_i \to \mathbb{R}$, $\forall i$ s.t. $\forall x, y, z \in M_i$:*

$$d_i(x, y) = 0 \iff x = y$$
$$d_i(x, y) = d_i(y, x)$$
$$d_i(x, z) \leq d_i(x, y) + d_i(y, z)$$

The metric on the product space corresponds to the L_2 norm of the metric on the subspaces.

Distances on a product manifold can then be factored into the sum of distances on the submanifolds, since the metric of the product manifold simply reduces to the sum of the met-
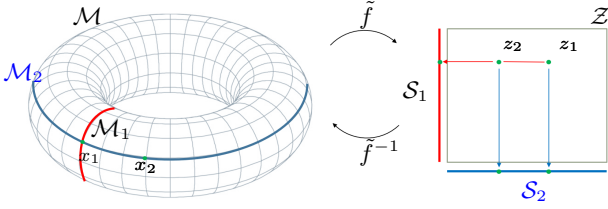
Figure 1. Let $\mathcal{X} \subset \mathbb{R}^3$ be our input space composed by points lying near a 2-dimensional torus $\mathcal{M}$. Since the torus is a product manifold of two circles, we can represent points on the torus as projections on the two submanifolds $\mathcal{M}_1$ and $\mathcal{M}_2$. Let $\tilde{f} : \mathcal{X} \to Z$ be an embedding into a subset of the plane $\mathcal{Z} \subset \mathbb{R}^2$. $\mathcal{Z}$ can be factored into two subspaces $\mathcal{S}_1 \subset \mathbb{R}$ and $\mathcal{S}_2 \subset \mathbb{R}$, which reflect the metric structure of $\mathcal{M}_1, \mathcal{M}_2$.

rics of its constituents $\mathfrak{g}(u,v) = \sum_i^K \mathfrak{g}_i(u_i, v_i)$, resulting in a product metric space (Lee, 2012).

2.2. Disentangled representations

We define a representation as *disentangled* if the variation of one generative factor in the data corresponds to a change in *exactly one* submanifold. Let us be given two data samples, and assume one of them is transformed by some unknown process. Then, the change in the distance between their projections onto the product manifold, measured before and after transforming the sample, is nonzero only on one submanifold. In other words, changing one factor of variation at a time will correspond to moving along a trajectory on a specific submanifold, while standing still on all the others. More formally:

Definition 2 *Let $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_k$ be a product manifold embedded in high dimensional space $\mathcal{X}$. A representation z in some space $\mathcal{Z} = S_1 \times \ldots \times S_k$, s.t. $dim(\mathcal{Z}) \ll dim(\mathcal{X})$, is* disentangled *w.r.t. $\mathcal{M}$ if there exists a smooth invertible function, with smooth inverse (a diffeomorphism) $\tilde{g} : \mathcal{Z} \to \mathcal{M}$ s.t. $\forall i \in 1, \ldots, k;\ \forall x_1, x_2 \in \mathcal{M}$:*

$$d_{\mathcal{M}_i}(x_1^i, x_2^i) > 0 \implies d_{\mathcal{S}_i}(s_1^i, s_2^i) > 0$$

$$d_{\mathcal{M}_i}(x_1^i, x_2^i) = 0 \implies d_{\mathcal{S}_i}(s_1^i, s_2^i) = 0 \implies s_1^i = s_2^i$$

where x_j^i is the projection of x_j on $\mathcal{M}_i$ and $s_j^i = \Pi_i \tilde{g}^{-1}(x_j)$ with Π_i being the projection onto the subspace $\mathcal{S}_i \subset \mathcal{Z}$.

The existence of $\tilde{g}^{-1}$ is guaranteed by the Whitney immersion theorem (Whitney, 1936), which states that for any manifold of dimension m there exists a one-to-one mapping in a Euclidean space of dimension at least $2m$ ($2m - 1$ if the manifold is smooth). Note that, if one has access to the metric of $\mathcal{M}_i$, a stricter definition of disentanglement can be enforced, requiring that the metric on the subspaces fully recovers the one on the submanifolds, i.e. $d_{\mathcal{M}_i}(x_1^i, x_2^i) \approx d_{\mathcal{S}_i}(s_1^i, s_2^i)$. The Nash-Kuiper embedding theorem (Nash, 1956) ensures the existence of the isometry

$\tilde{g}^{-1}$ in the latter case. Unfortunately, knowing the metric of $\mathcal{M}_i$ is unlikely in a practical setting.

Example. To better understand the definition we refer the reader to Figure 1, where we also introduce some notation. Let $\mathcal{X} \supset \mathcal{M}$ be our observation space, equipped with the standard Euclidean metric $d_e(\cdot, \cdot) = || \cdot - \cdot ||_2$. We consider the embedding function $\tilde{f} : \mathcal{X} \to \mathcal{Z}$, s.t. $\tilde{f} \approx \tilde{g}^{-1}$ on $\mathcal{M}$, where $\mathcal{Z}$ is equipped with the metric induced via $\tilde{f}$ and $dim(\mathcal{Z}) \ll dim(\mathcal{X})$. We consider the factorization $\mathcal{Z} = \mathcal{S}_1 \times \ldots \times \mathcal{S}_K$, where each subspace ideally corresponds to a parametric space for each of the $\mathcal{M}_i \subset \mathcal{M}$. We name the projections onto $\mathcal{S}_i$ as $\Pi_i : \mathcal{Z} \to \mathcal{S}_i\ \forall i \in \{1 \ldots K\}$. In Figure 1, the subsets $\mathcal{S}_1, \mathcal{S}_2$ approximate the metric structure of $\mathcal{M}_1, \mathcal{M}_2$ in the sense of Definition 2: z_1, z_2 projected onto $\mathcal{S}_1$ will correspond to the same point (i.e. their distance will be zero), reflecting the structure of $\mathcal{M}_1$, while their projection onto $\mathcal{S}_2$ will approximate $d_{\mathcal{M}_2}(x_1, x_2)$.

