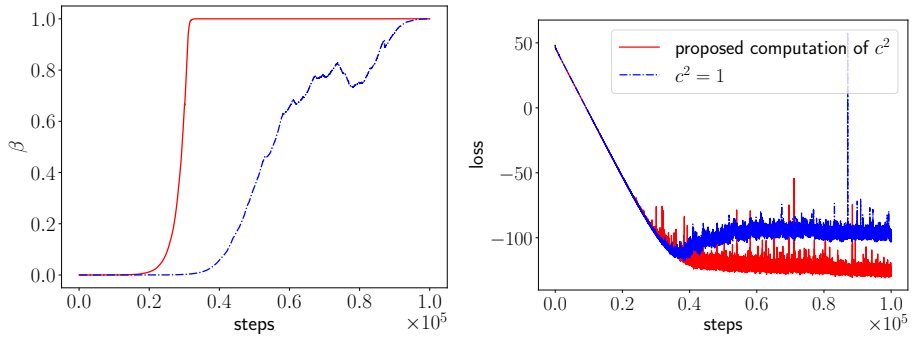
A. Appendix

A.1. Additional Results on the Human Motion Dataset

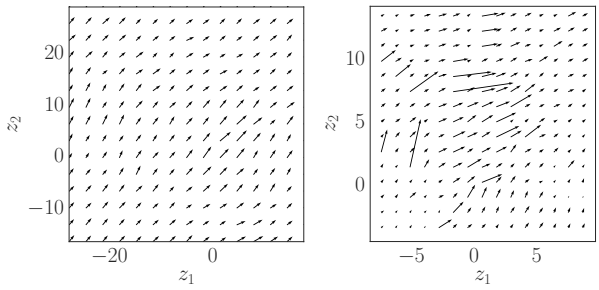
A.2. Influence of c^2



(a) β (b) Losses of the training data set

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



(a) VHP-FMVAE. (b) VHP-VAE.

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)), ξ, 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, ξ. 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	$\mathbf{1.03 \pm 0.16}$	$\mathbf{0.59 \pm 0.11}$
	VHP-VAE	1.36 ± 0.27	0.47 ± 0.14

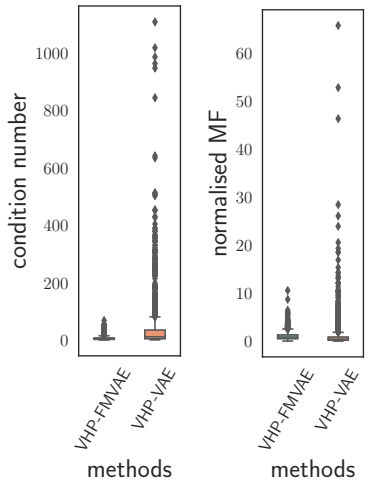


Figure 13. Human motion data with a 5D latent spac: if both the condition number and the normalised MF values are close to one, it indicates that $\mathbf{G}(\mathbf{z}) \propto \mathbb{1}$. The box-plots are based on 3,000 generated samples.

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 β-parametrisation of the Lagrange multiplier $\beta = \frac{1}{\lambda}$ in our experiments.

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

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

where f_β is defined as

$$f_\beta(\beta, \delta; \tau) = \big(1 - H(\delta)\big) \cdot \tanh\big(\tau \cdot (\beta - 1)\big) - H(\delta). \tag{15}$$

H is the Heaviside function and τ a slope parameter.

