

---

# Learning Flat Latent Manifolds with VAEs

---

Nutan Chen<sup>1</sup> Alexej Klushyn<sup>1</sup> Francesco Ferroni<sup>2</sup> Justin Bayer<sup>1</sup> Patrick van der Smagt<sup>1</sup>

## Abstract

Measuring the similarity between data points often requires domain knowledge, which can in parts be compensated by relying on unsupervised methods such as latent-variable models, where similarity/distance is estimated in a more compact latent space. Prevalent is the use of the Euclidean metric, which has the drawback of ignoring information about similarity of data stored in the decoder, as captured by the framework of Riemannian geometry. We propose an extension to the framework of variational auto-encoders allows learning *flat latent manifolds*, where the Euclidean metric is a proxy for the similarity between data points. This is achieved by defining the latent space as a Riemannian manifold and by regularising the metric tensor to be a scaled identity matrix. Additionally, we replace the compact prior typically used in variational auto-encoders with a recently presented, more expressive hierarchical one—and formulate the learning problem as a constrained optimisation problem. We evaluate our method on a range of data-sets, including a video-tracking benchmark, where the performance of our unsupervised approach nears that of state-of-the-art supervised approaches, while retaining the computational efficiency of straight-line-based approaches.

## 1. Introduction

Measuring the distance between data points is a central ingredient of many data analysis and machine learning applications. Several kernel methods (KernelPCA (Schölkopf et al., 1997), KernelNMF (Li & Ding, 2006), etc.), and other non-parametric approaches such as k-nearest neighbours (Altman, 1992) rely on the availability of a suitable distance

<sup>1</sup>Machine Learning Research Lab, Volkswagen Group, Munich, Germany <sup>2</sup>Autonomous Intelligent Driving GmbH, Munich, Germany. Correspondence to: Nutan Chen <nutan.chen@gmail.com>.

*Proceedings of the 37<sup>th</sup> International Conference on Machine Learning*, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

function. Computer vision pipelines, e.g. tracking over time, perform matching based on similarity scores.

But designing a distance function can be challenging: it is not always obvious to write down mathematical formulae that accurately express a notion of similarity. Learning such functions has hence been proven as a viable alternative to manual engineering in this respect (NCA (Goldberger et al., 2005), metric learning (Xing et al., 2003), etc.). Often, these methods rely on the availability of pairs labelled as similar or dissimilar. A different route is that of exploiting the structure that latent-variable models learn. The assumption that a set of high-dimensional observations is explained by points in a much simpler latent space underpins these approaches. In their respective probabilistic versions, a latent prior distribution is transformed non-linearly to give rise to a distribution of observations. The hope is that simple distances, such as the Euclidean distance measured in latent space, implement a function of similarity. Yet, these approaches do not incorporate the variation of the observations with respect to the latent points. For example, the observations will vary much more when a path in latent space will cross a class boundary.

In fact, recent approaches to non-linear latent variable models, such as the generative adversarial network (Goodfellow et al., 2014) or the variational auto-encoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014), regularise the latent space to be compact, i.e. to remove low-density regions. This is in contrast to the aforementioned hope that Euclidean distances appropriately reflect similarity.

The above insight leads us to the development of *flat manifold* variational auto-encoders. This class of VAEs defines the latent space as Riemannian manifold and regularises the Riemannian metric tensor to be a scaled identity matrix. In this context, a *flat manifold* is a Riemannian manifold, which is isometric to the Euclidean space. To not compromise the expressiveness, we relax the compactness assumption and make use of a recently introduced hierarchical prior (Klushyn et al., 2019). As a consequence, the model is capable of learning a latent representation, where the Euclidean metric is a proxy for the similarity between data points. This results in a computational efficient distance metric which is practical for applications in real-time scenarios.

## 2. Variational Auto-Encoders with Flat Latent Manifolds

### 2.1. Background on Learning Hierarchical Priors in VAEs

Latent-variable models are defined as

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}, \quad (1)$$

where  $\mathbf{z} \in \mathbb{R}^{N_z}$  represents latent variables and  $\mathbf{x} \in \mathbb{R}^{N_x}$  the observable data. The integral in Eq. (1) is usually intractable but it can be approximated by maximising the evidence lower bound (ELBO) (Kingma & Welling, 2014; Rezende et al., 2014):

$$\begin{aligned} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} [\log p_{\theta}(\mathbf{x})] &\geq \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \left[ \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] \right. \\ &\quad \left. - \text{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \right], \end{aligned} \quad (2)$$

where  $p_{\mathcal{D}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i)$  is the empirical distribution of the data  $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ . The distribution parameters of the approximate posterior  $q_{\phi}(\mathbf{z}|\mathbf{x})$  and the likelihood  $p_{\theta}(\mathbf{x}|\mathbf{z})$  are represented by neural networks. The prior  $p(\mathbf{z})$  is usually defined as a standard normal distribution. This model is commonly referred to as the variational auto-encoder (VAE).

However, a standard normal prior often leads to an over-regularisation of the approximate posterior, which results in a less informative learned latent representation of the data (Tomczak & Welling, 2018; Klushyn et al., 2019). To enable the model to learn an informative latent representation, Klushyn et al. (2019) propose to use a flexible hierarchical prior  $p_{\Theta}(\mathbf{z}) = \int p_{\Theta}(\mathbf{z}|\zeta) p(\zeta) d\zeta$ , where  $p(\zeta)$  is the standard normal distribution. Since the optimal prior is the aggregated posterior (Tomczak & Welling, 2018), the above integral is approximated by an importance-weighted (IW) bound (Burda et al., 2015) based on samples from  $q_{\phi}(\mathbf{z}|\mathbf{x})$ . This leads to a model with two stochastic layers and the following upper bound on the KL term:

$$\begin{aligned} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \text{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) &\leq \mathcal{F}(\phi, \Theta, \Phi) \\ &\equiv \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} \left[ \log q_{\phi}(\mathbf{z}|\mathbf{x}) \right. \\ &\quad \left. - \mathbb{E}_{\zeta_1:K \sim q_{\Phi}(\zeta|\mathbf{z})} \left[ \log \frac{1}{K} \sum_{i=1}^K \frac{p_{\Theta}(\mathbf{z}|\zeta_i) p(\zeta_i)}{q_{\Phi}(\zeta_i|\mathbf{z})} \right] \right], \end{aligned} \quad (3)$$

where  $K$  is the number of importance samples. Since it has been shown that high ELBO values do not necessarily correlate with informative latent representations (Alemi et al., 2018; Higgins et al., 2017)—which is also the case for hierarchical models (Sønderby et al., 2016)—different optimisation approaches have been introduced (Bowman

et al., 2016; Sønderby et al., 2016). Klushyn et al. (2019) follow the line of argument in (Rezende & Viola, 2018) and reformulate the resulting ELBO as the Lagrangian of a constrained optimisation problem:

$$\begin{aligned} \mathcal{L}_{\text{VHP}}(\theta, \phi, \Theta, \Phi; \lambda) &\equiv \\ \mathcal{F}(\phi, \Theta, \Phi) + \lambda (\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [C_{\theta}(\mathbf{x}, \mathbf{z})] - \kappa^2), \end{aligned} \quad (4)$$

with the optimisation objective  $\mathcal{F}(\phi, \Theta, \Phi)$ , the inequality constraint  $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [C_{\theta}(\mathbf{x}, \mathbf{z})] \leq \kappa^2$ , and the Lagrange multiplier  $\lambda$ .  $C_{\theta}(\mathbf{x}, \mathbf{z})$  is defined as the reconstruction-error-related term in  $-\log p_{\theta}(\mathbf{x}|\mathbf{z})$ . Thus, we obtain the following optimisation problem:

$$\min_{\Theta, \Phi} \min_{\theta} \max_{\lambda} \min_{\phi} \mathcal{L}_{\text{VHP}}(\theta, \phi, \Theta, \Phi; \lambda) \quad \text{s.t. } \lambda \geq 0. \quad (5)$$

Building on that, the authors propose an optimisation algorithm—including a  $\lambda$ -update scheme—to achieve a tight lower bound on the log-likelihood. This approach is referred to as variational hierarchical prior (VHP) VAE.