Furthermore, we can regard x_2 as the result of a transformation T applied to x_1, namely a translation on the submanifold $\mathcal{M}_2$ (in general, each pair of samples from $\mathcal{M}$ can be seen in this sense). This allows us to describe the example of Figure 1 via the commutative diagrams in Figure 2.

Figure 2. Commutative diagrams describing the relation between the transformation T on $\mathcal{M}$ that moves x_1 to x_2, the corresponding transformation T_z on $\mathcal{Z}$, and after the projections Π_1 (left diagram) and Π_2 (right diagram) on the subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively.

Can we learn disentangled representations? Our main objective is to obtain a disentangled representantion (with respect to Definition 2) without having any access to neither the metrics, nor the manifold $\mathcal{M} = \mathcal{M}_1 \times \ldots \times \mathcal{M}_k$ or its subspaces. We only assume to observe pairs of samples (x_1, x_2) in the high-dimensional data space $\mathcal{X}$.

Our assumption is that w.l.o.g. x_2 results from a transformation of x_1, $x_2 = T_i(x_1)$, corresponding to a translation over the submanifold $\mathcal{M}_i$ for some i. Moreover, the projection of (x_1, x_2) on the submanifold $\mathcal{M}_i$ is invariant to transformations T_j over $\mathcal{M}_j$. Specifically, the distance of the projection of the sampled pair on the corresponding submanifolds $\mathcal{M}_j$ is zero $\forall j \neq i$.[1]

Then, our objective is to learn the mapping $\tilde{f} : \mathcal{X} \to \mathcal{Z}$, s.t. $\tilde{f} \approx g^{-1}$ on $\mathcal{M}$, where $\mathcal{Z}$ is a low-dimensional product

---

[1]In practice, we relax this assumption to the case where multiple transformations can occur at the same time, i.e. $x_2 = T^*(x_1)$, where $T^* = T_1 \circ \ldots \circ T_k$ and at least one $T_i = Id$, $i = 1 \ldots k$.
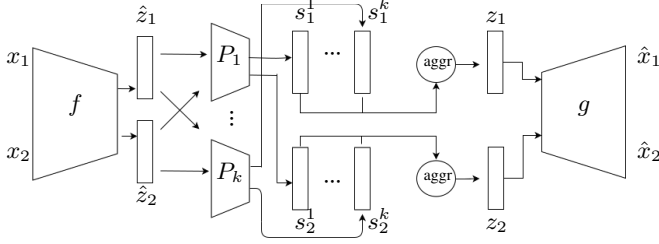
Figure 3. The architecture of our model. We process data in pairs (x_1, x_2), which are embedded into a lower dimensional space $\mathcal{Z}$ via a twin network f. The image (z_1, z_2) is then mapped into k smaller spaces $\mathcal{S}_1, \dots, \mathcal{S}_k \subset \mathcal{Z}$ via the nonlinear operators P_i. The resulting vectors are aggregated in $\mathcal{Z}$, with $aggr = +$, and mapped back to the input data space by the decoder g, to get $(\hat{x}_1, \hat{x}_2)$. As we do not impose any constraint on f and g, the intermediate module of the architecture could be attached to any autoencoder model. For a zoomed-in version see the Appendix.

space which acts as a parametric space for $\mathcal{M}$. Therefore, the mapping $\tilde{f}$ composed with the projections $\Pi_i \; \forall i \in 1, \dots, k$ must be *invariant* with respect to transformations T_j, and *equivariant* with respect to the transformation T_i. We answer the titular question in the following.

Relation with (Locatello et al., 2020a). The theoretical setting of Locatello et al. (2020a) is a special case of our framework, where their $\mathcal{S}_i$ are only 1-dimensional, the $s \in S_i$ parametrize a fixed (tipically Gaussian) distribution $P(s_i)$, and $d(a, b) = KL(P(s')\|P(s)) \; \forall s, s' \in \mathcal{S}$ is the Kullback-Leibler divergence. Assuming to have a sufficiently dense sampling of the data, and given that the map $\tilde{g}$ in Definition 2 is a diffeomorphism, these conditions are sufficient to prove identifiability for our model, reconnecting to the exact case proved in Locatello et al. (2020a).

3. Method

To implement our framework in practice, we approximate $\tilde{f}$ and its inverse with an encoder-decoder model where the encoder $f \approx \tilde{g}^{-1}$, and the decoder $g \approx \tilde{g}$. We impose the product space structure $\mathcal{Z} = \mathcal{S}_1 \times \dots \times \mathcal{S}_k$ on $\mathcal{Z}$ by adding a projection module in the latent space. This module is composed of nonlinear operators $P_i \; \forall i \in 1, \dots, k$, one for each factor of variation (we discuss the choice of k in the experimental section). The P_i act as nonlinear projectors, mapping from an *entangled*, intermediate latent space $\hat{\mathcal{Z}}$ onto the *disentangled* subspaces $\mathcal{S}_i \subset \mathcal{Z}$. We remark that the subspaces $\mathcal{S}_i$ act as parametric spaces for the latent submanifolds $\mathcal{M}_i \subset \mathcal{M}$, where each $\mathcal{M}_i$ characterizes a single factor of variation, in the sense of Definition 2.

