Pattern Recognition 134 (2023) 109130

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Unsupervised learning of global factors in deep generative models

Ignacio Peis*, Pablo M. Olmos, Antonio Artés-Rodríguez

Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain

ARTICLE INFO

Article history:
Received 12 July 2021
Revised 14 September 2022
Accepted 22 October 2022
Available online 25 October 2022

Keywords:
VAE
Deep generative models
Global factors
Unsupervised learning
Disentanglement
Representation learning

ABSTRACT

We present a novel deep generative model based on non i.i.d. variational autoencoders that captures global dependencies among observations in a fully unsupervised fashion. In contrast to the recent semi-supervised alternatives for global modeling in deep generative models, our approach combines a mixture model in the local or data-dependent space and a global Gaussian latent variable, which lead us to obtain three particular insights. First, the induced latent global space captures interpretable disentangled representations with no user-defined regularization in the evidence lower bound (as in β-VAE and its generalizations). Second, we show that the model performs domain alignment to find correlations and interpolate between different databases. Finally, we study the ability of the global space to discriminate between groups of observations with non-trivial underlying structures, such as face images with shared attributes or defined sequences of digits images.

1. Introduction

Since its first proposal by [21], Variational Autoencoders (VAEs) have evolved into a vast amount of variants. To name some representative examples, we can include VAEs with latent mixture models priors [9], adapted to model time-series [4,8,11], trained via deep hierarchical variational families [32,35], with enhanced, parametric and robust priors [20,35,36], that include advanced techniques for gradient estimation [7,33] or that naturally handle heterogeneous data types and missing data [26,27,29,31].

The large majority of VAE-like models are designed over the assumption that data is i.i.d., which remains a valid strategy for simplifying the learning and inference processes in generative models with latent variables. A different modelling approach may drop the i.i.d. assumption with the goal of capturing a higher level of dependence between samples. Inferring such kind of higher level dependencies can directly improve current approaches to find interpretable disentangled generative models [5], to perform domain alignment [15] or to ensure fairness and unbiased data [3].

The main contribution of this paper is to show that a deep probabilistic VAE non i.i.d. model with both local and global latent variable can capture meaningful and interpretable correlation among data points in a completely unsupervised fashion. Namely, weak supervision to group the data samples is not required. In the following we refer to our model as Unsupervised Global VAE (UG-VAE). We combine a clustering inducing mixture model prior in the local space, that helps to separate the fundamental data features that an i.i.d. VAE would separate, with a global latent variable that modulates the properties of such latent clusters depending on the observed samples, capturing fundamental and interpretable data features. We demonstrate such a result using both CelebA, MNIST and the 3D FACES dataset in [30]. Furthermore, we show that the global latent space can explain common features in samples coming from two different databases without requiring any domain label for each sample, establishing a probabilistic unsupervised framework for domain alignment. Up to our knowledge, UG-VAE is the first VAE model in the literature that performs unsupervised domain alignment using global latent variables.

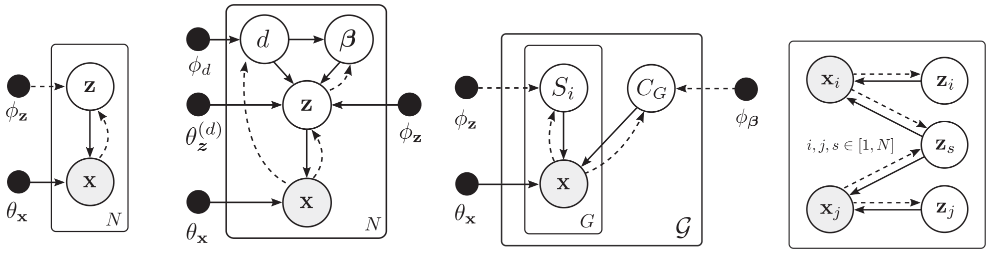
Finally, we demonstrate that, even when the model parameters have been trained using an unsupervised approach, the global latent space in UG-VAE can discriminate groups of samples with non-trivial structures, separating groups of people with black and blond hair in CelebA or series of numbers in MNIST. In other words, if weak supervision is applied at test time, the posterior distribution of the global latent variable provides with an informative representation of the user defined groups of correlated data.

The principal contributions of this work are:

- To our knowledge, we propose the first deep generative model for generating groups of samples with shared properties learned in a fully-unsupervised fashion, named UG-VAE.
- We demonstrate that the information captured in the structured latent space of UG-VAE is highly interpretable in compar-

* Corresponding author.:.
E-mail address: ipeis@tsc.uc3m.es (I. Peis).

https://doi.org/10.1016/j.patcog.2022.109130



(a) VAE (b) GMVAE (c) ML-VAE (d) NestedVAE

Fig. 1. Comparison of four deep generative models. Dashed lines represent the graphical model of the associated variational family. The Vanilla VAE (a), the GMVAE (b), and semi-supervised variants for grouped data; ML-VAE (c) and NestedVAE (d).

ison with other related methods, leading to an improved disentanglement in both local and global spaces.

- We demonstrate that, by simply training UG-VAE with mini-batches of samples from several datasets, the structured latent space aligns them and captures common interpretable properties without any label or supervision.
- We demonstrate that, although the training unsupervised, the global space is able to effectively separate the global posterior of different groups when weak supervision is included at test time for grouping observations with a given label or attribute.

2. Related work

Non i.i.d. deep generative models are getting recent attention but the literature is still scarse. First we find VAE models that implement non-parametric priors: in [14] the authors make use of a global latent variable that induces a non-parametric Beta process prior, and more efficient variational mechanism for this kind of IBP prior are introduced in [38]. Second, both [22] and [34] proposed non i.i.d. exchangable models by including correlation information between datapoints via an undirected graph. Third, conditional dependencies with supervised classes are modeled in [1] with the aim at performing natural clustering at the latent space and disentangle class-dependent factors. Finally, some other works rely on simpler generative models (compared to these previous approaches), including global variables with fixed-complexity priors, typically a multi-variate Gaussian distribution, that aim at modelling the correlation between user-specified groups of correlated samples (e.g. images of the same class in MNIST, or faces of the same person). In [5] or [18], authors apply weak supervision by grouping image samples by identity, and include in the probabilistic model a global latent variable for each of these groups, along with a local latent variable that models the distribution for each individual sample. In [24], authors use co-supervision for achieving stationary state in learning graphs for multi-view clustering. Below we specify the two most relevant lines of research, in relation to our work.

VAEs with mixture priors Several previous works have demonstrated that incorporating a mixture in the latent space leads to learn significantly better models. In [19] authors introduce a latent GMM prior with nonlinear observations, where the means are learned and remain invariant with the data. The GMVAE proposal by Dilokthanakul et al. [9] aims at incorporating unsupervised clustering in deep generative models for increasing interpretability. In the VAMP VAE model [35], the authors define the prior as a mixture with components given by approximated variational posteriors, that are conditioned on learnable pseudo-inputs. This approach leads to an improved performance, avoiding typical local optima difficulties that might be related to irrelevant latent dimensions.

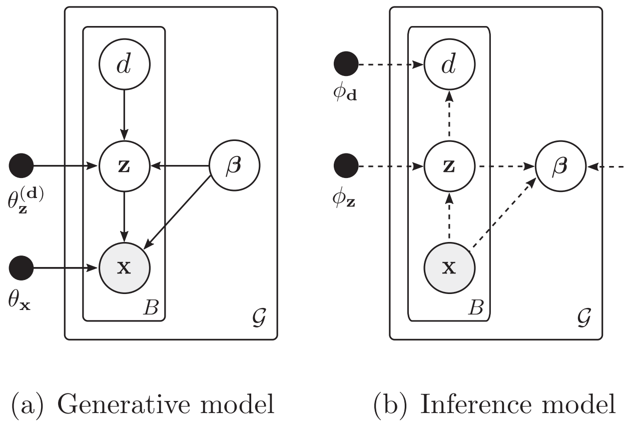
Semi-supervised deep models for grouped data In contrast to the i.i.d. vanilla VAE model in Fig. 1(a), and its augmented version for unsupervised clustering, GMVAE, in Fig. 1(b), the graphical model of the Multi-Level Variational Autoencoder (ML-VAE) in [5] is shown in Fig. 1(c), where G denotes the number of groups. ML-VAE includes a local Gaussian variable S_i that encodes style-related information for each sample, and a global Gaussian variable C_G is shared within a group of samples. For instance, they feed their algorithm with batches of face images from the same person, modeling content shared within the group that characterize a person. This approach leads to learning disentangled representations at the group and observations level, in a content-style fashion. Nevertheless, the groups are user-specified, hence resulting in a semi-supervised modelling approach. In [37] authors use weak supervision for pairing samples. They implement two outer VAEs with shared weights for the reconstruction, and a Nested VAE that reconstructs latent representation off one to another, modelling correlations across pairs of samples. The graphical model for Nested VAE is depicted in Fig. 1(d). Despite the fact that semi-supervision is proved to improve performance for some deep generative models [12,13], it requires prior knowledge about the data that we do not assume in this work.

3. Unsupervised global VAE

We present UG-VAE, a deep generative VAE framework for modeling non-i.i.d. data with global dependencies. It generalizes the ML-VAE graphical model in Fig. 1(c), combining the global model with a mixture prior to *(i)* remove the group supervision, *(ii)* include a clustering-inducing prior in the local space, and *(iii)* propose a more structured variational family. The latent discrete variable d is expected to represent the inferred group with no supervision needed.