### 2.2. Learning Flat Latent Manifolds with VAEs

The VHP-VAE is able to learn a latent representation that corresponds to the topology of the data manifold (Klushyn et al., 2019). However, it is not guaranteed that the (Euclidean) distance between encoded data in the latent space is a sufficient distance metric in relation to the observation space. In this work, we aim to measure the distance/difference of observed data directly in the latent space by means of the Euclidean distance of the encodings.

Chen et al. (2018a); Arvanitidis et al. (2018) define the latent space of a VAE as a Riemannian manifold. This approach allows for computing the observation-space length of a trajectory  $\gamma : [0, 1] \rightarrow \mathbb{R}^{N_z}$  in the latent space:

$$L(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^T \mathbf{G}(\gamma(t)) \dot{\gamma}(t)} dt, \quad (6)$$

where  $\mathbf{G} \in \mathbb{R}^{N_z \times N_z}$  is the Riemannian metric tensor, and  $\dot{\gamma}(t)$  the time derivative. We define the *observation-space distance* as the shortest possible path

$$D = \min_{\gamma} L(\gamma) \quad (7)$$

between two data points. The trajectory  $\gamma = \arg \min_{\gamma} L(\gamma)$  that minimises  $L(\gamma)$  is referred to as the (minimising) geodesic. In the context of VAEs,  $\gamma$  is transformed by a continuous function  $f(\gamma(t))$ —the decoder—to the observation space. The metric tensor is defined as  $\mathbf{G}(\mathbf{z}) = \mathbf{J}(\mathbf{z})^T \mathbf{J}(\mathbf{z})$ , where  $\mathbf{J}$  is the Jacobian of the decoder.

To measure the *observation-space distance* directly in the latent space, distances in the observation space should be proportional to distances in the latent space:

$$D \propto \|\mathbf{z}(1) - \mathbf{z}(0)\|_2, \quad (8)$$

where we define the Euclidean distance as the distance metric. This requires that the Riemannian metric tensor is  $\mathbf{G} \propto \mathbb{1}$ . As a consequence, the Euclidean distance in the latent space corresponds to the *observation-space distance*. We refer to a manifold with this property as *flat manifold* (Lee, 2006). To obtain a *flat latent manifold*, the model typically needs to learn complex latent representations of the data (see experiments in Sec. 4). Therefore, we propose the following approach: (i) to enable our model to learn complex latent representations, we apply a flexible prior (VHP), which is learned by the model (empirical Bayes); and (ii) we regularise the curvature of the decoder such that  $\mathbf{G} \propto \mathbb{1}$ .

For this purpose, the VHP-VAE, introduced in Sec. 2.1, is extended by a Jacobian-regularisation term. We define the regularisation term as part of the optimisation objective, which is in line with the constrained optimisation setting. The resulting objective function is

$$\mathcal{L}(\theta, \phi, \Theta, \Phi; \lambda, \eta, c^2) = \mathcal{L}_{\text{VHP}}(\theta, \phi, \Theta, \Phi; \lambda) + \eta \left( \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\|\mathbf{G}(\mathbf{z}) - c^2 \mathbb{1}\|_2^2] \right), \quad (9)$$

where  $\eta$  is a hyper-parameter determining the influence of the regularisation and  $c$  the scaling factor. Additionally, we use a stochastic approximation (first order Taylor expansion) of the Jacobian to improve the computational efficiency (Rifai et al., 2011b):

$$\mathbf{J}_t(\mathbf{z}) = \lim_{\sigma \rightarrow 0} \frac{1}{\sigma} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2)} [f(\mathbf{z} + \epsilon e_t) - f(\mathbf{z})], \quad (10)$$

where  $\mathbf{J}_t \in \mathbb{R}^{N_x}$  is the Jacobian of the  $t$ -th latent dimension and  $e_t$  a standard basis vector. This approximation method allows for a faster computation of the gradient and avoids the second-derivative problem of piece-wise linear layers (Chen et al., 2018a).

However, the influence of the regularisation term in Eq. (9) on the decoder function is limited to regions where data is available. To overcome this issue, we propose to use *mixup*, a data-augmentation method (Zhang et al., 2018), which was introduced in the context of supervised learning. We extend this method to the VAE framework (unsupervised learning) by applying it to encoded data in the latent space. This approach allows augmenting data by interpolating between two encoded data points  $\mathbf{z}_i$  and  $\mathbf{z}_j$ :

$$g(\mathbf{z}_i, \mathbf{z}_j) = (1 - \alpha) \mathbf{z}_i + \alpha \mathbf{z}_j, \quad (11)$$

with  $\mathbf{x}_i, \mathbf{x}_j \sim p_{\mathcal{D}}(\mathbf{x})$ ,  $\mathbf{z}_i \sim q_{\phi}(\mathbf{z}|\mathbf{x}_i)$ ,  $\mathbf{z}_j \sim q_{\phi}(\mathbf{z}|\mathbf{x}_j)$ , and  $\alpha \sim U(-\alpha_0, 1 + \alpha_0)$ . In contrast to (Zhang et al., 2018), where  $\alpha \in [0, 1]$  limits the data augmentation to only convex combinations, we define  $\alpha_0 > 0$  to take into account the outer edge of the data manifold. By combining *mixup* in Eq. (11) with Eq. (9), we obtain the objective

function of our *flat manifold* VAE (FMVAE):

$$\begin{aligned} \mathcal{L}_{\text{VHP-FMVAE}}(\theta, \phi, \Theta, \Phi; \lambda, \eta, c^2) &= \mathcal{L}_{\text{VHP}}(\theta, \phi, \Theta, \Phi; \lambda) + \\ &\eta \mathbb{E}_{\mathbf{x}_{i,j} \sim p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{\mathbf{z}_{i,j} \sim q_{\phi}(\mathbf{z}|\mathbf{x}_{i,j})} [\|\mathbf{G}(g(\mathbf{z}_i, \mathbf{z}_j)) - c^2 \mathbb{1}\|_2^2]. \end{aligned} \quad (12)$$

Inspired by batch normalisation, we define the squared scaling factor to be the mean over the batch samples and diagonal elements of  $\mathbf{G}$  (see App. A.2 for empirical support):

$$c^2 = \frac{1}{N_z} \mathbb{E}_{\mathbf{x}_{i,j} \sim p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{\mathbf{z}_{i,j} \sim q_{\phi}(\mathbf{z}|\mathbf{x}_{i,j})} [\text{tr}(\mathbf{G}(g(\mathbf{z}_i, \mathbf{z}_j)))]. \quad (13)$$

The optimisation algorithm Alg. 1, and further details about the optimisation process can be found in App. A.4.