Our proposed architecture is illustrated in Figure 3. As we model $\mathcal{Z} = \mathcal{S}_1 \times \dots \times \mathcal{S}_k$ as a product vector space, a latent vector in $\mathcal{Z}$ is obtained by aggregating the subspaces through a concatenation operation. Note that, in princi-

ple, the subspaces $\mathcal{S}_i$ could have different dimensionalities, which we are interested in learning. In practice, we set the same dimensionalities $dim(\mathcal{S}_i) = dim(\mathcal{Z}) = dim(\hat{\mathcal{Z}})$ for each subspace $\mathcal{S}_i$ and add a sparsity and orthogonality constraint to them, enforced by a specific term in our loss function. This results in the projectors $\mathcal{P}_i$ mapping onto disjoint subsets of the dimension indices for each subspace, thus allowing us to approximate the concatenation operation through the sum of the subspaces (proof in Appendix), which corresponds to set the operator $aggr = +$. We remark that other choices for $aggr$ may be possible according to the desired structure to impose on $\mathcal{Z}$.

3.1. Losses

We model f and g as neural networks parametrized by a set of parameters θ and γ respectively, while the nonlinear operators P_i are parametrized by ω_i. The model is trained by minimizing the composite energy $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{rec} + \beta_1(\mathcal{L}_{dis} + \mathcal{L}_{spar}) + \beta_2\mathcal{L}_{cons} + \beta_3\mathcal{L}_{reg} \quad (1)$$

w.r.t. the parameters θ, γ, ω_i, balanced by regularization parameters $\beta_1, \beta_2, \beta_3$. We now provide an explicit formula for each term and describe their role in the optimization.

Reconstruction loss

$$\mathcal{L}_{rec} = \|x - g_\gamma(\text{aggr}(P_{1,\omega_1}f_\theta(x), \dots, P_{k,\omega_k}f_\theta(x)))\|_2^2 \quad (2)$$

The reconstruction term captures the global structure of the manifold $\mathcal{M}$ by enforcing $f \approx g^{-1}$. We remark that invertibility also implies bijectivity on the data manifold.
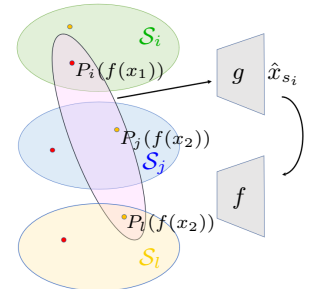
Consistency losses

$$\mathcal{L}_{cons} = \sum_{i=1}^{k} \|P_{i,\omega_i}(f_\theta(\hat{x}_{s_i})) - s_i\|_2^2, \quad (3)$$

with $s_i = P_i f(x_1)$ and $\hat{x}_{s_i} = g(\text{aggr}(P_i f(x_1), P_{j\neq i} f(x_2)))$.

The minimization of this loss makes the nonlinear operator P_i invariant to changes in the subspaces $S_j, \forall j \neq i$. This invariance is induced by aggregating, at training time, the representations $P_i(x_1)$ with the images of x_2 through P_j, $\forall j \neq i$ in the latent space $\mathcal{Z}$. This combination creates a new latent vector that differs from the encoding of x_1 only in the subspaces $S_j, \forall j \neq i$ as illustrated in the inset Figure. Forcing $P_i(f(x_1))$ and the image of the resulting latent vector via the composition $P_i \circ f \circ g$ to coincide in $\mathcal{S}_i$, is equivalent to

require that P_i is invariant to changes in the other subspaces $\mathcal{S}_j$; this promotes injectivity of the P_i.

Distance loss. This term, defined below, ensures the second property of Definition 2 and, coupled with the consistency constraint, is the key loss for constraining disentanglement in the representation.

Given a pair (x_1, x_2) and its latent representation (z_1, z_2), assume the existence of an oracle $\mathcal{O}$, that, acting on the latent vectors, tells us exactly which subspace $\mathcal{S}_i$ of $\mathcal{Z}$ contains the difference in the input pair. If $x_2 = T_i(x_1)$ for a fixed T_i acting on $\mathcal{M}_i$ then $\mathcal{O}(\hat{z}_1, \hat{z}_2) = i$. The oracle $\mathcal{O}$ can be implemented in practice by allowing a higher degree of supervision, i.e. by incorporating labels into the sampled data pairs, as done in (Zhan et al., 2019; Locatello et al., 2020b). However, in real settings, typically we do *not* have access to labels, thus $\mathcal{O}$ has to be estimated.

To estimate $\mathcal{O}$ in our weakly-supervised setting, we proceed as follows: (i) We compute a distance between the projected pair $d(s_1^i, s_2^i)$ in each subspace $\mathcal{S}_i$; (ii) we estimate the oracle as $\tilde{\mathcal{O}}(\hat{z}_1, \hat{z}_2) = \arg\max_i d(P_i(\hat{z}_1), P_i(\hat{z}_2))$, that corresponds to the subspace $\mathcal{S}_i$ where the projections differ the most. The distance in (i) should be general enough to be compared in the different subspaces. We select the Euclidean distance normalized by the average length of a vector in each subspace to provide a reliable measure, also considering different dimensions. We indicate with $\delta_i = d(P_{i,\omega_i}(\hat{z}_1), P_{i,\omega_i}(\hat{z}_2))$ the distance between the images on the i-th subspace, $\forall i \in 1, \ldots, k$.