3.1. Generative model

Fig. 2 represents the generative graphical model of UG-VAE. The global variable $\boldsymbol{\beta} \in \mathbb{R}^g$ modulates the prior, inducing shared properties within a group of B samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_B\} \subseteq \mathbb{R}^D$, while the local variable $\mathbf{z}$ encodes the local properties for each datapoint. Although we use this notation for the global latent space, we would like to remark that $\boldsymbol{\beta}$ is not a parameter as the β defined in [17]. We denote by $\mathcal{G}$ the number of groups we jointly use to amortize the learning of the model parameters. During amortized variational training, groups are simply random data mini-batches from the training dataset, being $\mathcal{G}$ the number of data mini-batches. We could certainly take $B = N$ (the training set size) and hence $\mathcal{G} = 1$, but this leads to a less interpretable global latent space (too much data to correlate with a single global realization), and a slow training process. On the contrary, a small batch size might result in



(a) Generative model (b) Inference model

Fig. 2. Generative (left) and inference (right) of UG-VAE.

highly dispersed global properties, difficult to capture and again hardly interpretable. The difficulty in choosing a proper value for the batch size limits the potential of learning useful representations, and arises in the lack of agnostic metrics for performing objective validations. Although some useful representation metrics (e.g. [16]) could be used for validating B, we show results in Section 4 demonstrating that reasonable batch sizes values that are widely employed in similar works (namely $B = 128$) successfully learn disentangled global representations.

Conditioned to $\boldsymbol{\beta}$, data samples are distributed according to a Gaussian mixture local (one per data) latent variable $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_B\} \subseteq \mathbb{R}^d$, and $\mathbf{d} = \{d_1, \dots, d_B\} \subseteq \{1, \dots, K\}$ are independent discrete categorical variables with uniform prior distributions. This prior, along with the conditional distribution $p(\mathbf{z}_i|d_i, \boldsymbol{\beta})$, defines a Gaussian mixture latent space, which helps to infer similarities between samples from different batches (by assigning them to the same cluster), and thus, d_i plays a similar role than the semi-supervision included in [5] by grouping. Our experimental results demonstrate that this level of structure in the local space is crucial to acquire interpretable information at the global space.

The joint distribution for a single group is therefore defined by:

$$p_\theta(\mathbf{X}, \mathbf{Z}, \mathbf{d}, \boldsymbol{\beta}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\beta})\, p(\mathbf{Z}|\mathbf{d}, \boldsymbol{\beta})\, p(\mathbf{d})\, p(\boldsymbol{\beta}) \tag{1}$$

where the likelihood term of each sample is a Gaussian distribution, whose parameters are obtained from a concatenation of $\mathbf{z}_i$ and $\boldsymbol{\beta}$ as input of a decoder network:

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\beta}) = \prod_{i=1}^{B} p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\beta}) = \prod_{i=1}^{B} \mathcal{N}\left(\boldsymbol{\mu}_{\theta_x}([\mathbf{z}_i, \boldsymbol{\beta}]), \boldsymbol{\Sigma}_{\theta_x}([\mathbf{z}_i, \boldsymbol{\beta}])\right) \tag{2}$$

In contrast with [19], where the parameters of the clusters are learned but shared by all the observations, in UG-VAE, the parameters of each component are obtained with networks fed with $\boldsymbol{\beta}$. Thus, the prior of each local latent continuous variable is defined by a mixture of Gaussians, where d_i defines the component and $\boldsymbol{\beta}$ is the input of a NN that outputs its parameters:

$$p(\mathbf{Z}|\mathbf{d}, \boldsymbol{\beta}) = \prod_{i=1}^{B} p(\mathbf{z}_i|d_i, \boldsymbol{\beta}) = \prod_{i=1}^{B} \mathcal{N}\left(\boldsymbol{\mu}_{\theta_z}^{(d_i)}(\boldsymbol{\beta}), \boldsymbol{\Sigma}_{\theta_z}^{(d_i)}(\boldsymbol{\beta})\right) \tag{3}$$

hence we trained as many NNs as discrete categories. This local space encodes samples in representative clusters to model local factors of variation. The prior of the discrete latent variable is defined as uniform:

$$p(\mathbf{d}) = \prod_{i=1}^{B} \mathrm{Cat}(\boldsymbol{\pi}) \quad \pi_k = 1/K \tag{4}$$

and the prior over the continuous latent variable β follows an isotropic Gaussian, $p(\boldsymbol{\beta}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

By making use of the presented generative model, we propose a flexible hierarchy of both global and local mixture of components that, while being trained on random-mini batches, it is able to exploit the augmented degrees of freedom for capturing group-features in the latent space in an unsupervised manner. A graphical representation of how UG-VAE structures the information in the latent space is provided in Fig. 3. Further, as shown in Section 3.2, the posterior approximation results from an individual contribution of each data point that favors group separation across latent spaces.

3.2. Inference model

The graphical model of the proposed variational family is shown in Fig. 2(b):

$$q_\phi(\mathbf{Z}, \mathbf{d}, \boldsymbol{\beta}|\mathbf{X}) = q(\mathbf{Z}|\mathbf{X})\, q(\mathbf{d}|\mathbf{Z}) q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}) \tag{5}$$

where we employ an encoder network that maps the input data into the local latent posterior distribution, which is defined as a Gaussian:

$$q(\mathbf{Z}|\mathbf{X}) = \prod_{i=1}^{B} q(\mathbf{z}_i|\mathbf{x}_i) = \prod_{i=1}^{B} \mathcal{N}(\boldsymbol{\mu}_{\phi_z}(\mathbf{x}_i), \boldsymbol{\Sigma}_{\phi_z}(\mathbf{x}_i)) \tag{6}$$

Given the posterior distribution of $\mathbf{z}$, the categorical posterior distribution of d_i is parametrized by a NN that takes $\mathbf{z}_i$ as input

$$q(\mathbf{d}|\mathbf{Z}) = \prod_{i=1}^{B} q(d_i|\mathbf{z}_i) = \prod_{i=1}^{B} \mathrm{Cat}(\pi_{\phi_d}(\mathbf{z}_i)) \tag{7}$$

The approximate posterior distribution of the global variable $\boldsymbol{\beta}$ is computed as a product of local contributions per datapoint within a randomly sampled batch. This strategy, as demonstrated by Bouchacourt et al. [5], outperforms other approaches like, for example, a mixture of local contributions, as it allows to accumulate group evidence. For each sample, a NN encodes $\mathbf{x}_i$ and the Categorical parameters $\pi_{\phi_d}(\mathbf{z}_i)$ in a local Gaussian:

$$\begin{aligned} q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}) &= \mathcal{N}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right) \\ &= \prod_{i=1}^{B} \mathcal{N}\left(\boldsymbol{\mu}_{\phi_\beta}([\mathbf{x}_i, \pi_{\phi_d}(\mathbf{z}_i)]), \boldsymbol{\Sigma}_{\phi_\beta}([\mathbf{x}_i, \pi_{\phi_d}(\mathbf{z}_i)])\right) \end{aligned} \tag{8}$$

If we denote by $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ the parameters obtained by networks $\boldsymbol{\mu}_{\phi_\beta}$ and $\boldsymbol{\Sigma}_{\phi_\beta}$, respectively, the parameters of the global Gaussian distribution are given, following [6], by:

$$\begin{aligned} \boldsymbol{\Lambda}_\beta &= \boldsymbol{\Sigma}_\beta^{-1} = \sum_{i=1}^{B} \boldsymbol{\Lambda}_i \\ \boldsymbol{\mu}_\beta &= (\boldsymbol{\Lambda}_\beta)^{-1} \sum_{i=1}^{B} \boldsymbol{\Lambda}_i \boldsymbol{\mu}_i \end{aligned} \tag{9}$$