By using augmented data, we regularise  $\mathbf{G}$  to be a scaled identity matrix for the *entire* latent space enclosed by the data manifold. As a consequence, the function  $f(\mathbf{z})$  (decoder) is—up to the scaling factor  $c$ —distance-preserving since  $D_{\mathbf{x}}(f(\mathbf{z}_i), f(\mathbf{z}_j)) \approx c D_{\mathbf{z}}(\mathbf{z}_i, \mathbf{z}_j)$ , where  $D_{\mathbf{x}}$  and  $D_{\mathbf{z}}$  refer to the distance in the observation and latent space, respectively.

### 3. Related Work

**Interpretation of the VAE’s latent space.** In general, the latent space of VAEs is considered to be Euclidean (e.g. Kingma et al., 2016; Higgins et al., 2017), but it is not constrained to be Euclidean. This can be problematic if we require a precise metric that is based on the latent space. Some recent works (Mathieu et al., 2019; Grattarola et al., 2018) adapted the latent space to be non-Euclidean to match the data structure. We solve the problem from another perspective: we enforce the latent space to be Euclidean.

**Jacobian and Hessian regularisation.** In (Rifai et al., 2011a), the authors proposed to regularise the Jacobian and Hessian of the encoder. However, it is more difficult to augment data in the observation space than in the latent space. In (Hadjeres et al., 2017), the Jacobian of the decoder was regularised to be as small as possible/zero. On the contrary, we regularise the the Riemannian metric tensor to be a scaled identity matrix, and hence the Jacobian to be constant, and hence the Hessian to be zero. (Nie & Patel, 2019) regularised the Jacobian with respect to the weights for GANs. In terms of supervised learning, (Jakubovitz & Giryes, 2018) used Jacobian regularisation to improve the robustness for classification.

**Metric learning.** Various metric learning approaches for both deep supervised and unsupervised models were proposed. For instance, deep metric learning (Hoffer & Ailon, 2015) used a triplet network for supervised learning. (Karaletsos et al., 2016) introduced an unsupervised metric learning method, where a VAE is combined with triplets. How-

ever, a human oracle is still required. By contrast, our approach is completely based on unsupervised learning, using the Euclidean distance in the latent space as a distance metric. Our proposed method is similar to the metric learning methods such as Large Margin Nearest Neighbour (Weinberger & Saul, 2009), which pulls target neighbours together and pushes impostors away. The difference is that our approach is an unsupervised method.

**Constraints in latent space.** Constraints on time (e.g. Wang et al., 2007; Chen et al., 2016; 2015) allow obtaining similar distance metrics in the latent space. Additionally, due to the missing data between different sequence steps, constraints on time cannot guarantee that the metric is correct between different sequences. However, our method can be used for general data-sets.

**Data augmentation.** The latent space is formed arbitrarily in regions where data is missing. Zhang et al. (2018) proposed *mixup*, an approach for augmenting data and labels for supervised learning. Various follow-up studies of *mixup* were developed, such as (Verma et al., 2019; Beckham et al., 2019). (Verma et al., 2019) considered *mixup* of hidden representations of training data to flatten the class-specific state distribution. We extend *mixup* to the VAE framework (unsupervised learning) by applying it to encoded data in the latent space of generative models. This facilitates the regularisation of regions where no data is available. As a consequence, similarity of data points can be measured in the latent space by applying the Euclidean metric.

**Geodesic.** Recent studies on geodesics for generative models (e.g. Tosi et al., 2014; Chen et al., 2018a; Arvanitidis et al., 2018) are focusing on methods for computing/finding the geodesic in the latent space. By contrast, we use the geodesic/Riemannian distance for influencing the learned latent manifold. (Frenzel et al., 2019) projected the latent space to a new latent space, where the geodesic is equivalent to the Euclidean interpolation. However, these two separate processes—VAEs and projection—probably hinder the model to find the latent features autonomously. Another difference is the assumption of previous work is that distances, defined by geodesics, can only be measured by following the data manifold. This is useful in scenarios such as avoiding unseen barriers between two data points, e.g., (Chen et al., 2018b), but it does not allow measuring distances between different categories. In this work, we focus on learning a general distance metric.

## 4. Experiments

We test our method on artificial pendulum images, human motion data, MNIST, and MOT16. We measure the performance in terms of equidistances, interpolation smoothness, and distance computation. Additionally, our method is ap-

plied to a real-world environment—a video-tracking benchmark. Here, the tracking and re-identification capabilities are evaluated.

The Riemannian metric tensor has many intrinsic properties of a manifold and measures local angles, length, surface area, and volumes (Bronstein et al., 2017). Therefore, the models are quantified based on the Riemannian metric tensor by computing condition numbers and magnification factors. The condition number, which shows the ratio of the most elongated to the least elongated direction, is defined as  $k(\mathbf{G}) = \frac{S_{\max}(\mathbf{G})}{S_{\min}(\mathbf{G})}$ , where  $S_{\max}$  is the largest eigenvalue of  $\mathbf{G}$ . The magnification factor  $MF(\mathbf{z}) \equiv \sqrt{\det \mathbf{G}(\mathbf{z})}$  (Bishop et al., 1997) depicts the sensitivity of the likelihood functions. When projecting from the Riemannian (latent) to the Euclidean (observation) space, the MF can be considered a scaling coefficient. Since we cannot directly compare the MFs of different models, the MFs are normalised/divided by their means. The closer the conditional number and the normalised MF are to one, the more invariant is the model with respect to the Riemannian metric tensor. In other words: the conditional number and the normalised MF are metrics to evaluate whether  $\mathbf{G}(\mathbf{z})$  is approximately constant and proportional to  $\mathbb{1}$ .

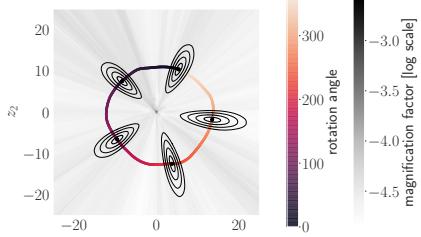
In order to make the visualisations of the magnification factor in Sec. 4.1 (Fig. 1) and Sec. 4.2 (Fig. 3 & Fig. 7) comparable, we define the respective upper range of the colour-bar as  $\frac{\max(MF_{VAE-VHP}(\text{grid\_area})) \cdot \text{mean}(MF_{VHP-FMVAE}(\text{data}))}{\text{mean}(MF_{VHP-FMVAE}(\text{data}))}$ .  $MF(\text{data})$  and  $MF(\text{grid\_area})$  are computed with training data and by using a grid area, respectively.

To be in line with previous literature (e.g. Higgins et al., 2017; Sønderby et al., 2016), we use the  $\beta$ -parametrisation of the Lagrange multiplier  $\beta = \frac{1}{\lambda}$  in our experiments.