We constrain the images onto the subspaces not selected by $\tilde{\mathcal{O}}$ to be close to each other, because they should ideally correspond to the same point as visualized in the inset Figure. To avoid the collapse of multiple subspaces into a single point, we insert a contrastive term (Hadsell et al., 2006) to balance the loss, which encourages the projected points in $\mathcal{S}_i$ to move away from each other. The resulting distance loss is therefore written as:

$$\mathcal{L}_{dis} = \sum_{i=1}^{k}(1-\alpha_i)\delta_i^2 + \alpha_i \max(m-\delta_i, 0)^2\,, \quad (4)$$

where $\alpha_i = 1$ if $\tilde{\mathcal{O}}(\hat{z}_1, \hat{z}_2) = i$ and $\alpha_i = 0$ otherwise; m is a fixed margin, which constrains the points to be at least at distance m from each other, and prevents the contrastive term from being unbounded.

Sparsity Loss

$$\mathcal{L}_{spar} = \sum_{i=1}^{k} \|P_{i,\omega_i}(f_\theta(x)) \odot \sum_{j\neq i}^{k} P_{j,\omega_j}(f_\theta(x))\|_1\,, \quad (5)$$

where $\odot$ denotes the element-wise product.

This loss combines an orthogonality and an ℓ_1-sparsity constraint on the latent subspaces. It allows us to set only the dimensionality of the latent space $\mathcal{Z}$ as an hyperparameter, while the algorithm infers the dimensionalities of the subspaces $\mathcal{S}_i, \forall i$. By minimizing this energy, we force the images of the P_i to have few non-zero entries. Together with the choice of the sum as aggregation operator for the $\mathcal{S}_i, i = 1 \ldots k$, this imposes the structure of a Cartesian product space on $\mathcal{Z} = \mathcal{S}_1 \times \ldots \mathcal{S}_k$, as we prove in the Appendix. Furthermore, this loss is equivalent to setting a constraint on the size of the learned latent space, which is needed to prevent unbounded optimization.

Training strategy. We train in two phases. First, we aim to learn the global structure of the manifold, relying solely on the reconstruction loss. Once we have given a structure to the latent space, we can start factorizing it into subspaces as the other terms in the loss enter the minimization. It is crucial that the space $\mathcal{Z}$ will continue to change as the other losses take over, since decoupling completely the reconstruction from the disentanglement could lead to not-separable spaces. The regularization parameters $\beta_i, i = 1, 2, 3$ are used to impose this behavior. For further details, we refer to the Appendix.

Regularization loss. In the first phase of the training, where we mainly aim to achieve good reconstruction quality, we have no guarantees that the information of the factors of variation is correctly spread among the subspaces, and this may lead to a single subspace encoding multiple factors of variation. To avoid this problem, we introduce a penalty that ensures the choice of the oracle $\tilde{\mathcal{O}}$ is equally distributed among the subspaces. In practice, the indicator variables in Eq.4 are approximated in each batch of N samples with a matrix $\mathbf{A}$ of dimensions $N \times k$, obtained by applying a weighted softmax to the distance matrix of pairs in each of the k subspaces. This matrix acts as a differentiable mask that implements the oracle $\tilde{\mathcal{O}}$ in practice. In the reconstruction phase, we activate a penalty on the matrix:

$$\mathcal{L}_{reg} = \sum_{j=1}^{k}\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{A}_{n,j} - \frac{1}{k}\right)^2\,, \quad (6)$$

which ensures that the selection of each subspace has the same probability.

4. Experimental results

To validate our model we perform experiments on both synthetic and real datasets. In the synthetic experiments, observations are generated as a deterministic function of the known factors of variation. For the real settings, since the data are collected directly from real observations, the parametric function generating the data is not known, and could have a different number of parameters than the number

of factors that we aim to disentangle. To put ourselves in a setting comparable to the competing methods, in all our experiments we use a latent space of dimension $d = 10$, unless otherwise specified, and $k = 10$ latent subspaces.

Synthetic data. We adopted 4 widely used synthetic datasets in order to evaluate the effectivness of our method, namely *DSprites* (Higgins et al., 2017), *Shapes3D* (Kim & Mnih, 2018), *Cars3D* (Reed et al., 2015), *SmallNORB* (LeCun et al., 2004) . All these datasets contain images that are parametrically rendered based on some known factors of variation, the goal is thus to obtain a disentangled latent space that reflects these parameters. We resized the input images to a dimension of 64×64 pixels. During training, we give as input to our method two images which differ only by a single randomly sampled factor of variation, and this is the *only* assumption that we make on the input data. We do not provide any information on which of the factors is changing between the two images. We show quantitative results on these datasets in Tables 1,2,3,4 comparing our method to the state-of-the-art model Ada-GVAE of (Locatello et al., 2020a), in terms of the metrics described below.

Real data. As a benchmark for real data, we test our method on FAUST (Bogo et al., 2014), a dataset composed by 100 3D human scans of 10 subjects in 10 different poses. Differently from the synthetic case, the data does not derive from a parametric model. This dataset is particularly challenging, since each of the two factors of variation (pose and identity) is difficult to embed in a 1-dimensional space; indeed, human parametric models define pose and identity using dozens of parameters (Loper et al., 2015). To speedup the training, we remeshed each shape to 2.5K points.