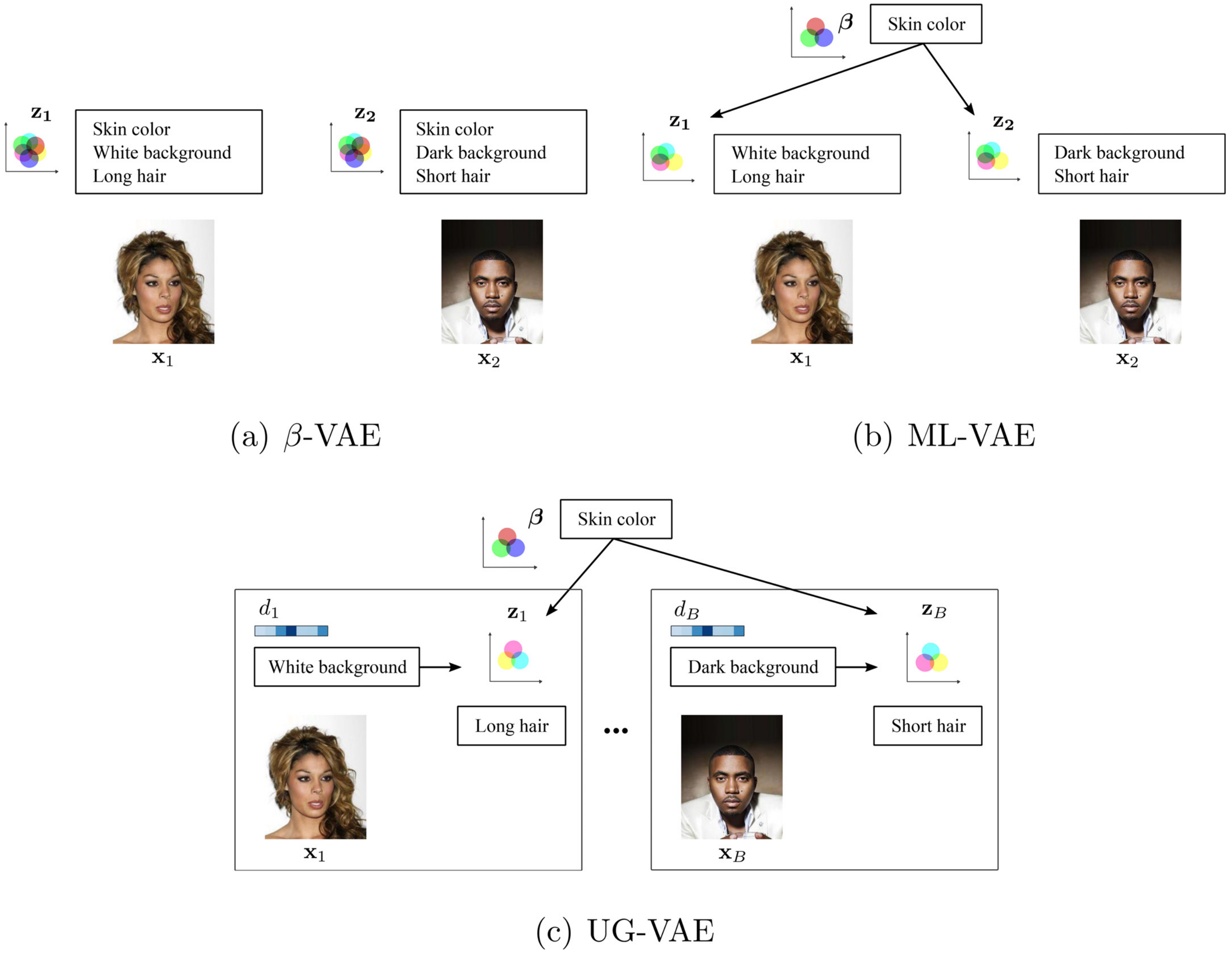
where $\boldsymbol{\Lambda}_\beta = \boldsymbol{\Sigma}_\beta^{-1}$ is defined as the precision matrix, which we model as a diagonal matrix.

3.3. Evidence lower bound

Overall, the evidence lower bound reads as follows:

$$\begin{aligned} \mathcal{L}(\theta, \phi;\, \mathbf{X}, \mathbf{Z}, \mathbf{d}, \boldsymbol{\beta}) = {} & \mathbb{E}_{q(\boldsymbol{\beta})}\left[\mathcal{L}_i(\theta, \phi;\, \mathbf{x}_i, \mathbf{z}_i, \mathbf{d}, \boldsymbol{\beta})\right] \\ & - \mathbb{E}_{q(\mathbf{d})}\left[D_{KL}\left(q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}) \| p(\boldsymbol{\beta})\right)\right] \end{aligned} \tag{10}$$

The resulting ELBO is an expansion of the ELBO for a standard GM-VAE with a new regularizer for the global variable. As the reader



(a) β-VAE

(b) ML-VAE

(c) UG-VAE

Fig. 3. Illustration of the inductive bias introduced by the generative model in β-VAE (a), ML-VAE (b) and the proposed UG-VAE (c). β-VAE compresses all features in a single latent space, while ML-VAE uses a global component that needs to be supervised during training. In UG-VAE, we propose a flexible hierarchy of both global and local mixture of components that enables learning group-features in the latent space in an unsupervised manner.

may appreciate, the ELBO for UG-VAE does not include extra hyperparameters to enforce disentanglement, like other previous works as β-VAE, and thus, no extra validation is needed apart from the parameters of the networks architecture, the number of clusters and the latent dimensions. We denote by $\mathcal{L}_i$ each local contribution to the ELBO:

$$\mathcal{L}_i(\theta,\phi;\,\mathbf{x}_i,\mathbf{z}_i,\mathbf{d},\boldsymbol{\beta}) = \mathbb{E}_{q(\boldsymbol{d}_i,\mathbf{z}_i)}\big[\log p(\mathbf{x}_i|\mathbf{z}_i,d_i,\boldsymbol{\beta})\big] -\mathbb{E}_{q(\boldsymbol{d}_i)}\Big[D_{KL}\big(q(\mathbf{z}_i|\mathbf{x}_i)\|p(\mathbf{z}_i|d_i,\boldsymbol{\beta})\big)\Big] - D_{KL}(q(d_i|\mathbf{z}_i)\|p(d_i))) \tag{11}$$

The first part of (10) is an expectation over the global approximate posterior of the so-called local ELBO. This local ELBO differs from the vanilla ELBO proposed by Kingma and Welling [21] in the regularizer for the discrete variable d_i, which is composed by the typical reconstruction term of each sample and two KL regularizers: one for $\mathbf{z}_i$, expected over d_i, and the other over d_i. The second part in (10) is a regularizer on the global posterior. The expectations over the discrete variable d_i are tractable and thus, analytically marginalized.

In contrast with GMVAE (Fig. 1(b)), in UG-VAE, $\boldsymbol{\beta}$ is shared by a group of observations, therefore the parameters of the mixture are the same for all the samples in a batch. In this manner, within each optimization step, the encoder $q(\boldsymbol{\beta}|\mathbf{X},\mathbf{Z})$ only learns from the global information obtained from the product of Gaussian contributions of every observation, with the aim at configuring the mixture to improve the representation of each datapoint in the batch, by means of $p(\mathbf{Z}|\mathbf{d},\boldsymbol{\beta})$ and $p(\mathbf{X}|\mathbf{Z},\boldsymbol{\beta})$. Hence, the control of the mixture is performed by using global information. In contrast with ML-VAE (whose encoder $q(C_G|\mathbf{X})$ is also global, but the model does not include a mixture), in UG-VAE, the $\boldsymbol{\beta}$ encoder incorporates information about which component each observation belongs to, as the weights of the mixture inferred by $q(\mathbf{d}|\mathbf{Z})$ are used to obtain $q(\boldsymbol{\beta}|\mathbf{X},\mathbf{Z})$. Thus, while each cluster will represent different local features, moving $\boldsymbol{\beta}$ will affect all the clusters. In other words, modifying $\boldsymbol{\beta}$ will have some effect in each local cluster. As the training progresses, the encoder $q(\boldsymbol{\beta}|\mathbf{X},\mathbf{Z})$ learns which information emerging from each batch of data allows to move the cluster in a way that the ELBO increases.

4. Experiments

In this section we demonstrate the ability of the UG-VAE model to infer global factors of variation that are common among samples, even when coming from different datasets. In all cases, we have not validated in depth all the networks used, we have merely rely on encoder/decoder networks proposed in state-of-the-art VAE

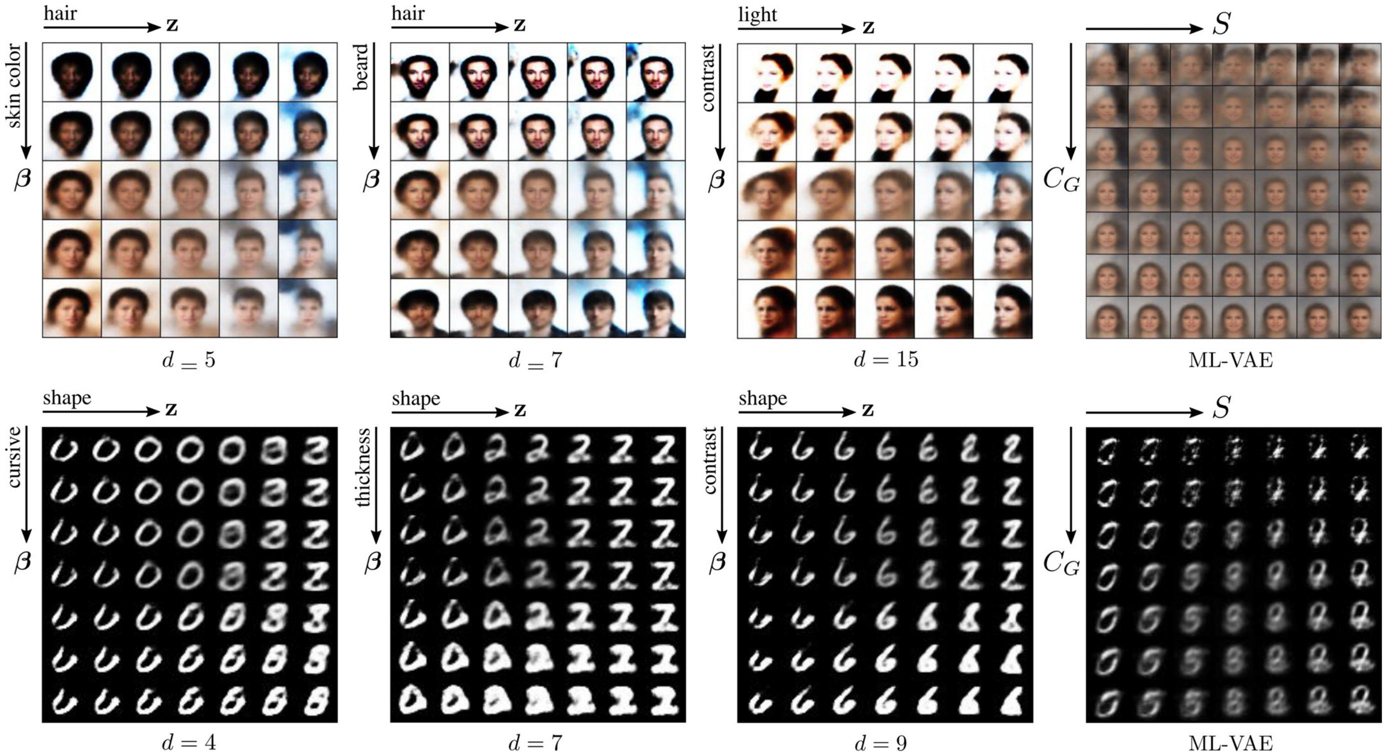


Fig. 4. Sampling from UG-VAE (first three columns) and ML-VAE (last column) for CelebA (top) and MNIST (bottom). We include samples from 3 local clusters of UG-VAE from a total of $K = 20$ for CelebA and $K = 10$ for MNIST. In CelebA (top), the global latent variable disentangles in skin color, beard and face contrast, while the local latent variable controls hair and light orientation. In MNIST (bottom), $\boldsymbol{\beta}$ controls cursive grade, contrast and thickness of handwriting, while $\mathbf{z}$ varies digit shape. In ML-VAE (right column), both spaces are unimodal and the disentanglement is hardly interpretable when we feed the data without semi-supervision.

papers such as [5,21] or [17]. Our results must be hence regarded as a proof of concept about the flexibility and representation power of UG-VAE, rather than fine-tuned results for each case. Hence there is room for improvement in all cases. Details about network architecture and training parameters are provided in Appendix B.

