Learning disentangled representations with the Wasserstein Autoencoder

Anonymous Author(s)

Abstract

Disentangled representation learning has undoubtedly benefited from objective function surgery. However, tuning the trade-off between reconstruction performances and disentanglement in the latent remains a delicate balancing act. Building on earlier successes of penalizing high total correlation in the latent variables, we propose TCWAE (Total Correlation Wasserstein Autoencoder). Working in the WAE paradigm enables the separation of the total-correlation term, thus providing disentanglement control over the learned representation, while offering more flexibility in the choice of the reconstruction cost. We propose two variants using different KL estimators and perform extensive quantitative comparisons on toy data sets. We also study the advantages of using different reconstruction costs on more difficult data sets.

1 Introduction

Learning representations of data is at the heart of deep learning; the ability to interpret those representations empowers practitioners to improve the performance and robustness of their models [4, 35]. In the case where the data is underpinned by independent latent generative factors, a good representation should encode information about the data in a semantically meaningful manner with statistically independent latent variables encoding for each factor. Bengio et al. [4] define a disentangled representation as having the property that a change in one dimension corresponds to a change in one factor of variation, while being relatively invariant to changes in other factors. While many attempts to formalize this concept have been proposed [14, 8, 7], finding a principled and reproducible approach to assess disentanglement is still an open problem [23].

Recent successful unsupervised learning methods have shown how simply modifying the ELBO objective, either re-weighting the latent regularization terms or directly regularizing the statistical dependencies in the latent, can be effective in learning disentangled representation. Higgins et al. [13] and Burgess et al. [5] control the information bottleneck capacity of Variational Autoencoders (VAEs, [19, 27]) by heavily penalizing the latent regularization term. Chen et al. [6] perform ELBO surgery to isolate the terms at the origin of disentanglement in β-VAE, improving the reconstruction-disentanglement trade-off. Esmaeili et al. [9] further improve the reconstruction capacity of β-TCVAE by introducing structural dependencies both between groups of variables and between variables within each group. Alternatively, directly regularizing the aggregated posterior to the prior with density-free divergences [39] or moments matching [20], or simply penalizing a high Total Correlation (TC, [38]) in the latent [17] has shown good disentanglement performances.

In fact, information theory has been a fertile ground to tackle representation learning. Achille and Soatto [1] re-interpret VAEs from an Information Bottleneck view [31], re-phrasing it as a trade-off between sufficiency and minimality of the representation, regularizing a pseudo TC between the aggregated posterior and the true conditional posterior. Similarly, Gao et al. [11] use the principle of total Correlation Explanation (CorEX) [36] and maximize the mutual information between the observation and a subset of anchor latent points. Maximizing the mutual information (MI) between the observation and the latent has been broadly used [34, 15, 3, 33], showing encouraging results in

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

representation learning. However, Tschannen et al. [33] argued that MI maximization alone cannot explain the disentanglement performances of these methods.

Building on the Optimal Transport (OT) problem [37], Tolstikhin et al. [32] introduced the Wasserstein Autoencoder (WAE), an alternative to VAE for learning generative models. Similarly to VAE, WAE maps the data into a (low-dimensional) latent space while regularizing the averaged encoding distribution. This is in contrast with VAEs where the posterior is regularized at each data point, and allows the encoding distribution to capture significant information about the data while still matching the prior when averaged over the whole data set. Interestingly, by directly regularizing the aggregated posterior, WAE hints at more explicit control on the way the information is encoded, and thus better disentanglement. The reconstruction term of the WAE allows for any cost function on the observation space, opening the door to better suited reconstruction terms, for example when working with continuous RGB data sets where the Euclidean distance or any metric on the observation space can result in more accurate reconstructions of the data.

In this work, following the success of regularizing the TC in disentanglement, we propose to use the Kullback-Leibler (KL) divergence as the latent regularization function in the WAE. We introduce the Total Correlation WAE (TCWAE) with an explicit dependency on the TC of the aggregated posterior. Using two different estimators for the KL terms, we perform extensive comparison with succesful methods on a number of data sets. Our results show that TCWAEs achieve competitive disentanglement performances while improving modelling performance by allowing flexibility in the choice of reconstruction cost.