Evaluation metrics. To quantitatively evaluate our experiments we adopt a set of evaluation metrics widely used in the disentanglement literature: namely, the β-VAE score and the Factor VAE score, which measure disentanglement as the accuracy of a linear classifier that predicts the index of a fixed factor of variation and a majority vote classifier respectively (Higgins et al., 2017; Kim & Mnih, 2018); the DCI disentanglement score, which measures the entropy of the distribution on the single dimension of the representation vector (Eastwood & Williams, 2018); and the mutual information gap (MIG) (Chen et al., 2018) which for each dimension of the representation vector, first measures the mutual information w.r.t. the other dimensions and then takes the gap between the highest and second coordinate. Note that the aforementioned metrics have been defined to measure the disentanglement in a variational framework, wherein the factors of variation are assumed to be one-dimensional. This is not the case in our method, where each disentangled subspace is *not* required to be one-dimensional. Nevertheless, we can still apply these metrics in the aggregated latent space. This allows us to perform a direct comparison with

Table 1. Median disentanglement scores on Shapes3D.

METRIC	β-VAE	β-TCVAE	Ada-GVAE	Ours
BetaVAE	98.6%	99.8%	**100%**	**100%**
FactorVAE	83.9%	86.8%	**100%**	**100%**
DCI Disent.	58.8%	70.9%	94.6%	**99.0%**
MIG	22.0%	27.1%	56.2%	**63.7%**
MIG-PCA	-	-	-	73.5%
MIG-KM	-	-	-	69.2%

Table 2. Median disentanglement scores on DSprites.

METRIC	β-VAE	β-TCVAE	Ada-GVAE	Ours
BetaVAE	82.3%	86.4%	92.3%	**98.9%**
FactorVAE	66.0%	73.6%	84.7%	**88.5%**
DCI Disent.	18.6%	30.4%	47.9%	**49.7%**
MIG	10.2%	18.0%	**26.6%**	25.7%
MIG-PCA	-	-	-	20.9%
MIG-KM	-	-	-	17.9%

Table 3. Median disentanglement scores on Cars3D.

METRIC	β-VAE	β-TCVAE	Ada-GVAE	Ours
BetaVAE	100%	100%	**100%**	**100%**
FactorVAE	87.9%	90.2%	**90.2%**	89.9%
DCI Disent.	22.5%	27.8%	**54.0%**	48.9%
MIG	8.8%	12.0%	15.0%	**25.9%**
MIG-PCA	-	-	-	20.6%
MIG-KM	-	-	-	17.8%

Table 4. Median disentanglement scores on SmallNorb.

METRIC	β-VAE	β-TCVAE	Ada-GVAE	Ours
BetaVAE	74.0%	76.5%	87.9%	**99.2%**
FactorVAE	49.5%	54.2%	55.5%	**88.2%**
DCI Disent.	28.0%	30.2%	33.8%	**49.4%**
MIG	21.4%	21.0%	25.6%	**25.8%**
MIG-PCA	-	-	-	20.8%
MIG-KM	-	-	-	17.8%

methods working in the variational setting.

We also introduce two new metrics, which adapt the Mutual Information Gap (MIG) score to a multi-dimensional space. The former, which we denote by MIG-PCA, projects the latent subspaces onto the principal axis of variation through PCA (F.R.S., 1901). This way we can directly reconduct ourselves to the variational setting, where each factor is encoded by just one dimension of the latent space, and use directly the MIG score. The latter, denoted by MIG-KM, is based on the use of k-means in the multi-dimensional subspace to derive a one-dimensional discrete random variable. In practice, on each subspace, we extract b centroids using k-means on the considered evaluation samples. Each sample
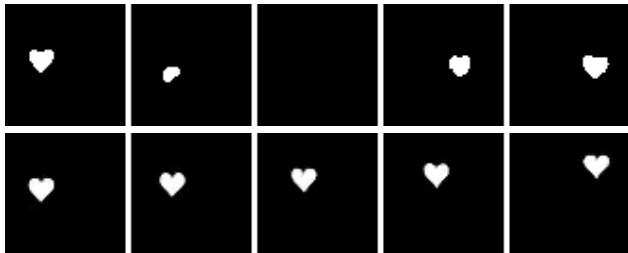
Figure 4. Interpolation example on the prevalent latent dimension / subspace of the changing factor on a pair of images from *Dsprites*; we compare Ada-GVAE (top) with our method (bottom). To highlight the importance of having multi-dimensional disentangled subspaces, we trained the network by aggregating the translation factors along the *x* and *y* dimensions into a single one, making it multi-dimensional by construction. Even if reaching a similar MIG score (Ada-GVAE: 50.2%, Ours: 49.3%), the interpolation obtained with our method is better behaved.

Table 5. Comparison of disentanglement metrics and reconstruction score at varying latent space dimensions on the FAUST dataset. We trained our model by fixing the number of latent subspaces to 2, while allowing different dimensions of the global latent space to 2, 4, 8, 16.