4.1. Unsupervised learning of global factors

In this section we first asses the interpretability of the global disentanglement features inferred by UG-VAE over both CelebA and MNIST. In Fig. 4 we show samples of the generative model as we explore both the global and local latent spaces. We perform a linear interpolation with the aim at exploring the hypersphere centered at the mean of the distribution and with radius σ_i for each dimension i. Instead of finding influential latent factors [25] and interpolate them (fixing the rest), we choose to maximize the variation range across every dimension, moving diagonally through the latent space. Rows correspond to an interpolation on the global $\boldsymbol{\beta}$ between $[-1, 1]$ on every dimension ($p(\boldsymbol{\beta})$ follows a standard Gaussian). As the local $p(\mathbf{z}|d, \boldsymbol{\beta})$ ((3)) depends on d and $\boldsymbol{\beta}$, if we denote $\boldsymbol{\mu}_z = \boldsymbol{\mu}_z^{(d)}(\boldsymbol{\beta})$, the local interpolation goes from $[\mu_{z0} - 3, \mu_{z1} - 3, \dots \mu_{zd} - 3]$ to $[\mu_{z0} + 3, \mu_{z1} + 3, \dots, \mu_{zd} + 3]$. The range of ±3 for the local interpolation is determined to cover the variances $\boldsymbol{\Sigma}_z^{(d)}(\boldsymbol{\beta})$ that we observe upon training the model for MNIST and CelebA. The every image in Fig. 4 correspond to samples from a different cluster (fixed values of d), in order to facilitate the interpretability of the information captured at both local and global levels. By using this set up, we demonstrate that the global information tuned by $\boldsymbol{\beta}$ is different and clearly interpretable inside each cluster. In order to visually remark the advantage of capturing global correlations among samples with UG-VAE, we include in Fig. 5 an interpolation in the latent space of β-VAE. We explore from $\mathbf{z} = [-1, -1, \dots, -1]$ to $\mathbf{z} = [1, 1, \dots, 1]$, given that the prior is an isotropic Gaussian. As the reader may appreciate, only one row is included as β-VAE does not include global space. In this case, moving diagonally through the latent space start from a blond woman and ends in a brunette woman with the same angle face. Thus, the local space is in charge of encoding both content and style aspects. Although in β-VAE, authors analyze the disentanglement in each dimension of the latent space, we do not study whether each dimension of $\mathbf{z}$ represents an interpretable generative factor in UG-VAE or not, as it is out of the scope for this work. The novelty lies on the fact that, apart from the local disentanglement, our model adds an extra point of interpretability through the disentanglement in the global space.

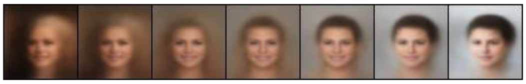


Fig. 5. Interpolation in the prior latent space of β-VAE with $\beta = 10$, using the same networks architecture than in the local part of UG-VAE. Interpolation consists on 7 steps from $\mathbf{z} = [-1, -1, \dots, -1]$ to $\mathbf{z} = [1, 1, \dots, 1]$.

The total number of clusters is set to $K = 20$ for CelebA and $K = 10$ for MNIST. Three of these components are presented in Fig. 4. We can observe that each row (each value of $\boldsymbol{\beta}$) induces a shared generative factor, while $\mathbf{z}$ is in charge of variations inside this common feature. For instance, in CelebA (top), features like skin color, presence of beard or face contrast are encoded by the global variable, while local variations like hair style or light direction are controlled by the local variable. In a simple dataset like MNIST (bottom), results show that handwriting global features as cursive style, contrast or thickness are encoded by $\boldsymbol{\beta}$, while the local $\mathbf{z}$ defines the shape of the digit. The characterization of whether these generative factors are local/global is based on an interpretation of the effect that varying $\mathbf{z}$ and $\boldsymbol{\beta}$ provokes in each image within a batch, and in the whole batch of images, respectively. In A.1, we reproduce the same figures for the all the clusters, in which we can appreciate that there is a significant fraction of clusters with visually interpretable global/local features.

We stress here again the fact that the UG-VAE training is fully unsupervised: data batches during training are completely randomly chosen from the training dataset, with no structured correlation whatsoever. Unlike other approaches for disentanglement, see [17] or [28], variational training in UG-VAE does not come with additional ELBO hyperparameters that need to be tuned to find a proper balance among terms in the ELBO.

One of the main contributions in the design of UG-VAE is the fact that, unless we include a clustering mixture prior in the local space controlled by the global variable $\boldsymbol{\beta}$, unsupervised learning of global factors is non-informative. To illustrate such a result, in Fig. 4 (last column) we reproduce the same results but for a probabilistic model in which the discrete local variable d is not included. Namely, we use the ML-VAE in Fig. 2(c) but we trained it with random data batches. In this case, the local space is uni-modal given $\boldsymbol{\beta}$ and we show interpolated values between -1 to 1. Note that the disentanglement effect of variations in both $\boldsymbol{\beta}$ and $\boldsymbol{z}$ is mild and hard to interpret.

It remains a challenge for generative models to obtain a quantitative appropriate metric for evaluating the quality of the generated images. In this work, we employ the FID (Frechet Inception Distance), proposed in [16], which summarizes the distance between the Inception feature vectors for real and generated images in the same domain, with the advantage that it is correlated with the better quality of the generated images. In Table 1 we include the score for samples from UG-VAE, ML-VAE and β-VAE. In both CelebA and MNIST, UG-VAE obtain lower distance and thus outperforms the other methods in the quality of the generated samples. We show empirically that the way the information is structured in the latent space of UG-VAE allows an improved generation of images. The reasons are: (i) differently from our model and ML-VAE, in β-VAE the global information shared by groups of samples is not captured. (ii) UG-VAE latent space is much more expressive than ML-VAE, where the conditional prior $p(\mathbf{z}|\boldsymbol{\beta})$ is unimodal. In other words, the prior $p(\mathbf{z}|d, \boldsymbol{\beta})$ in UG-VAE is a generalization of ML-VAE ($K = 1$). Therefore, in UG-VAE the latent space is augmented, which increases the representation capacity of the model.

Table 1
FID score between subsets of 1280 images from the test sets of CelebA and MNIST and 1280 images generated with UG-VAE, ML-VAE and β-VAE. Results are provided as the *mean* ± *std* FID score of 9 repetitions.

Method	UG-VAE	ML-VAE	β-VAE
CelebA	$\mathbf{162.3 \pm 1.2}$	204.7 ± 2.4	173.5 ± 0.6
MNIST	$\mathbf{63.6 \pm 2.4}$	108.9 ± 4.5	133.2 ± 0.8

In the following quantitative analysis, and with the intention at showing the wide spectrum for factor representation capacity provided by UG-VAE, we compute the FID metric between groups of CelebA images that share a given attribute, and samples generated by UG-VAE from a selected component d, in order to visualize whether the attributes are correlated with some of the components. These results are given in Fig. 6. As one may appreciate, UG-VAE is able to encode human perceptible factors within each component of the mixture, and images from the same attribute present different FID scores for the set of clusters. Within unimodal models (for instance, ML-VAE or β-VAE), such rich representation is not possible. We are able to obtain a set of basis faces that are tuned by the global variable. Incorporating the mixture allows β to control the distribution of clusters for representing groups that can