Algorithm 1 VHP-FMVAE

Initialise $t = 1$
Initialise $\beta \ll 1$
Initialise INITIALPHASE = TRUE
while training **do**
  Read current data batch $\mathbf{x}_{\text{ba}}$
  Sample from variational posterior $\mathbf{z}_{\text{ba}} \sim q_\phi(\mathbf{z}|\mathbf{x}_{\text{ba}})$
  Shuffle the samples from variational posterior $\mathbf{z}'_{\text{ba}} = \text{shuffle}(\mathbf{z}_{\text{ba}})$
  Augment data $\mathbf{z}^{\text{aug}}_{\text{ba}} = g(\mathbf{z}_{\text{ba}}, \mathbf{z}'_{\text{ba}})$
  Compute $c^2 = \frac{1}{\text{batch\_size}} \sum_i \frac{1}{N_{\mathbf{z}}} \left[\text{tr}(\mathbf{G}(\mathbf{z}^{\text{aug}}_i))\right]$
  Compute $\hat{\mathrm{C}}_{\text{ba}}$ (batch average)
  $\hat{\mathrm{C}}_t = (1 - \alpha) \cdot \hat{\mathrm{C}}_{\text{ba}} + \alpha \cdot \hat{\mathrm{C}}_{t-1}$, $(\hat{\mathrm{C}}_0 = \hat{\mathrm{C}}_{\text{ba}})$
  if $\hat{\mathrm{C}}_t < \kappa^2$ **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 θ, ϕ
  else
    $\beta \leftarrow \beta \cdot \exp\left[\nu \cdot f_\beta(\beta_{t-1}, \hat{\mathrm{C}}_t - \kappa^2; \tau) \cdot (\hat{\mathrm{C}}_t - \kappa^2)\right]$
    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 ν in (Klushyn et al., 2019). We train each dataset on a single GPU.

DATASET	OPTIMISER	ARCHITECTURE	
PENDULUM	ADAM	INPUT	16×16×1
	$1e$-4	LATENTS	2
		$q_\phi(\mathbf{z}\|\mathbf{x})$	FC 256, 256. RELU ACTIVATION.
		$p_\theta(\mathbf{x}\|\mathbf{z})$	FC 256, 256. RELU ACTIVATION. GAUSSIAN.
		$q_\Phi(\zeta\|\mathbf{z})$	FC 256, 256, RELU ACTIVATION.
		$p_\Theta(\mathbf{z}\|\zeta)$	FC 256, 256, RELU ACTIVATION.
		OTHERS	κ = 0.025, ν = 1, K = 16, $\eta = 1000$.
CMU HUMAN	ADAM	INPUT	50
	$1e$-4	LATENTS	2, 5
		$q_\phi(\mathbf{z}\|\mathbf{x})$	FC 256, 256, 256, 256. RELU ACTIVATION.
		$p_\theta(\mathbf{x}\|\mathbf{z})$	FC 256, 256, 256, 256. RELU ACTIVATION. GAUSSIAN.
		$q_\Phi(\zeta\|\mathbf{z})$	FC 256, 256, 256, 256, RELU ACTIVATION.
		$p_\Theta(\mathbf{z}\|\zeta)$	FC 256, 256, 256, 256, RELU ACTIVATION.
		OTHERS	κ = 0.03, ν = 1, K = 32, $\eta = 8000$.
MNIST	ADAM	INPUT	28×28×1
	$1e$-4	LATENTS	2
		$q_\phi(\mathbf{z}\|\mathbf{x})$	FC 256, 256, 256, 256. RELU ACTIVATION.
		$p_\theta(\mathbf{x}\|\mathbf{z})$	FC 256, 256, 256, 256. RELU ACTIVATION. BERNOULLI.
		$q_\Phi(\zeta\|\mathbf{z})$	FC 256, 256, 256, 256. RELU ACTIVATION.
		$p_\Theta(\mathbf{z}\|\zeta)$	FC 256, 256, 256, 256. RELU ACTIVATION.
		OTHERS	κ = 0.245 , ν = 1, K = 16, $\eta = 8000$.
MOT16	ADAM	INPUT	64×64×3
	$3e$-5	LATENTS	128
		$q_\phi(\mathbf{z}\|\mathbf{x})$	VGG16 (SIMONYAN & ZISSERMAN, 2015)
		$p_\theta(\mathbf{x}\|\mathbf{z})$	CONV2DT+CONV2D 256, 128, 64, 32, 16. RELU ACTIVATION. GAUSSIAN.
		$q_\Phi(\zeta\|\mathbf{z})$	FC 512, 512. RELU ACTIVATION.
		$p_\Theta(\mathbf{z}\|\zeta)$	FC 512, 512. RELU ACTIVATION.
		OTHERS	κ = 0.8 , ν = 1, K = 8, $\eta = 300$ or 3000.