	MIG			
	2	4	8	16
Ada-GVAE	15.1%	**11.1%**	**4.6%**	2.3%
Ours	**18.2%**	4.39%	3.08%	**2.68%**
	MIG-KM			
	2	4	8	16
Ours	-	10.6%	10.3%	9.5%
	Reconstruction error			
	2	4	8	16
Ada-GVAE	7.22e-3	5.11e-3	3.39e-3	3.30e-3
Ours	**2.56e-3**	**2.04e-3**	**1.24e-3**	**1.26e-3**

Table 6. Performance comparison of our method with and without fixing the number of changing factors to exactly one. We report median results on the *Shapes3D* dataset.

#factors	Beta VAE	DCI Dis.	MIG	MIG-PCA	MIG-KM
One	100%	99.0%	63.7%	73.5%	69.2%
Variable	98.9%	94.9%	62.3%	70.5%	66.9%

is then assigned to one of these b centroids, thus deriving a probability distribution over the centroids. The Mutual Information is then computed between this distribution and the distribution of the ground-truth labels in order to derive the MIG score. This latter metric is particularly important in the real data scenario, where each factor of variation is likely to be multi-dimensional, while MIG-PCA still assumes the factor of variation to be represented in a linear subspace.

Implementation details. We performed our experimental evaluation on a machine equipped with an NVIDIA RTX 2080ti, within the Pytorch framework. The architecture for encoder and decoder on images are convolutional, with the same exact settings as (Locatello et al., 2020a). For the experiments on FAUST we used a PointNet (Qi et al., 2017) architecture and a simple MLP as a decoder. Detailed information on the architecture, including the hyperparameter choice, are reported in the Appendix.

4.1. Analysis

In the following we analyze the core properties of our approach.

Inferring the number of latent subspaces. We stress that our method is tailored for learning the number of subspaces, and therefore the number of generative factors, as well as the dimensionality of each of them. The only hyperparameter that needs to be set is the dimensionality of the latent space $\mathcal{Z}$ to be factorized. To infer the number of subspaces we over-estimate this parameter, with the desirable consequence that the network will activate only some of them, corresponding to the number of ground truth generative factors. The remaining ones will collapse to a point, as we can observe in the qualitative example of Figure 5 accompanied by the quantitative result in Table 7.

Expressiveness of higher dimensional latent subspaces. We further report experimental evidence that having multi-dimensional subspaces allows to obtain a better disentangled representation. We show this on the challenging FAUST dataset (Bogo et al., 2014), where the factors of variation are not reducible to just one dimension. As a test for our method, we confront ourselves with the challenging task of separating pose and style for human 3D models in different poses and identities.

In Table 5 we ran this experiment while growing progressively the dimensionality of the latent space $\mathcal{Z}$ from 2 to 16. Crucially, we show that adopting progressively larger multi-dimensional subspaces does not degrade the disentanglement quality as measured by the MIG-KM score, while the reconstruction error goes down faster than with one dimension. The standard MIG score shows that in the one-dimensional case the performance of disentanglement degrades. A qualitative example is reported in Figure 6. A qualitative comparison with (Locatello et al., 2020a) on a different dataset is also shown in Figure 4.

We remark here that the constraint of observing data sampled in pairs is a reasonable assumption, because data come often in this form in the acquisition stage; examples include

Table 7. Average standard deviation on the latent subspaces corresponding to Fig. 5. The lower scores correspond to collapsed subspaces.

Subspace	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	$\mathcal{S}_6$	$\mathcal{S}_7$	$\mathcal{S}_8$	$\mathcal{S}_9$	$\mathcal{S}_{10}$
Average std	0.061	0.043	0.063	0.026	2.6e-5	2.1e-5	1.5e-5	0.058	0.063	2.1e-5
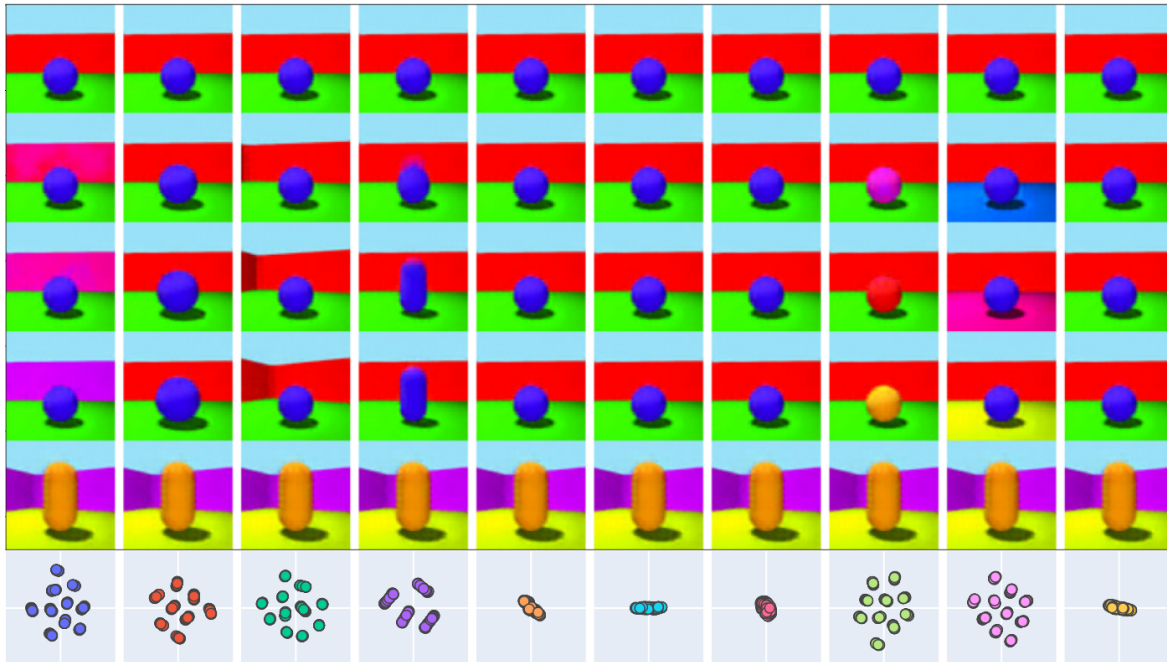
Figure 5. Analysis of the latent subspaces at training convergence on the *Shapes3D* dataset. We project each latent subspace (ordered by column number) on a 2-dimensional space, using t-SNE (last row). On the rows we show the interpolation (rows 2,3) between two samples which differ on all factors of variation (rows 1,4). We show that the latent subspaces (columns 1,2,3,4,8,9) which encode information about the generative factors (wall hue, size, camera angle, shape, object hue, floor hue) are correctly characterized by the latent subspace projections, and the number of clusters is almost always equal to the number of possible values that the generative factor can assume. For example, the number of clusters in the first subspace corresponds to the ten possible color hue values for the wall, and the number of clusters in the 4th subspace corresponds to the number of possible shapes in the dataset. The latent spaces which do not encode any information (columns 5,6,7,10) collapse to one point (represented by a single, small cluster in the projection), as confirmed by the quantitative evaluation in Table 7.