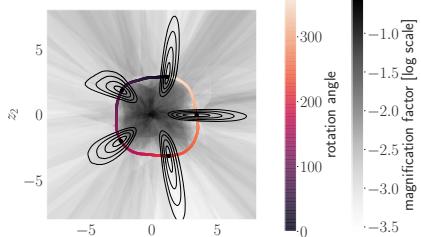
### 4.1. Artificial Pendulum Data-set

The pendulum data-set (Klushyn et al., 2019; Chen et al., 2018a) consists of  $16 \times 16$ -pixel images generated by a pendulum simulator. The pendulum has one degree of freedom, and the joint is located in the centres of the images. We generated  $15 \cdot 10^3$  images with joint angles uniformly in the ranges of  $[0, 360)$  degrees. Additionally, we added 0.05 Gaussian noise to each pixel.

As seen in Fig. 1, without regularisation, the contour lines are denser in the centre of the latent space. The reason is that, in contrast to the VHP-VAE, the regularisation term in the VHP-FMVAE *smoothes* the latent space ( $\mathbf{G} \approx c \mathbb{1}$ )—visualised by the MF and the equidistance plots. In Fig. 2, VHP-FMVAE and VAE-VHP are compared in terms of condition number and normalised MF. In both cases the VHP-FMVAE outperforms the VHP-VAE.

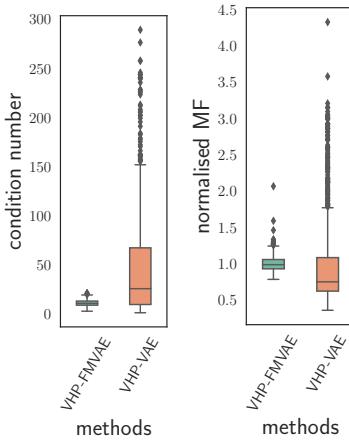


(a) VHP-FMVAE



(b) VHP-VAE

**Figure 1.** Latent representation of pendulum data: the contour plots illustrate curves of equal *observation-space distance* to the respective encoded data point. Distances are calculated using Eq. (6). The grey-scale displays  $MF(\mathbf{z})$ . Note: round, homogeneous contour plots indicate that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ .

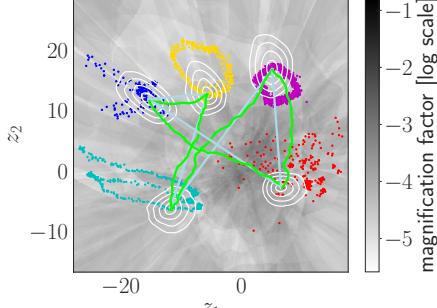


**Figure 2.** Pendulum data: if both the condition number and the normalised MF values are close to one, it indicates that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ . The box-plots are based on 1,000 generated samples.

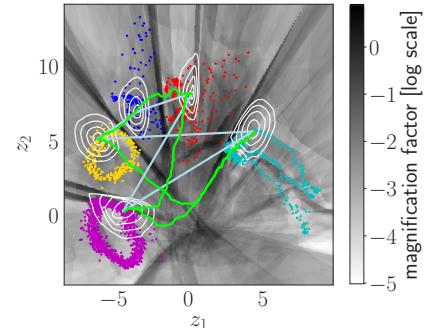
#### 4.2. Human Motion Capture Database

To evaluate our approach on the CMU human motion data-set (<http://mocap.cs.cmu.edu>), we select five different movements: walking (subject 35), jogging (subject 35), balancing (subject 49), punching (subject 143), and kicking (subject 74). After data pre-processing, the input data is a 50-dimensional vector of the joint angles. Note that the data-set is not balanced: walking, for example, has

more data points than jogging.



(a) VHP-FMVAE



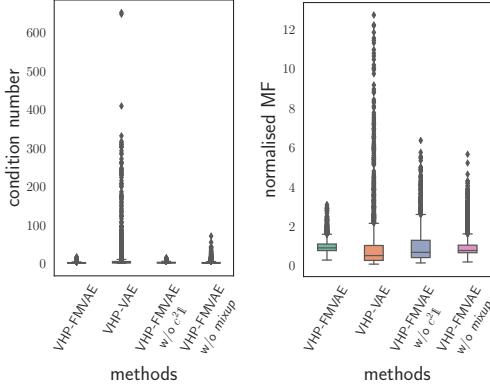
(b) VAE-VHP

**Figure 3.** Latent representation of human motion data: the contour plots illustrate curves of equal *observation-space distance* to the respective encoded data point. The grey-scale displays  $MF(\mathbf{z})$ . Note: round, homogeneous contour plots indicate that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ . In case of the VHP-FMVAE (a), Jogging is a large-range movement compared with walking, so that jogging is reasonably distributed on a larger area in the latent space than walking. By contrast, in case of the VHP-VAE (b), the latent representation of walking is larger than the one of jogging. Additionally, geodesics are compared to the corresponding Euclidean interpolations. The Euclidean interpolations in (a) are much closer to the geodesics.

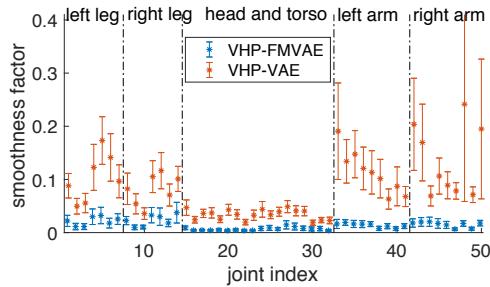
**Table 1.** Verification of the distance metric. The table shows the length ratio of the Euclidean interpolation to the geodesic. Additionally, we list the ratio of the related distances in the observation space.

DATA-SET	METHOD	OBSERVATION	LATENT
HUMAN	VHP-FMVAE	<b><math>1.02 \pm 0.06</math></b>	<b><math>0.93 \pm 0.03</math></b>
	VHP-VAE	$1.23 \pm 0.20$	$0.82 \pm 0.10$
MNIST	VHP-FMVAE	<b><math>1.01 \pm 0.08</math></b>	<b><math>0.92 \pm 0.05</math></b>
	VHP-VAE	$1.13 \pm 0.22$	$0.70 \pm 0.31$

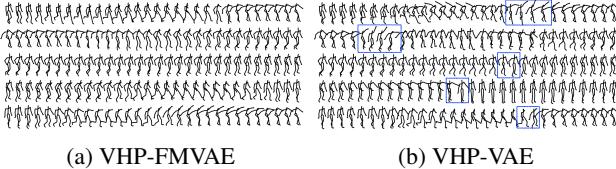
**Equidistance plots.** In Fig. 3, we randomly select a data point from each class as centres of the equidistance plots.



**Figure 4.** Human motion data: if both the condition number and the normalised MF values are close to one, it indicates that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ . The box-plots are based on 3,000 generated samples.



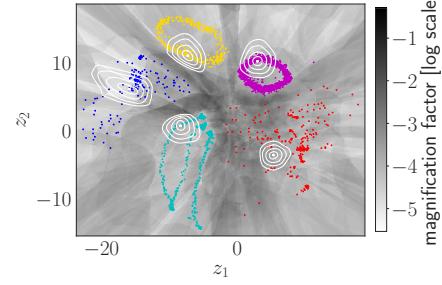
**Figure 5.** Smoothness measure of the human-movement interpolations. The mean and standard deviation are displayed for each joint: the smaller the value, the smoother the interpolation.