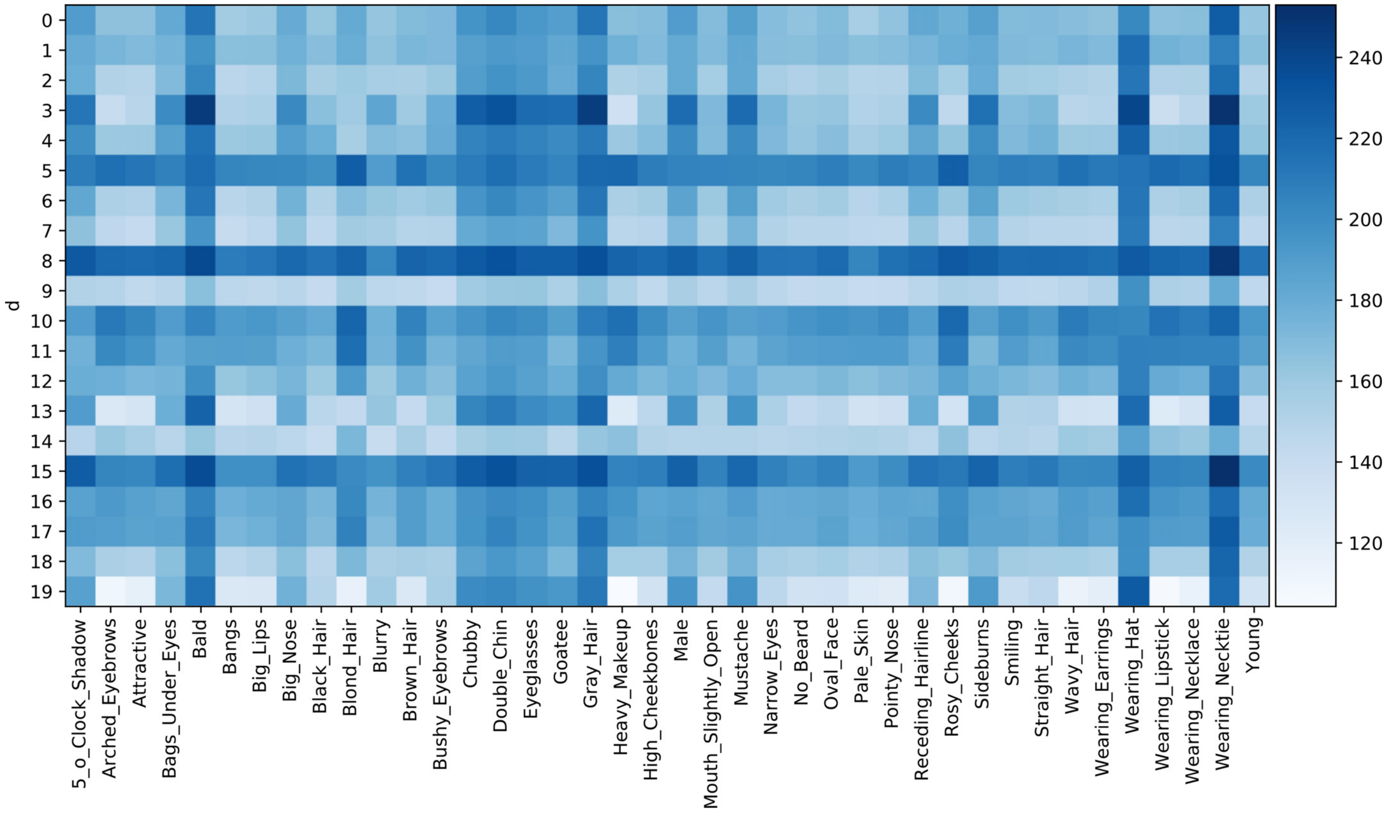
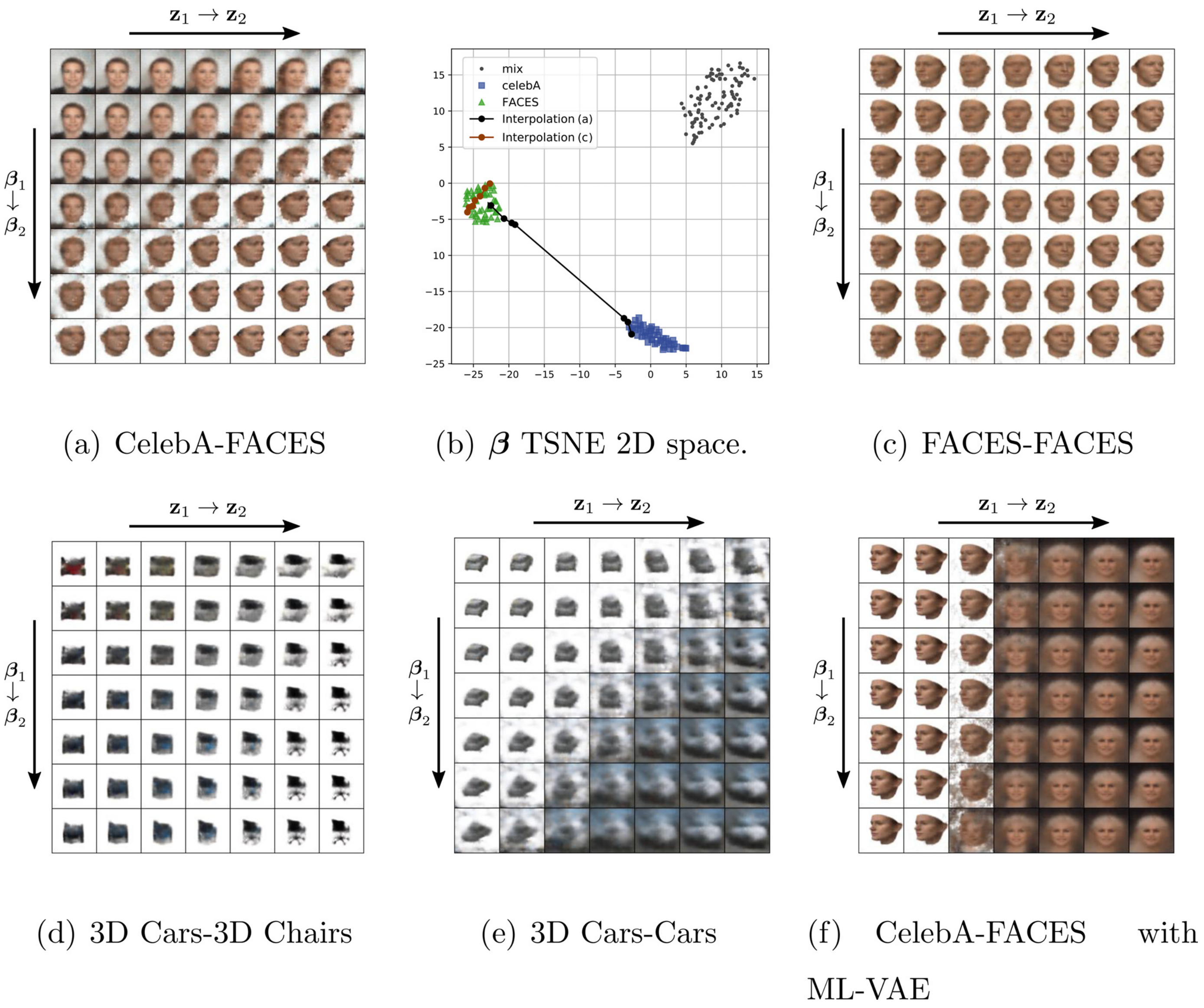


Fig. 6. FID score between subsets of 1280 images from CelebA with a given attribute and 1280 images generated with UG-VAE from a fixed cluster d.



(a) CelebA-FACES

(b) $\boldsymbol{\beta}$ TSNE 2D space.

(c) FACES-FACES

(d) 3D Cars-3D Chairs

(e) 3D Cars-Cars

(f) CelebA-FACES with ML-VAE

Fig. 7. Interpolation in local (columns) and global (rows) posterior spaces, fusing several datasets, using UG-VAE from (a) to (e). In (a) the interpolation goes between the posteriors of a sample from CelebA dataset and a sample from FACES dataset. In (b) we plot the t-SNE map of the samples from each dataset. In (c) the interpolation goes between samples from the same dataset. In (d) and (e) we include interpolations from 3D Cars to Chairs, and for 3D Cars to Cars Dataset, respectively. In (f) we reproduce the interpolation using the latent space of ML-VAE.

be compared to the semi-supervision applied in ML-VAE by grouping.