temporal consistency of video frames (Lai et al., 2018), object identity consistency of a rendered object from different camera angles, etc.

Robustness to simultaneous changes of multiple factors. Although the proposed framework assumes that the observed pairs change in only one factor of variation, we show that, in practice, our method performs well also in settings where more than one factor may randomly vary simultaneously. The only assumption we make is that at least one factor is shared among the elements of the sampled pairs. To build the pairs of observations we follow the same process as before, where we sample a random number between 1 and $k - 1$ transformed factors between the elements of the sampled pair. Our method proves to be robust to this challenging setting, with performance comparable to the case with one fixed factor of variation, as we show in Table 6.

Ablation study. We justify the importance of the consistency loss and the distribution loss by running some ablation experiments. We fix a random seed and run experiments on the *Shapes3D* dataset with a complete run, a run without the regularization loss, a run without the consistency loss, and a run without both. The importance of using all our loss terms is highlighted in Table 8, where we report the β-VAE, Factor-VAE, DCI and MIG scores for these setups. This is further supported by the rank correlation matrix between losses and disentanglement scores shown in Figure 7.

5. Conclusions

In this paper, we introduced a new definition of disentanglement based on the notion of metric space, which generalizes existing approaches. Relying on this definition, we proposed
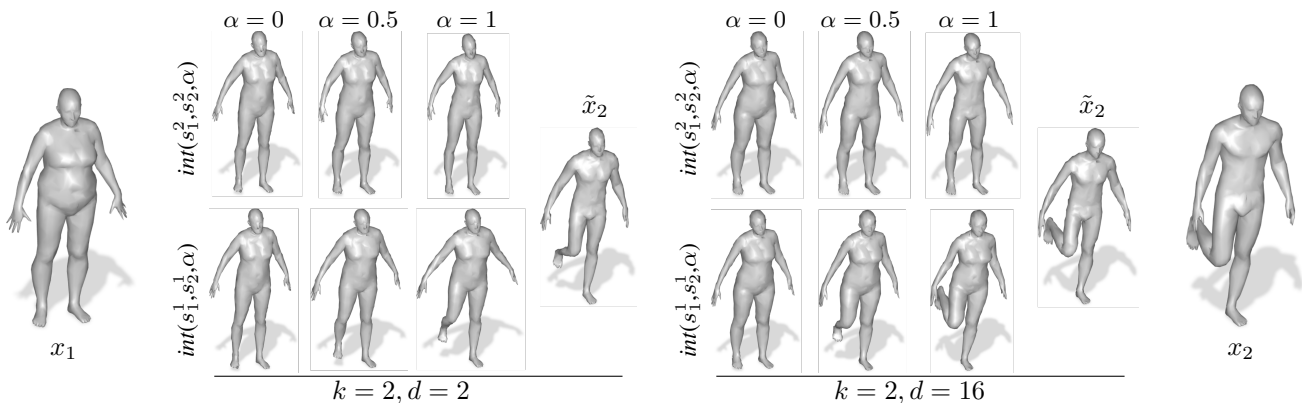
Figure 6. Disentangled interpolation on a pair of FAUST shapes with different dimensions of the latent space. For each of the two latent space dimensions ($d = 2$ on the left and $d = 16$ on the right) we encode the two input shapes x_1 and x_2 as subspace latent vectors (s_1^1, s_1^2) for x_1 and (s_2^1, s_2^2) for x_2. We then interpolate separately the 2 latent subspaces between the two shapes, and obtain the decoded shapes $g(aggr(s_1^1, int(s_1^2, s_2^2, \alpha)))$ (top row) and $g(aggr(int(s_1^1, s_2^1, \alpha), s_1^2))$ (bottom row), with $int(a, b, \alpha) = (1 - \alpha)a + \alpha b$. We can see how, even if both configurations are able to achieve the disentanglement between pose and style of the subjects, the smaller latent space ($d = 2$) is not able to encode the finer shape details, resulting in smoother and less accurate reconstructions.