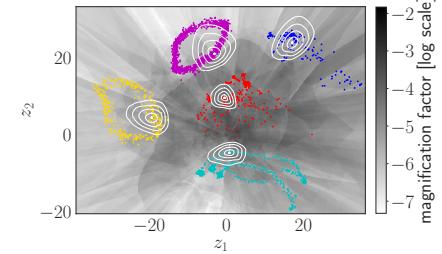
**Figure 6.** Human-movement reconstructions of Euclidean interpolations in the latent space. Discontinuities in the motions are marked by blue boxes.

In case of our proposed method, the equidistance plots are homogeneous, while in case of the VHP-VAE, the equidistance contour lines are distorted in regions of high MF values. Thus, the mapping from latent to observation space learned by the VHP-FMVAE is approximately distance preserving. Additionally, we use the condition number and the normalised MF to evaluate  $\mathbf{G}$  based on 3,000 random samples. In contrast to the VHP-VAE, both the condition number and the normalised MF values of the VHP-FMVAE are close to one, which indicates that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ .

**Smoothness.** We randomly sample 100 pair points and



(a) VHP-FMVAE without *mixup*



(b) VHP-FMVAE without the identity term  $c^2\mathbb{1}$

**Figure 7.** Influence of the data augmentation and the identity term  $c^2\mathbb{1}$  on the learned latent representation of human movement data. The movements are coloured as in Fig. 3. (a) If not applying *mixup*, regions, where data is missing (e.g., between two movements), have a high MF and distorted equidistance contours. (b) regularising the metric tensor, and hence the Jacobian to be zero, does not allow the model to learn a *flat latent manifold*. The equidistance contours are scaled differently at various locations in the latent space. Without  $c^2\mathbb{1}$  term as in (Hadjeres et al., 2017), it cannot reduce the distance for points with high similarities. For instance, the walking is not squeezed as in Fig. 3a in the latent space. Therefore, the walking is not distributed smaller than jogging.

linearly interpolate between each pair. The second derivative of each trajectory is defined as the smoothness factor. Fig. 5 illustrates that the VHP-FMVAE significantly outperforms the VAE-VHP in terms of smoothness. Fig. 6 shows five examples of the interpolated trajectories.

**Verification of the distance metric.** To verify that the Euclidean distance in the latent space corresponds to the geodesic distance, we approximate the geodesic by using a graph-based approach (Chen et al., 2019). The graph of the baseline has 14,400 nodes, which are sampled in the latent space using a uniform distribution. Each node has 12 neighbours. In Fig. 3, five geodesics each are compared to the corresponding Euclidean interpolations. Tab. 1 shows the ratios of Euclidean distances in latent space to geodesics distances, as well as the related ratios in the observation space. To compute the ratios, we randomly sampled 100 pairs of points and interpolated between each pair. If the ratio of the distances is close to one, the Euclidean interpolation approximates the geodesic. The VHP-FMVAE outperforms

the VAE-VHP.

**Influence of the data augmentation and the identity term  $c^2\mathbb{1}$ .** Fig. 4 and Fig. 7a show the influence of the data augmentation (see Sec. 2.2). Without data augmentation, the influence of the regularisation term is limited to regions where data is available, as verified by the high MF values between the different movements. As an additional experiment, Fig. 4 and Fig. 7b illustrates the influence of the identity term  $c^2\mathbb{1}$ . If we remove it, the regularisation term becomes  $\|\mathbf{G}(g(\mathbf{z}_i, \mathbf{z}_j))\|_2^2$ . As a consequence, the model is not able to learn a *flat latent manifold*.

### 4.3. MNIST

The binarised MNIST data-set (Larochelle & Murray, 2011) consists of 50,000 training and 10,000 test images of handwritten digits (zero to nine) with  $28 \times 28$  pixels in size.

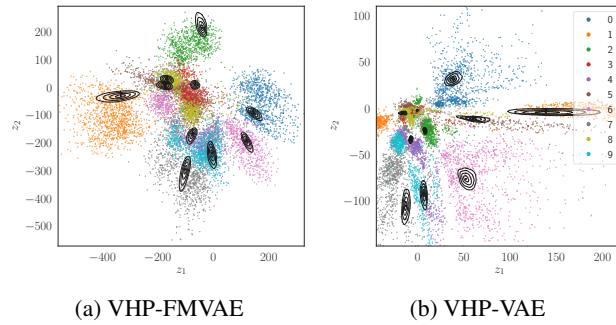


Figure 8. Latent representation of MNIST data: the contour plots illustrate curves of equal *observation-space distance* to the respective encoded data point (denoted by a black dot).

Both of our evaluation metrics the condition number and the normalised MF show that the VHP-FMVAE outperforms the VAE-VHP (see Fig. 8 and Fig. 9). In contrast to the

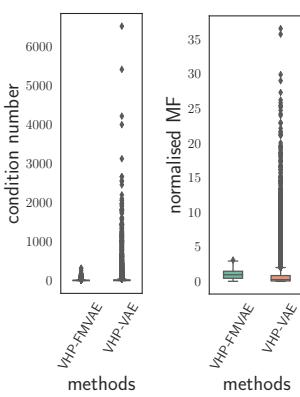


Figure 9. MNIST data: if both the condition number and the normalised MF values are close to one, it indicates that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ . The box-plots are based on 10,000 generated samples.

VHP-VAE, the VHP-FMVAE learns a latent space, where Euclidean distances are close to geodesic distances (see Tab. 1). This indicates that  $\mathbf{G}(\mathbf{z})$  is approximately constant.

### 4.4. MOT16 Object-Tracking Database

We evaluate our approach on the MOT16 object-tracking database (Milan et al., 2016), which is a large-scale person re-identification data-set, containing both static and dynamic scenes from diverse cameras.



Figure 10. Example identity switches between overlapping tracks. For vanilla SORT, track 3260 gets occluded and when subsequently visible, it gets assigned a new ID 3421. For deeSORT and VHP-VAE-SORT, the occluding track gets assigned the same ID as the track it occludes (42/61), and subsequently keeps this (erroneous) track. For VHP-FMVAE-SORT, the track 42 gets occluded, but is re-identified correctly when again visible.

Table 2. Comparisons between different descriptors for the purposes of object tracking and re-identification (Ristani et al., 2016). The bold and the red numbers denote the best results among all methods and among non-supervised methods, respectively.