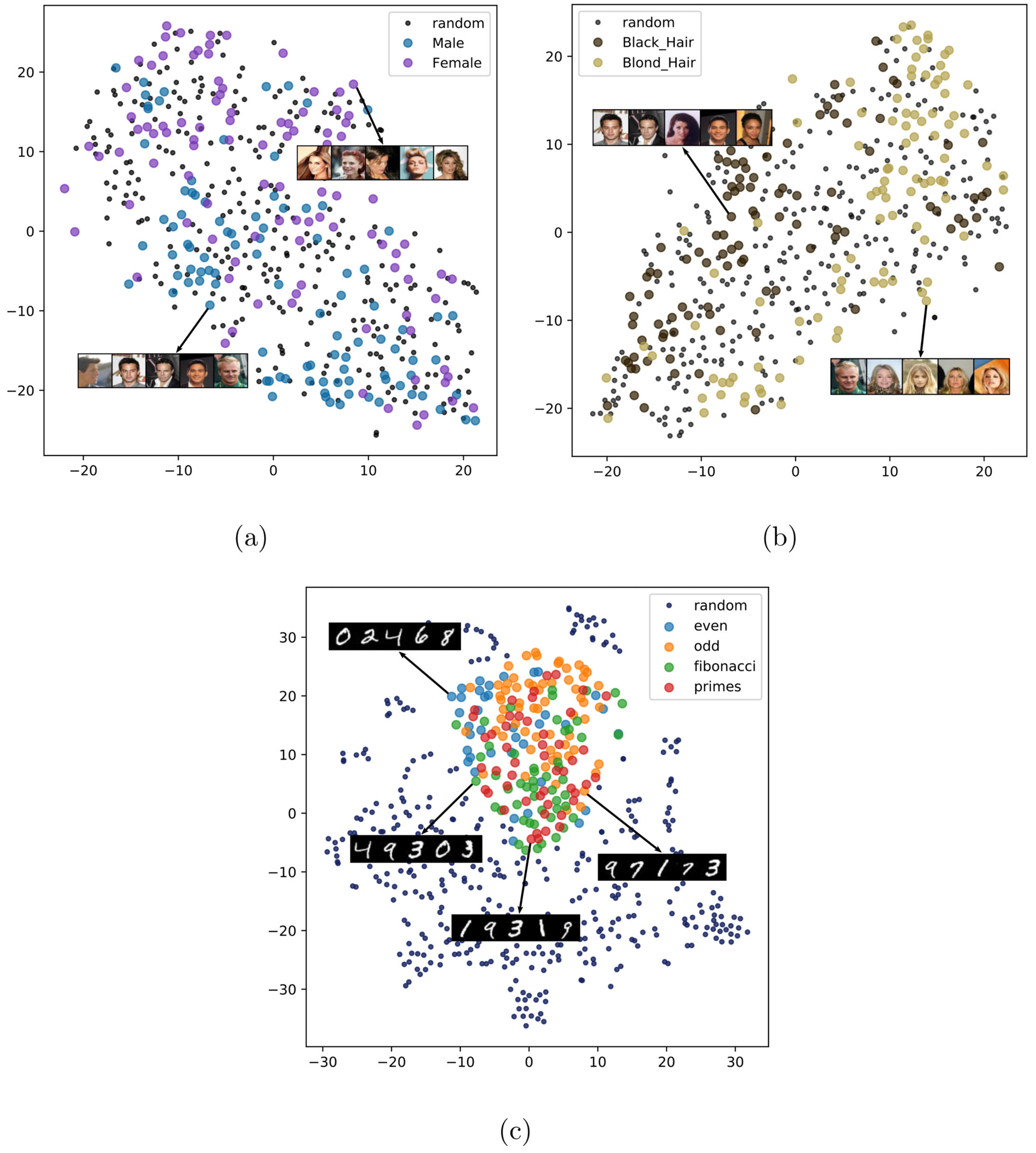
4.2. Domain alignment

In this section, we evaluate the UG-VAE performance in an unsupervised domain alignment setup. During training, the model is fed with data batches that include random samples coming from two different datasets. In particular, we train our model with a mixed dataset between CelebA and 3D FACES [30], a dataset of 3D scanned faces, with a proportion of 50% samples from each dataset inside each batch.

Upon training with random batches, in Fig. 7, we perform the following experiment using domain supervision to create test data batches. We create two batches containing only images from CelebA and 3D FACES. Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be the mean global posterior computed using (8) associated for each batch. For two particular images in these two batches, let $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ be the mean local posterior of these two images, computed using (3). Fig. 7(a) shows samples of the UG-VAE model when we linearly interpolate between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ (rows) and between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ (columns).[1] Certainly $\boldsymbol{\beta}$ is capturing the domain knowledge. For fixed $\mathbf{z}$, e.g. $\boldsymbol{z}_1$ in the first column, the interpolation between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ is transferring the CelebA image into the 3D FACES domain (note that background is turning white, and the image is rotated to get a 3D effect). Alternatively, for fixed $\boldsymbol{\beta}$, e.g. $\boldsymbol{\beta}_1$ in the first row, interpolating between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ modifies the first image into one that keeps the domain but resembles features of the image in the second domain, as face rotation.

In Fig. 7(b) we show the 2D t-SNE plot of the posterior distribution of $\boldsymbol{\beta}$ for batches that are random mixtures between datasets (grey points), batches that contain only CelebA faces (blue squares), and batches that contain only 3D faces (green triangles). We also add the corresponding points of the $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ interpolation in

[1] Note that since both $\boldsymbol{\beta}$ and $\boldsymbol{z}$ are deterministically interpolated, the discrete variable d plays no role to sample from the model.



(a)

(b)

(c)

Fig. 8. 2D t-SNE projection of the UG-VAE $\boldsymbol{\beta}$ posterior distribution of structured batches of 128 CelebA images. UG-VAE is trained with completely random batches of 128 train images.

Fig. 7(a). In Fig. 7(c), we reproduce the experiment in (a) but interpolating between two images and values of $\boldsymbol{\beta}$ that correspond to the same domain (brown interpolation line in Fig. 7(b)). As expected, the interpolation of $\boldsymbol{\beta}$ in this case does not change the domain, which suggests that the domain structure in the global space is smooth, and that the interpolation along the local space $\mathbf{z}$ modifies image features to translate one image into the other.

In Fig. 7(d,e) experiments with more datasets are included. When mixing the 3DCars dataset [10] with the 3D Chairs dataset [2], in Fig. 7(d), we find that certain correlations between cars and chairs are captured. Interpolating between a racing car and an office desk chair leads to a white car in the first domain (top right) and in a couch (bottom left). In Fig. 7(e), when using the 3D Cars along with the Cars Dataset [23], rotations in the cars are induced.

Table 2
Batch classification accuracy using samples of the posterior $\boldsymbol{\beta}$ distribution.

Batch categories	Classifier	Train accuracy	Test accuracy
Black (0) vs blond (1)	Linear SVM	1.0	0.95
	RBF SVM	1.0	0.98
Black (0) vs blond (1) vs random (2)	Linear SVM	0.91	0.54
	RBF SVM	0.85	0.56
Male (0) vs female (1)	Linear SVM	1.0	0.85
	RBF SVM	1.0	0.85
Male (0) vs female (1) vs random (2)	Linear SVM	0.84	0.66
	RBF SVM	0.89	0.63

Finally, in 7(f) we show that, as expected, the rich structured captured by UG-VAE is lost when we do not include the clustering effect in the local space, i.e. if we use ML-VAE with unsupervised random data batches, and all the transition between domains is performed within the local space.

4.3. UG-VAE representation of structured non-trivial data batches

In the previous subsection, we showed that the UG-VAE global space is able to separate certain structure in the data batches (e.g. data domain) even though during training batches did not present such an explicit correlation. Using UG-VAE trained over CelebA with unsupervised random batches of 128 images as a running example, in this section we want to further demonstrate this result.

In Fig. 8 we show the t-SNE 2D projection of structured batches using the posterior $\boldsymbol{\beta}$ distribution in (8) over CelebA and MNIST test images. In Fig. 8(a), we display the distribution of batches containing only men and women, while in Fig. 8(b) the distribution of batches containing people with black or blond hair. In both cases we show the distribution of randomly constructed batches as the ones in the training set. To some extend, in both cases we obtain separable distributions among the different kinds of batches. A quantitative evaluation can be found in Table 2. We have employed samples from the $\boldsymbol{\beta}$ distribution to train a supervised classifier that discriminates between different types of batches. When random batches are not taken as a class, the separability is evident. When random batches are included, it is expected that the classifier struggles to differentiate between a batch that contains 90% of male images and a batch that only contain male images, hence the drop in accuracy for the multi-case problem.

An extension with similar results when using structured grouped batches from MNIST dataset for testing our model is exposed in Fig. 8(c). In this experiment, the groups are digits that belong to certain mathematical series, including even numbers, odd numbers, Fibonacci series and prime numbers. We prove that UG-VAE is able to discriminate among their global posterior representations.

5. Conclusion

In this paper we have presented UG-VAE, an unsupervised deep generative model able to capture both local and global factors from batches of data samples. Unlike similar approaches in the literature, by combining a structured clustering prior in the local latent space with a Gaussian global prior and a structured variational family, we have demonstrated that interpretable group features can be inferred from the global space in a completely unsupervised fashion. Model training does not require artificial manipulation of the ELBO to force latent interpretability, which makes UG-VAE stand out w.r.t. most of the current disentanglement approaches using VAEs.

The ability of UG-VAE to infer diverse features from the training set is further demonstrated in a domain alignment setup, where we show that the global space allows interpolation between domains, and also by showing that images in correlated batches of data, related by non-trivial features such as hair color or gender in CelebA, define identifiable structures in the posterior global space.

The code is publicly available at https://github.com/ipeis/UG-VAE. The package includes the UG-VAE model, and all the experiments of this paper for reproducibility purposes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partly supported by Spanish government (AEI/MCI) under grants PID2021-123182OB-I00, PID2021-125159NB-I00 and RTI2018-099655-B-100, by Comunidad de Madrid under grant IND2022/TIC-23550, by the European Union (FEDER) and the European Research Council (ERC) through the European Union's Horizon 2020 research and innovation program under Grant 714161, and by Comunidad de Madrid and FEDER through IntCARE-CM. The work of Ignacio Peis has been also supported by by Spanish government (MIU) under grant FPU18/00516.