Loss terms	Beta VAE	Factor VAE	DCI Dis.	MIG	MIG-PCA	MIG-KM
Complete	**100%**	**100%f**	96.1%	**81.0%**	**86.0%**	**80.3%**
W/o $\mathcal{L}_{reg}$	99.5%	99.8%	77.2%	48.2%	58.6%	51.5%
W/o $\mathcal{L}_{cons}$	**100%**	97.1%	**97.0%**	65.4%	81.0%	71.0%
W/o $\mathcal{L}_{reg}, \mathcal{L}_{cons}$	93.4%	87.9%	78.2%	56.2%	55.5%	53.8%

Rank correlation

Table 8 & Figure 7. Analysis of the importance of the adopted loss terms on the *Shapes3D* dataset. The table (**left**) reports ablation results; these show the disentanglement scores obtained by training our model with a fixed initialization seed, and removing some of the loss terms. We can see how all the losses contribute to achieve a better disentanglement. A similar conclusion can be drawn looking at the Rank correlation matrix (**right**), showing the rank correlation between the losses and the disentanglement scores obtained after training 10 models. All the disentanglement losses (first 4 columns) are negatively correlated with disentanglement scores. The disentanglement loss $\mathcal{L}_{dis}$ seems to be less correlated with respect to the others, which is due to the fact that it always reaches a similar low value in all the runs.

a simple recipe to compute disentangled representations in practice, in a weakly supervised setting. In contrast with previous work, we demonstrated that disentanglement representations benefit from the choice of modelling generative factors as possibly multi-dimensional subspaces, especially when the true factors of variation live in a space with dimension greater than 1. In many cases, the proposed solution outperforms state-of-the-art competitors on synthetic datasets and more challenging real-world scenarios. Moreover, our method can be easily adapted to many standard neural network models as a plugin module.

Future work. Exploiting the possible contribution of the proposed model in existing autoencoders is a natural direction raised by this paper. In the future, we aim to investigate possible alternative constraints to incorporate additional properties in the latent space or its factorization.

References

Belkin, M. and Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pp. 585–591. MIT Press, 2001.

Bengio, Y., Courville, A., and Vincent, P. Representa-

tion learning: A review and new perspectives. *arXiv 1206.5538*, 2014.

Bogo, F., Romero, J., Loper, M., and Black, M. J. FAUST: dataset and evaluation for 3d mesh registration. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3794–3801. IEEE Computer Society, 2014.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2615–2625, 2018.

Comon, P. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, 1994. Higher Order Statistics.

Cosmo, L., Norelli, A., Halimi, O., Kimmel, R., and Rodolà, E. Limp: Learning latent shape representations with metric preservation priors. *Lecture Notes in Computer Science*, pp. 19–35, 2020. ISSN 1611-3349.

Dubrovina, A., Xia, F., Achlioptas, P., Shalah, M., Groscot, R., and Guibas, L. J. Composite shape modeling via latent space factorization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8139–8148. IEEE, 2019.

Eastwood, C. and Williams, C. K. I. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

F.R.S., K. P. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Hadsell, R., Chopra, S., and Lecun, Y. Dimensionality reduction by learning an invariant mapping. pp. 1735 – 1742, 2006. ISBN 0-7695-2597-0.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D. J., and Lerchner, A. Towards a definition of disentangled representations. *ArXiv*, abs/1812.02230, 2018.

Hu, Q., Szabó, A., Portenier, T., Favaro, P., and Zwicker, M. Disentangling factors of variation by mixing them. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3399–3407. IEEE Computer Society, 2018.

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3765–3773, 2016.

Hyvärinen, A., Sasaki, H., and Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429 – 439, 1999. ISSN 0893-6080.

Kazemi, H., Iranmanesh, S. M., and Nasrabadi, N. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 848–856, 2019.

Kim, H. and Mnih, A. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2654–2663. PMLR, 2018.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. B. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2539–2547, 2015.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Lai, W., Huang, J., Wang, O., Shechtman, E., Yumer, E., and Yang, M. Learning blind video temporal consistency. *CoRR*, abs/1808.00449, 2018.

LeCun, Y., Fu Jie Huang, and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pp. II–104 Vol.2, 2004.

Lee, J. *Introduction to manifolds*. Springer, 2012.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 2020a.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variation using few labels. *arXiv 1905.01258*, 2020b.

Locke, J. *An Essay Concerning Human Understanding*. Oxford University Press, 1689.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16, 2015.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, art. arXiv:1802.03426, 2018.

Nash, J. The imbedding problem for riemannian manifolds. *Annals of Mathematics*, 63(1):20–63, 1956. ISSN 0003486X.

Pfau, D., Higgins, I., Botev, A., and Racanière, S. Disentangling by subspace diffusion. *arXiv 2006.12982*, 2020.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 77–85. IEEE Computer Society, 2017.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1252–1260, 2015.

Roweis, S. T. and Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290 (5500):2323–2326, 2000.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Shukla, A., Bhagat, S., Uppal, S., Anand, S., and Turaga, P. Product of orthogonal spheres parameterization for disentangled representation learning, 2019.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. ISSN 0036-8075.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Whitney, H. Differentiable manifolds. *Annals of Mathematics*, 37(3):645–680, 1936. ISSN 0003486X.

Zhan, G., Zhao, Y., Zhao, B., Yuan, H., Chen, B., and Dong, H. DLGAN: disentangling label-specific fine-grained features for image manipulation. *CoRR*, abs/1911.09943, 2019.