METHOD	TYPE	IDF <sub>1</sub> ↑	IDP↑	IDR↑	RECALL↑	PRECISION↑	FAR↓	MT↑
VHP-FMVAE-SORT $\eta = 300$ (OURS)	UNSUPERVISED	63.7	77.0	54.3	65.0	92.3	<b>1.12</b>	158
VHP-FMVAE-SORT $\eta = 3000$ (OURS)	UNSUPERVISED	<b>64.2</b>	<b>77.6</b>	<b>54.8</b>	65.1	<b>92.3</b>	1.13	162
VHP-VAE-SORT	UNSUPERVISED	60.5	72.3	52.1	65.8	91.4	1.28	<b>170</b>
SORT	N.A.	57.0	67.4	49.4	<b>66.4</b>	90.6	1.44	158
DEEPSORT	SUPERVISED	<b>64.7</b>	76.9	<b>55.8</b>	<b>66.7</b>	91.9	1.22	<b>180</b>

METHOD	PT↓	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↑	MOTAL↑
VHP-FMVAE-SORT $\eta = 300$ (OURS)	269	90	<b>5950</b>	38592	<b>616</b>	<b>1143</b>	59.1	81.8	59.7
VHP-FMVAE-SORT $\eta = 3000$ (OURS)	<b>265</b>	90	6026	38515	598	1163	<b>59.1</b>	81.8	<b>59.7</b>
VHP-VAE-SORT	266	<b>81</b>	6820	37739	693	1264	59.0	81.6	59.6
SORT	275	84	7643	<b>37071</b>	1486	1515	58.2	<b>81.9</b>	59.5
DEEPSORT	<b>250</b>	87	6506	<b>36747</b>	<b>585</b>	1165	<b>60.3</b>	81.6	<b>60.8</b>

We compare with two baselines: SORT (Bewley et al., 2016) and DeepSORT (Wojke et al., 2017). SORT is a simple on-line and real-time tracking method, which uses bounding box intersection-over-union (IOU) for associating detections between frames and Kalman filters for the track predictions. It relies on good two-dimensional bounding box detections from a separate detector, and suffers from ID switching when tracks overlap in the image. DeepSORT extends the original SORT algorithm to integrate appearance information based on a deep appearance descriptor, which helps with re-identification in the case of such overlaps or missed detections. The deep appearance descriptor is trained using a *supervised* cosine metric learning approach (Wojke & Bewley, 2018). The candidate object locations of the pre-generated detections for both SORT, DeepSORT and our method are taken from (Yu et al., 2016). Further details regarding the implementation can be found in App. A.3.

We use the following metrics for evaluation. ↑ indicates that the higher the score is, the better the performance is. On the contrary, ↓ indicates that the lower the score is, the better the performance is.

- IDF<sub>1</sub>(↑): ID F<sub>1</sub> Score
- IDP(↑): ID Precision
- IDR(↑): ID Recall
- FAR(↓): False Alarm Ratio
- MT(↑): Mostly Tracked Trajectory
- PT(↓): Partially Tracked Trajectory
- ML(↓): Mostly Lost Trajectory
- FP(↓): False Positives
- FN(↓): False Negatives
- IDs(↓): Number of times an ID switches to a different previously tracked object
- FM(↓): Fragmentations
- MOTA(↑): Multi-object tracking accuracy
- MOTP(↑): Multi-object tracking precision
- MOTAL(↑): Log tracking accuracy

Tab. 2 shows that the performance of the proposed method is better than that of the model without Jacobian regularisation, and even close to the the performance of supervised learning. All methods depend on the same underlying detector for object candidates, and identical Kalman filter parameters. Compared to baseline SORT which does not utilise any appearance information, DeepSORT has 2.54 times, VHP-VAE-SORT has 2.14 times, VHP-FMVAE-SORT ( $\eta = 300$ ) has 2.41 times and VHP-FMVAE-SORT ( $\eta = 3000$ ) has 2.48 times fewer ID switches. Whilst the supervised DeepSORT descriptor has the least, using unsupervised VAEs with flat decoders has only 2.2% more switches, without the need for labels. Furthermore, by ensuring a quasi-Euclidean latent space, one can query nearest-neighbours efficiently via data-structures such as kDTrees. Fig. 10 shows an example of the results. In other examples of the videos, the VHP-FMVAE-SORT works similar as the DeepSORT. Videos of the results can be downloaded at: <http://tiny.cc/0s71cz>

## 5. Conclusion

In this paper, we have proposed a novel approach, which we call *flat manifold* variational auto-encoder. We have shown that this class of VAEs learns a latent representation, where the Euclidean metric is a proxy for the similarity between data points. This is realised by interpreting the latent space as a Riemannian manifold and by combining a powerful empirical Bayes prior with a regularisation method that constrains the Riemannian metric tensor to be a scaled identity matrix. Experiments on several datasets have shown the effectiveness of our proposed algorithm for measuring similarity. In case of the MOT16 object-tracking database, the performance of our unsupervised method nears that of state-of-the-art supervised approaches.

## Acknowledgements

We would like to thank Botond Cseke and Alexandros Paraschos for helpful discussions.

## References

- Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. *ICML*, 2018.
- Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.
- Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. In *ICLR*, 2018.
- Beckham, C., Honari, S., Lamb, A. M., Verma, V., Ghadiri, F., Hjelm, R. D., Bengio, Y., and Pal, C. On adversarial mixup resynthesis. *NeurIPS*, 2019.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. Simple online and realtime tracking. In *IEEE ICIP*, pp. 3464–3468, 2016.
- Bishop, C. M., Svens’ en, M., and Williams, C. K. Magnification factors for the SOM and GTM algorithms. In *Proceedings Workshop on Self-Organizing Maps*, 1997.
- Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. *CoNLL*, 2016.
- Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
- Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance weighted autoencoders. *CoRR*, abs/1509.00519, 2015.
- Chen, N., Bayer, J., Urban, S., and Van Der Smagt, P. Efficient movement representation by embedding dynamic movement primitives in deep autoencoders. In *IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 434–440, 2015.
- Chen, N., Karl, M., and van der Smagt, P. Dynamic movement primitives in latent space of time-dependent variational autoencoders. In *IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 629–636, 2016.
- Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., and van der Smagt, P. Metrics for deep generative models. In *AISTATS*, pp. 1540–1550, 2018a.
- Chen, N., Klushyn, A., Paraschos, A., Benbouzid, D., and van der Smagt, P. Active learning based on data uncertainty and model sensitivity. *IEEE/RSJ IROS*, 2018b.
- Chen, N., Ferroni, F., Klushyn, A., Paraschos, A., Bayer, J., and van der Smagt, P. Fast approximate geodesics for deep generative models. In *ICANN*, 2019.
- Frenzel, M. F., Teleaga, B., and Ushio, A. Latent space cartography: Generalised metric-inspired measures and measure-based transformations for generative models. *arXiv preprint arXiv:1902.02113*, 2019.
- Goldberger, J., Hinton, G. E., Roweis, S. T., and Salakhutdinov, R. R. Neighbourhood components analysis. In *Advances in neural information processing systems*, pp. 513–520, 2005.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.
- Grattarola, D., Zambon, D., Alippi, C., and Livi, L. Learning graph embeddings on constant-curvature manifolds for change detection in graph streams. *arXiv preprint arXiv:1805.06299*, 2018.
- Hadjeres, G., Nielsen, F., and Pachet, F. GLSR-VAE: geodesic latent space regularization for variational autoencoder architectures. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, 2017.
- 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. *ICLR*, 2017.
- Hoffer, E. and Ailon, N. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92. Springer, 2015.
- Jakubovitz, D. and Giryes, R. Improving DNN robustness to adversarial attacks using Jacobian regularization. In *ECCV*, pp. 514–529, 2018.
- Karaletsos, T., Belongie, S., and Rätsch, G. Bayesian representation learning with oracle constraints. *ICLR*, 2016.
- Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *ICLR*, 2014.
- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improving Variational Inference with Inverse Autoregressive Flow. *NIPS*, 2016.
- Klushyn, A., Chen, N., Kurle, R., Cseke, B., and van der Smagt, P. Learning hierarchical priors in VAEs. *NeurIPS*, 2019.

- Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37, 2011.
- Lee, J. M. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.
- Li, T. and Ding, C. The relationships among various non-negative matrix factorization methods for clustering. In *International Conference on Data Mining*, pp. 362–371. IEEE, 2006.
- Mathieu, E., Lan, C. L., Maddison, C. J., Tomioka, R., and Teh, Y. W. Hierarchical representations with Poincaré variational auto-encoders. *NeurIPS*, 2019.
- Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.
- Nie, W. and Patel, A. Towards a better understanding and regularization of GAN training dynamics. In *UAI*, 2019.
- Rezende, D. J. and Viola, F. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.
- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, volume 32, pp. 1278–1286, 2014.
- Rifai, S., Dauphin, Y. N., Vincent, P., Bengio, Y., and Muller, X. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pp. 2294–2302, 2011a.
- Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. Higher order contractive auto-encoder. In *ECML-PKDD*, pp. 645–660. Springer, 2011b.
- Ristani, E., Solera, F., Zou, R. S., Cucchiara, R., and Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. *CoRR*, abs/1609.01775, 2016.
- Schölkopf, B., Smola, A., and Müller, K.-R. Kernel principal component analysis. In *International conference on artificial neural networks*, pp. 583–588. Springer, 1997.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.
- Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. *NIPS*, 2016.
- Tomczak, J. M. and Welling, M. VAE with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223, 2018.
- Tosi, A., Hauberg, S., Vellido, A., and Lawrence, N. D. Metrics for probabilistic geometries. In *UAI*, pp. 800–808, 2014.
- Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Courville, A., Lopez-Paz, D., and Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. *ICML*, 2019.
- Wang, J. M., Fleet, D. J., and Hertzmann, A. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007.
- Weinberger, K. Q. and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.
- Wojke, N. and Bewley, A. Deep cosine metric learning for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 748–756, 2018.
- Wojke, N., Bewley, A., and Paulus, D. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, pp. 3645–3649, 2017.
- Xing, E. P., Jordan, M. I., Russell, S. J., and Ng, A. Y. Distance metric learning with application to clustering with side-information. In *NIPS*, pp. 521–528, 2003.
- Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., and Yan, J. POI: multiple object tracking with high performance detection and appearance feature. *CoRR*, abs/1610.06136, 2016.
- Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

## A. Appendix

### A.1. Additional Results on the Human Motion Dataset

### A.2. Influence of $c^2$

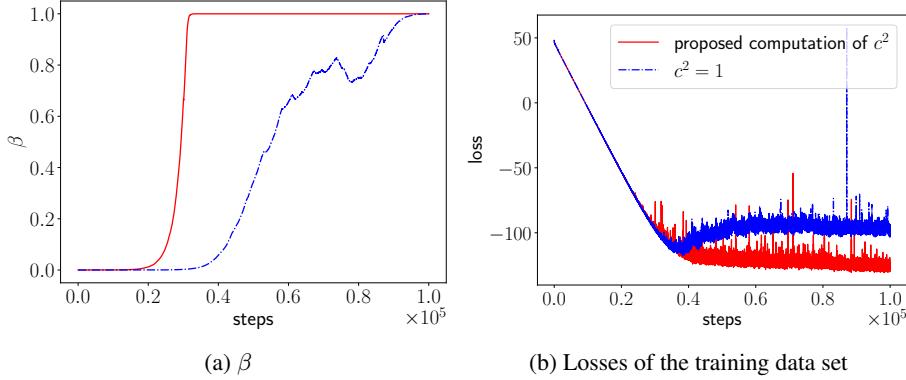


Figure 11. Comparison of different  $c^2$  using the human motion dataset. The model with the proposed computation of  $c^2$  converges faster than the model with  $c^2 = 1$ .

#### A.2.1. VECTOR FIELD

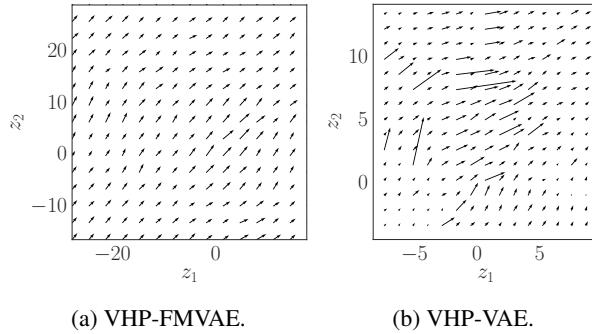


Figure 12. Vector field of the human motion dataset. The vector field is a vector of  $L_2$  norm over the output of Jacobian. The figures are corresponding to Fig. 3. The vector field of VHP-FMVAE is more regular than that of VAE-VHP.

#### A.2.2. RESULTS WITH A 5D LATENT SPACE

For the comparison of the geodesic in Sec. 4.2 (Tab. 1) and App. A.2.2 (Tab. 3), the singular regularisation hyper-parameter (see the definition in Eq. (17) of (Chen et al., 2018a)),  $\xi$ , of the graph-based geodesic is scaled by  $\xi_{\text{VHP-FMVAE}} = \frac{\text{mean}(s_i_{\text{VHP-FMVAE}}(\text{data}))^2 \cdot \xi_{\text{VHP-VAE}}}{\text{mean}(s_i_{\text{VHP-VAE}}(\text{data}))^2}$ .  $s_i$  denotes the singular of  $\mathbf{G}$ .  $s_i(\text{data})$  is computed with training data.

Table 3. Verification of the distance metric with a 5D latent space. The table shows the length ratio of the Euclidean interpolation to the geodesic. Additionally, we list the ratio of the related distances in the observation space. Note: the ratio also depends on the hyper-parameter of the graph-based approach,  $\xi$ . Given a pair of  $\{\xi_{\text{VHP-FMVAE}}, \xi_{\text{VHP-VAE}}\}$  as computed in App. A.2.2, the VHP-FMVAE outperforms the VHP-VAE.

DATA-SET	METHOD	OBSERVATION	LATENT
HUMAN	VHP-FMVAE	<b><math>1.03 \pm 0.16</math></b>	<b><math>0.59 \pm 0.11</math></b>
	VHP-VAE	$1.36 \pm 0.27$	$0.47 \pm 0.14$

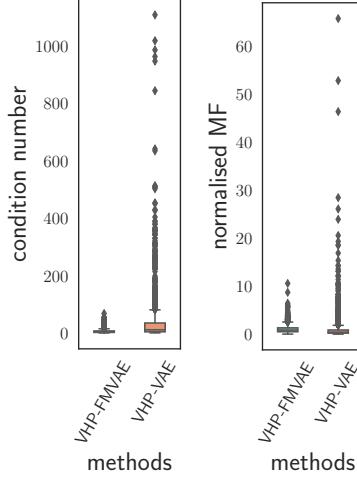


Figure 13. Human motion data with a 5D latent space: if both the condition number and the normalised MF values are close to one, it indicates that  $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$ . The box-plots are based on 3,000 generated samples.