2 Importance of Total correlation in disentanglement

2.1 Total correlation

The TC of a random vector $Z \in \mathcal{Z}$ under P is defined by

$$\mathbf{TC}(Z) \triangleq \sum_{d=1}^{d_Z} H_{p_d}(Z_d) - H_p(Z) \tag{1}$$

where $p_d(z_d)$ is the marginal density over only z_d and $H_p(Z) \triangleq -\mathbb{E}_p \log p(Z)$ is the Shannon differential entropy, which encodes the information contained in Z under P. Since

$$\sum_{d=1}^{d_Z} H_{p_d}(Z_d) \leq H_p(Z) \tag{2}$$

with equality when the marginals Z_d are mutually independent, the TC can be interpreted as the loss of information when assuming mutual independence of the Z_d; namely, it measures the mutual dependence of the marginals. Thus, in the context of disentanglement learning, we seek a low TC of the aggregated posterior, $p(z) = \int_{\mathcal{X}} p(z|x)\, p(x)\, dx$, which forces the model to encode the data into statistically independent latent codes. High MI between the data and the latent is then obtained when the posterior, $p(z|x)$, manages to capture relevant information from the data.

2.2 Total correlation in ELBO

We consider latent generative models $p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x|z)\, p(z)\, dz$ with prior $p(z)$ and decoder network, $p_\theta(x|z)$, parametrized by θ. VAEs approximate the intractable posterior $p(z|x)$ by introducing an encoding distribution (the encoder), $q_\phi(z|x)$, and learning simultaneously θ and ϕ when optimizing the variational lower bound, or ELBO, defined in Eq. (3):

$$\mathcal{L}_{ELBO}(\theta, \phi) \triangleq \mathop{\mathbb{E}}_{p_{\text{data}}(X)} \Big[\mathop{\mathbb{E}}_{q_\phi(Z|X)} [\log p_\theta(X|Z)] - \mathbf{KL}\Big(q_\phi(Z|X) \parallel p(Z)\Big) \Big] \leq \mathop{\mathbb{E}}_{p_{\text{data}}(X)} \log p_\theta(X) \tag{3}$$

Following [16], we can decompose the KL term in Eq. (3) as:

$$\frac{1}{N_{\text{batch}}} \sum_{n=1}^{N} \mathbf{KL}\Big(q_\phi(Z|x_n) \parallel p(Z)\Big) = \underbrace{\mathbf{KL}\Big(q(Z, N) \parallel q(Z)p(N)\Big)}_{\text{(i) index-code MI}} + \underbrace{\mathbf{KL}\Big(q(Z) \parallel p(Z)\Big)}_{\text{(ii) marginal KL}} \tag{4}$$

where $p(n) = \frac{1}{N}$, $q(z|n) = q(z|x_n)$, $q(z,n) = q(z|n)p(n)$ and $q(z) = \sum_{n=1}^{N} q(z|n)\, p(n)$. (i) refers to the *index-code mutual information* and represents the MI between the data and the latent under the join distribution $q(z,n)$, and (ii) to the *marginal KL* matching the aggregated posterior to the prior. While discussion on the impact of a high index-code MI on disentanglement learning is still open, the marginal KL term plays an important role in disentanglement. Indeed, it pushes the encoder network to match the prior when *averaged*, as opposed to matching the prior for each data point. Combined with a factorized prior $p(z) = \prod_d p_d(z_d)$, as it is often the case, the aggregated posterior is forced to factorize and align with the axis of the prior. More specifically, the marginal KL term in Eq. (4) can be decomposed the as sum of a TC term and a dimensionwise-KL term:

$$\mathbf{KL}\Big(q(Z) \parallel p(Z)\Big) = \mathbf{TC}\Big(q(Z)\Big) + \sum_{d=1}^{d_{\mathcal{Z}}} \mathbf{KL}\Big(q_d(Z_d) \parallel p_d(Z_d)\Big) \tag{5}$$

Thus maximizing the ELBO implicitly minimizes the TC of the aggregated posterior, enforcing the aggregated posterior to disentangle as Higgins et al. [13] and Burgess et al. [5] observed when strongly penalizing the KL term in Eq. (3). Chen et al. [6] leverage the KL decomposition in Eq. (5) by refining the heavy latent penalization to the TC only. However, the index-code MI term in Eq. (4) seems to have little to no role in disentanglement (see ablation study of Chen et al. [6]), potentially arming the reconstruction performances [16].