Appendix A. Extended experiments

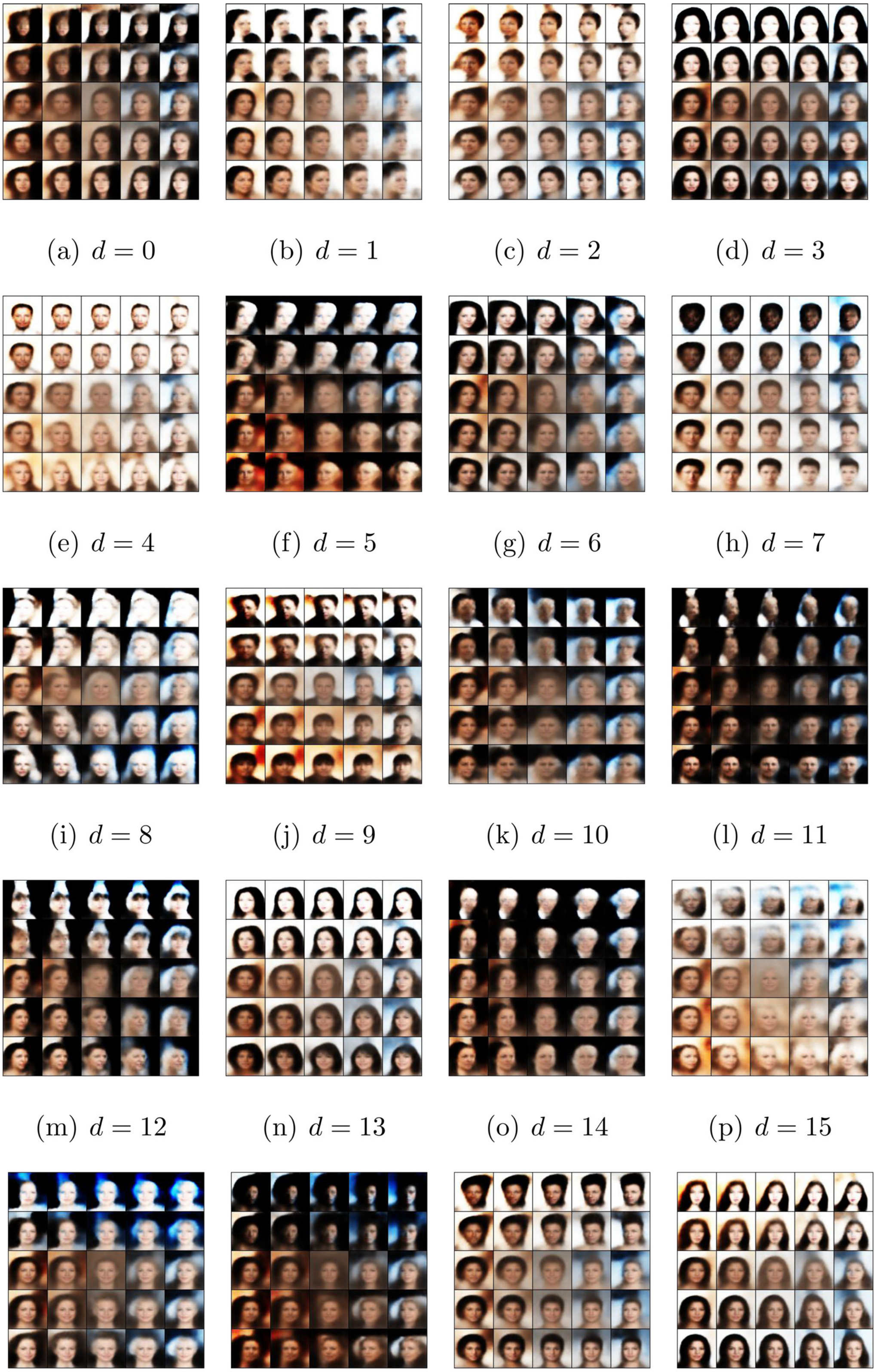
A1. Extended results for Section 4.1: Unsupervised learning of global factors

With the aim at evaluating whether a fraction of the clusters inferred by UG-VAE encode visually interpretable global/local features, in Fig. A.9 we include the results for CelebA for $K = 20$ clusters. We observe that a considerate proportion of the clusters captures disentangled generative factors. Moreover, considering the heterogeneity and variety in the generative factors of celebA faces (up to 40 different attributes), increasing the number of clusters might lead to capture more representative faces, and thus, generative global factors modulated by $\boldsymbol{\beta}$. In Fig. A.9, we appreciate that, apart from skin color, beard or image contrast, other generative factors controlled by the global variable are hair style (remarkable for components 9, 16, 17 or 18), sex (components 4 and 14), or background color (components 4, 16 and 17). In order to compare these results with a model trained on a small number of clusters, we include Fig. A.10 with samples from UG-VAE with $K = 4$. In this case, the model compresses the information of the whole dataset in only four modes, and thus, the variation of the samples within each cluster is higher.

A2. Extended results for Section 4.2: domain alignment

We include here the results of a interpolation in both the local space obtained when the number of components is $K = 1$, i. e., using the ML-VAE approach. As showed in Fig. A.11, when training ML-VAE with randomly grouped data, global space is not capable of capturing correlations between datasets, and the local space is in charge of encoding the transition from celebA to 3D FACES, which is performed within each row.

With the aim at reinforcing the robustness of UG-VAE in domain alignment, we include in Fig. A.12 the results of evaluating GMVAE with two clusters ($K = 2$) in a similar setup that in Section 4.2. As GMVAE does not have global variables, the interpolation only applies for the latent encodings in $\mathbf{z}$. Note that the interpolation is merely a gradual overlap between the two images.



(a) $d = 0$ (b) $d = 1$ (c) $d = 2$ (d) $d = 3$

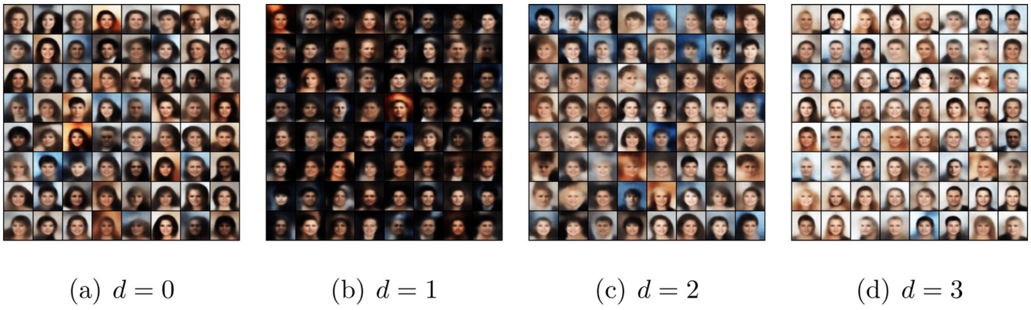
(e) $d = 4$ (f) $d = 5$ (g) $d = 6$ (h) $d = 7$

(i) $d = 8$ (j) $d = 9$ (k) $d = 10$ (l) $d = 11$

(m) $d = 12$ (n) $d = 13$ (o) $d = 14$ (p) $d = 15$

(q) $d = 16$ (r) $d = 17$ (s) $d = 185$ (t) $d = 185$

Fig. A1. Sampling from UG-VAE for CelebA. We include samples from each of the K = 20 clusters.



(a) $d = 0$ (b) $d = 1$ (c) $d = 2$ (d) $d = 3$

Fig. A2. Sampling from each cluster of UG-VAE for CelebA when $K = 4$.

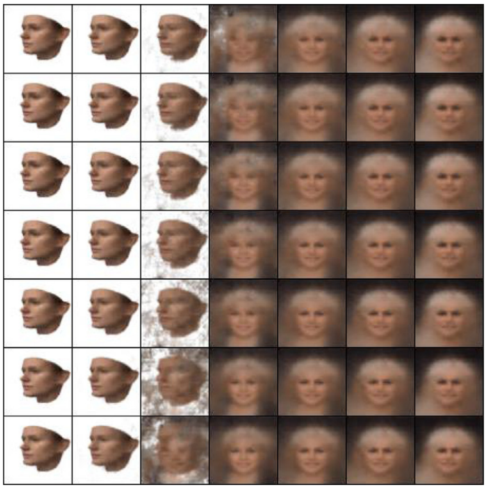


Fig. A3. ML-VAE interpolation in local (columns) and global (rows) posterior spaces, fusing celebA and FACES datasets.

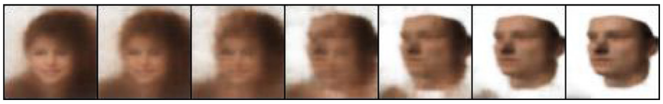


Fig. A4. Interpolation in the latent space of GMVAE with $K = 2$ for performing domain alignment, using the same network architecture than in the local part of UG-VAE. We interpolate between the encodings of images from CelebA and FACES dataset.

Namely, the model is not able to correlate the features of both images, regardless of their domain. On the other hand, with UG-VAE, by keeping fixed the global variable and interpolating in the local one, we maintain the domain but we translate the features of one image into the other. This analysis corroborates that the model finds this type of correlations in a clearly separated way.

Appendix B. Networks architecture

In this section we detail the architectures and parameters used for training the models exposed in the main paper. An extended overview is included in Table B.3.

Table B1
Architecture, parameters and hyperparameters for all the models trained for the experiments presented in the paper.

Dataset	Architecture				Params	Hyperparams
	Pre-encoder	Local encoder	Global encoder	Decoder		
CelebA	**h**: 5 CNN layers Filters: 32, 32, 64, 64, 256 Stride: all 4 Padding: All 1 ReLU activation Batch normalization	ϕ_z: Linear layer: $256 \to 2d$ First half $\boldsymbol{\mu}_z$ Second half diag($\boldsymbol{\Sigma}_z$) ϕ_d: Linear layers: $d \to 256 \to K$ Tanh activation Softmax output	ϕ_B: Linear layer: $256 + K \to 2g$ First half $\boldsymbol{\mu}_B$ Second half diag($\boldsymbol{\Sigma}_B$)	θ_z: Linear layers: $g \to 256 \to 2d$ First half $\boldsymbol{\mu}_z$ Second half diag($\boldsymbol{\Sigma}_z$) θ_x: Linear layer: $d + g \to 256$ 5 transpose CNN layers Filters: 64, 64, 32, 32, 3 Stride: 1, 4, 4, 4, 4 Padding: 0, 1, 1, 1, 1 ReLU activation Sigmoid output	d=20 $g = 50$ $K = 20$	$\sigma_x = 0.2$ $B = 128$
MNIST	**h**: Linear layer: $28 * 28 \to 256$ ReLU activation	ϕ_z: Linear layer: $256 \to 2d$ First half $\boldsymbol{\mu}_z$ Second half diag($\boldsymbol{\Sigma}_z$) ϕ_d: Linear layers: $d \to 256 \to K$ Tanh activation Softmax output	ϕ_B: Linear layer: $256 + K \to 2g$ First half $\boldsymbol{\mu}_B$ Second half diag($\boldsymbol{\Sigma}_B$)	θ_z: Linear layers: $g \to 256 \to 2d$ First half $\boldsymbol{\mu}_z$ Second half diag($\boldsymbol{\Sigma}_z$) θ_x: Linear layers: $d + g \to 256 \to 28 * 28$ ReLU activation Sigmoid output	$d = 10$ $g = 20$ $K = 10$	$\sigma_x = 0.2$ $B = 128$
CelebA + 3D FACES	Same than for CelebA				$d = 40$ $g = 40$ $K = 40$	$\sigma_x = 0.2$ $B = 128$
3D Cars-3D Chairs	Same than for CelebA				$d = 20$ $g = 20$ $K = 20$	$\sigma_x = 0.2$ $B = 128$
3D Cars-Cars	Same than for CelebA				$d = 20$ $g = 50$ $K = 20$	$\sigma_x = 0.2$ $B = 128$

References

[1] J. Antoran, A. Miguel, Disentangling and learning robust representations with natural clustering, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, 2019, pp. 694–699.