### A.3. Implementation of VHP-FMVAE-SORT

SORT (Simple Object Real-time Tracking) uses 2D detections from a neural network and associates measurements of each frame to tracks that are initiated, kept, or removed over time. The IOU overlap is used as a distance function between a given track box and measurement box, and all boxes are optimally associated using the Hungarian algorithm. DeepSORT is an extension of SORT wherein a “deep” association metric is added. This is learnt using a large person re-identification dataset, training a network that outputs a fixed vector output per object. This vector contains appearance information. During online application, the vector is used with nearest neighbor queries to establish measurement-to-track associations, instead of just the IOU overlap used by the vanilla SORT. In our paper, we train variational auto-encoders and use the hidden latent space representation as a drop-in replacement to the fixed vector outputted by supervised network of DeepSORT, effectively only running the encoder during evaluation.

We evaluate the performance of our model by replacing the appearance descriptor from DeepSORT with the latent space embedding from the various auto-encoders used, using the same size of 128. The hyperparameters used were held constant: the minimum detection confidence of 0.3, NMS max overlap of 0.7, max cosine distance 0.2, max appearance budget 100. We tested a VHP-FMVAE, and our regularised VHP-FMVAE with  $\eta = 300$  and  $\eta = 3000$ .

#### A.4. Optimisation Process

Note: to be in line with previous literature (e.g. Higgins et al., 2017; Sønderby et al., 2016), we use the  $\beta$ -parametrisation of the Lagrange multiplier  $\beta = \frac{1}{\lambda}$  in our experiments.

As introduced in (Klushyn et al., 2019), we apply the following  $\beta$ -update scheme:

$$\beta_t = \beta_{t-1} \cdot \exp [\nu \cdot f_\beta(\beta_{t-1}, \hat{C}_t - \kappa^2; \tau) \cdot (\hat{C}_t - \kappa^2)], \quad (14)$$

where  $f_\beta$  is defined as

$$f_\beta(\beta, \delta; \tau) = (1 - H(\delta)) \cdot \tanh(\tau \cdot (\beta - 1)) - H(\delta). \quad (15)$$

$H$  is the Heaviside function and  $\tau$  a slope parameter.

---

**Algorithm 1** VHP-FMVAE

```

Initialise  $t = 1$ 
Initialise  $\beta \ll 1$ 
Initialise INITIALPHASE = TRUE
while training do
    Read current data batch  $\mathbf{x}_{ba}$ 
    Sample from variational posterior  $\mathbf{z}_{ba} \sim q_\phi(\mathbf{z}|\mathbf{x}_{ba})$ 
    Shuffle the samples from variational posterior  $\mathbf{z}'_{ba} = \text{shuffle}(\mathbf{z}_{ba})$ 
    Augment data  $\mathbf{z}^{\text{aug}}_{ba} = g(\mathbf{z}_{ba}, \mathbf{z}'_{ba})$ 
    Compute  $c^2 = \frac{1}{\text{batch.size}} \sum_i \frac{1}{N_z} [\text{tr}(\mathbf{G}(\mathbf{z}_i^{\text{aug}}))]$ 
    Compute  $\hat{C}_{ba}$  (batch average)
     $\hat{C}_t = (1 - \alpha) \cdot \hat{C}_{ba} + \alpha \cdot \hat{C}_{t-1}, (\hat{C}_0 = \hat{C}_{ba})$ 
    if  $\hat{C}_t < \kappa^2$  then
        INITIALPHASE = FALSE
    end if
    if INITIALPHASE then
        Optimise  $\mathcal{L}_{\text{VHP-FMVAE}}(\theta, \phi, \Theta, \Phi; \beta, \eta, c^2)$  w.r.t  $\theta, \phi$ 
    else
         $\beta \leftarrow \beta \cdot \exp [\nu \cdot f_\beta(\beta_{t-1}, \hat{C}_t - \kappa^2; \tau) \cdot (\hat{C}_t - \kappa^2)]$ 
        Optimise  $\mathcal{L}_{\text{VHP-FMVAE}}(\theta, \phi, \Theta, \Phi; \beta, \eta, c^2)$  w.r.t  $\theta, \phi, \Theta, \Phi$ 
    end if
     $t \leftarrow t + 1$ 
end while

```

---

## A.5. Model Architectures

Table 4. Model architectures. FC refers to fully-connected layers. Conv2D and Conv2DT denote tow-D convolution layer and transposed two-D convolution layer, respectively. See the definition of  $\nu$  in (Klushyn et al., 2019). We train each dataset on a single GPU.

DATASET	OPTIMISER	ARCHITECTURE	
PENDULUM	ADAM $1e-4$	INPUT LATENTS $q_\phi(\mathbf{z} \mathbf{x})$ $p_\theta(\mathbf{x} \mathbf{z})$ $q_\Phi(\zeta \mathbf{z})$ $p_\Theta(\mathbf{z} \zeta)$ OTHERS	$16 \times 16 \times 1$ 2 FC 256, 256. RELU ACTIVATION. FC 256, 256. RELU ACTIVATION. GAUSSIAN. FC 256, 256. RELU ACTIVATION. FC 256, 256. RELU ACTIVATION. $\kappa = 0.025, \nu = 1, K = 16, \eta = 1000.$
CMU HUMAN	ADAM $1e-4$	INPUT LATENTS $q_\phi(\mathbf{z} \mathbf{x})$ $p_\theta(\mathbf{x} \mathbf{z})$ $q_\Phi(\zeta \mathbf{z})$ $p_\Theta(\mathbf{z} \zeta)$ OTHERS	50 2, 5 FC 256, 256, 256, 256. RELU ACTIVATION. FC 256, 256, 256, 256. RELU ACTIVATION. GAUSSIAN. FC 256, 256, 256, 256, RELU ACTIVATION. FC 256, 256, 256, 256, RELU ACTIVATION. $\kappa = 0.03, \nu = 1, K = 32, \eta = 8000.$
MNIST	ADAM $1e-4$	INPUT LATENTS $q_\phi(\mathbf{z} \mathbf{x})$ $p_\theta(\mathbf{x} \mathbf{z})$ $q_\Phi(\zeta \mathbf{z})$ $p_\Theta(\mathbf{z} \zeta)$ OTHERS	$28 \times 28 \times 1$ 2 FC 256, 256, 256, 256. RELU ACTIVATION. FC 256, 256, 256, 256. RELU ACTIVATION. BERNOLLI. FC 256, 256, 256, 256. RELU ACTIVATION. FC 256, 256, 256, 256. RELU ACTIVATION. $\kappa = 0.245, \nu = 1, K = 16, \eta = 8000.$
MOT16	ADAM $3e-5$	INPUT LATENTS $q_\phi(\mathbf{z} \mathbf{x})$ $p_\theta(\mathbf{x} \mathbf{z})$ $q_\Phi(\zeta \mathbf{z})$ $p_\Theta(\mathbf{z} \zeta)$ OTHERS	$64 \times 64 \times 3$ 128 VGG16 (SIMONYAN & ZISSERMAN, 2015) Conv2DT+Conv2D 256, 128, 64, 32, 16. RELU ACTIVATION. GAUSSIAN. FC 512, 512. RELU ACTIVATION. FC 512, 512. RELU ACTIVATION. $\kappa = 0.8, \nu = 1, K = 8, \eta = 300 \text{ or } 3000.$