[2] M. Aubry, D. Maturana, A.A. Efros, B.C. Russell, J. Sivic, Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 3762–3769.

[3] S. Barocas, M. Hardt, A. Narayanan, Fairness in Machine Learning, 1, NIPS Tutorial, 2017.

[4] F.M. Bianchi, L. Livi, K.O. Mikalsen, M. Kampffmeyer, R. Jenssen, Learning representations of multivariate time series with missing data, Pattern Recognit. 96 (2019) 106973.

[5] D. Bouchacourt, R. Tomioka, S. Nowozin, Multi-level variational autoencoder: Learning disentangled representations from grouped observations, in: Proceedings of the 30nd AAAI Conference on Artificial Intelligence, 2018.

[6] P. Bromiley, Products and convolutions of gaussian probability density functions, Tina-Vision Memo 3 (4) (2003) 1.

[7] Y. Burda, R. Grosse, R. Salakhutdinov, Importance weighted autoencoders, arXiv preprint arXiv:1509.00519(2015).

[8] J. Chung, K. Kastner, L. Dinh, K. Goel, A.C. Courville, Y. Bengio, A recurrent latent variable model for sequential data, in: Proceedings of the Advances in neural information processing systems, 2015, pp. 2980–2988.

[9] N. Dilokthanakul, P.A.M. Mediano, M. Garnelo, M.C.H. Lee, H. Salimbeni, K. Arulkumaran, M. Shanahan, Deep unsupervised clustering with gaussian mixture variational autoencoders, arXiv preprint arXiv:1611.02648 (2016).

[10] S. Fidler, S. Dickinson, R. Urtasun, 3d object detection and viewpoint estimation with a deformable 3d cuboid model, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 611–619.

[11] M. Fraccaro, S.K. Sønderby, U. Paquet, O. Winther, Sequential neural models with stochastic layers, Adv. Neural Inf. Process. Syst. 29 (2016).

[12] J. Gordon, J.M. Hernández-Lobato, Bayesian semisupervised learning with deep generative models, arXiv preprint arXiv:1706.09751(2017).

[13] J. Gordon, J.M. Hernández-Lobato, Combining deep generative and discriminative models for bayesian semi-supervised learning, Pattern Recognit. 100 (2020) 107156.

[14] P. Gyawali, Z. Li, C. Knight, S. Ghimire, B.M. Horacek, J. Sapp, L. Wang, Improving disentangled representation learning with the beta bernoulli process, in: Proceedings of the IEEE International Conference on Data Mining (ICDM), IEEE, 2019, pp. 1078–1083.

[15] C. Heinze-Deml, N. Meinshausen, Conditional variance penalties and domain shift robustness, arXiv preprint arXiv:1710.11469(2017).

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, β-VAE: Learning basic visual concepts with a constrained variational framework (2016).

[18] H. Hosoya, Group-based learning of disentangled representations with generalizability for novel contents, in: Proceedings of the IJCAI, 2019, pp. 2506–2513.

[19] M.J. Johnson, D.K. Duvenaud, A. Wiltschko, R.P. Adams, S.R. Datta, Composing graphical models with neural networks for structured representations and fast inference, Adv. Neural Inf. Process. Syst. 29 (2016) 2946–2954.

[20] W. Joo, W. Lee, S. Park, I.-C. Moon, Dirichlet variational autoencoder, Pattern Recognit. 107 (2020) 107514.

[21] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114(2013).

[22] I. Korshunova, J. Degrave, F. Huszár, Y. Gal, A. Gretton, J. Dambre, Bruno: a deep recurrent model for exchangeable data, in: Proceedings of the Advances in Neural Information Processing Systems, 2018, pp. 7190–7198.

[23] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: Proceedings of the 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.

[24] C. Liu, Z. Liao, Y. Ma, K. Zhan, Stationary diffusion state neural estimation for multiview clustering, arXiv preprint arXiv:2112.01334(2021).

[25] S. Liu, J. Liu, Q. Zhao, X. Cao, H. Li, D. Meng, H. Meng, S. Liu, Discovering influential factors in variational autoencoders, Pattern Recognit. 100 (2020) 107166.

[26] C. Ma, S. Tschiatschek, K. Palla, J.M. Hernandez-Lobato, S. Nowozin, C. Zhang, EDDI: efficient dynamic discovery of high-value information with partial VAE, in: Proceedings of the International Conference on Machine Learning, PMLR, 2019, pp. 4234–4243.

[27] C. Ma, S. Tschiatschek, R. Turner, J.M. Hernández-Lobato, C. Zhang, VAEM: a deep generative model for heterogeneous mixed type data, Adv. Neural Inf. Process. Syst. 33 (2020) 11237–11247.

[28] E. Mathieu, T. Rainforth, N. Siddharth, Y.W. Teh, Disentangling disentanglement in variational autoencoders, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 4402–4412.

[29] A. Nazabal, P.M. Olmos, Z. Ghahramani, I. Valera, Handling incomplete heterogeneous data using vaes, Pattern Recognit. (2020) 107501.

[30] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, A 3D face model for pose and illumination invariant face recognition, in: Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance, IEEE, 2009, pp. 296–301.

[31] I. Peis, C. Ma, J.M. Hernández-Lobato, Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo, Adv. Neural Inf. Process. Syst. 35 (2022).

[32] R. Ranganath, D. Tran, D. Blei, Hierarchical variational models, in: Proceedings of the 6th International Conference on Machine Learning, 2016, pp. 324–333.

[33] F.J.R. Ruiz, M.K. Titsias, T. Cemgil, A. Doucet, Unbiased gradient estimation for variational auto-encoders using coupled markov chains, in: Proceedings of the Uncertainty in Artificial Intelligence, PMLR, 2021, pp. 707–717.

[34] D. Tang, D. Liang, T. Jebara, N. Ruozzi, Correlated variational auto-encoders, in: Proceedings of the International Conference on Machine Learning, PMLR, 2019, pp. 6135–6144.

[35] J. Tomczak, M. Welling, Vae with a vampprior, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2018, pp. 1214–1223.

[36] A. Van Den Oord, O. Vinyals, et al., Neural discrete representation learning, Adv. Neural Inf. Process. Syst. 30 (2017).

[37] M.J. Vowels, N.C. Camgoz, R. Bowden, NestedVAE: isolating common factors via weak supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9202–9212.

[38] K. Xu, A. Srivastava, C. Sutton, Variational russian roulette for deep bayesian nonparametrics, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 6963–6972.

Ignacio Peis obtained his B.Sc. in Telecommunication Engineering from Universidad de Granada in 2016 and two M.Sc in Telecommunication Engineering and Multimedia and Communications from Universidad Carlos III de Madrid, in 2018. Since 2018 he is a PhD student in the Dept. of Signal Theory and Communications at the Universidad Carlos III de Madrid. His research interests lie on probabilistic machine learning, deep generative models, approximate inference, Bayesian inference and their methods and challenges: dealing with incomplete, heterogeneous data or temporal sequences. More details and publications are accessible at http://www.tsc.uc3m.es/~ipeis/.

Pablo M. Olmos was born in Granada, Spain, in 1984. He received the B.Sc./M.Sc. and Ph.D. degrees from the University of Sevilla in 2008 and 2011, respectively, all in telecommunication engineering. He is currently an Associate Professor with the Universidad Carlos III de Madrid. He has held appointments as a Visiting Researcher at Princeton University, École Polytechnique Fédérale de Lausanne, Notre Dame University, École Nationale Supérieure de l'Electronique et de ses Applications, and Nokia-Bell Labs. His research interests range from approximate inference methods for Bayesian machine learning to information theory and digital communications. A detailed CV and list of publications can be accessed at http://www.tsc.uc3m.es/olmos.

Antonio Artés-Rodríguez was born in Alhama de Almería, Spain, in 1963. He received the Ingeniero de Telecomunicación and Doctor Ingeniero de Telecomunicación degrees, both from the Universidad Politécnica de Madrid, Madrid, Spain, in 1988 and 1992, respectively. He is a Professor at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid. Prior to this, he held different teaching positions at Universidad de Vigo, Universidad Politécnica de Madrid, and Universidad de Alcalá, all of them in Spain. He has participated in more than 70 projects and contracts and has coauthored more that 50 journal papers and more than 100 international conference papers. His research interests include signal processing, machine learning, and information theory methods, and its application to health and sensor networks.