3 WAE naturally good at disentangling?

In this section we introduce the OT problem and the WAE objective, and discuss the compelling properties of WAEs for representation learning. Mirroring β-TCVAE decomposition, we derive the TCWAE objective.

3.1 WAE

The Kantorovich formulation of the OT between the true-but-unknown data distribution P_D and the model distribution P_θ, for a given cost function c, is defined by:

$$\mathrm{OT}_c(P_D, P_\theta) = \inf_{\Gamma \in \mathcal{P}(P_D, P_\theta)} \int_{\mathcal{X}\times\mathcal{X}} c(x,\tilde{x})\, \gamma(x,\tilde{x})\, dx d\tilde{x} \tag{6}$$

where $\mathcal{P}(P_D, P_\theta)$ is the space of all couplings of P_D and P_θ; namely, the space of joint distributions Γ on $\mathcal{X}\times\mathcal{X}$ whose densities γ have marginals p_D and p_θ. Tolstikhin et al. [32] derive the WAE objective by restraining this space and relaxing the hard constraint on the marginal using a soft constraint with a Lagrange multiplier (see Appendix A for more details):

$$W_{\mathcal{D},c}(\theta,\phi) \triangleq \mathop{\mathbb{E}}_{p_D(x)} \mathop{\mathbb{E}}_{q_\phi(z|x)} \mathop{\mathbb{E}}_{p_\theta(\tilde{x}|z)} c(x,\tilde{x}) + \lambda\, \mathcal{D}\Big(q(Z) \parallel p(Z)\Big) \tag{7}$$

where $\mathcal{D}$ is any divergence function and λ a relaxation parameter. The decoder, $p_\theta(\tilde{x}|z)$, and the encoder, $q_\phi(z|x)$, are optimized simultaneously by dropping the closed-form minimization over the encoder network, with standard stochastic gradient descent methods.

Similarly to the ELBO, objective (7) consists of a reconstruction cost term and a latent regularization term, preventing the latent codes to drift away from the prior. However, WAE explicitly penalizes the aggregate posterior. This motivates, following Section 2.2, the use of WAE in disentanglement learning. Rubenstein et al. [29] have shown promising disentanglement performances without modifying the objective (7). Another difference is the ground cost function in the reconstruction term. WAE allows for more flexibility in the reconstruction term with any cost function allowed, and in particular, it allows for cost functions better suited to the data at hand and for the use of deterministic decoder networks [32, 10].

3.2 TCWAE

In this section, for notation simplicity, we drop the explicit dependency of the distributions to their respective parameters.

Following Section 2.2 and Eq. (5), we chose the divergence function, $\mathcal{D}$, in Eq. (7), to be the KL divergence and assume a factorized prior (*e.g.* $p(z) = \mathcal{N}(0_{d_{\mathcal{Z}}}, \mathcal{I}_{d_{\mathcal{Z}}})$), obtaining the same

decomposition than in Eq. (5). Re-weighting each terms in Eq. (5) with hyper-parameters β and γ, and plugging into Eq. (7), we obtain our TCWAE objective:

$$W_{TC} \triangleq \mathop{\mathbb{E}}_{p(x_n)} \mathop{\mathbb{E}}_{q(z|x_n)} \Big[\mathop{\mathbb{E}}_{p(\tilde{x}_n|Z)} c(x_n, \tilde{x}_n) \Big] + \beta \mathbf{KL}\Big(q(Z) \parallel \prod_{d=1}^{d_\mathcal{Z}} q_d(Z_d)\Big) + \gamma \sum_{d=1}^{d_\mathcal{Z}} \mathbf{KL}\Big(q_d(Z_d) \parallel p_d(Z_d)\Big) \tag{8}$$

Given the positivity of the KL divergence, the TCWAE in Eq. (8) is an upper-bound of the WAE objective of Eq. (7) with $\lambda = \min(\beta, \gamma)$.

Eq. (8) can be directly related to the β-TCVAE objective of Chen et al. [6]:

$$-\mathcal{L}_{\beta-TC} \triangleq \mathop{\mathbb{E}}_{p(x_n)} \mathop{\mathbb{E}}_{q(z|x_n)} \Big[-\log p(x_n|Z) \Big] + \beta \mathbf{KL}\Big(q(Z) \parallel \prod_{d=1}^{d_\mathcal{Z}} q_d(Z_d)\Big) + \gamma \sum_{d=1}^{d_\mathcal{Z}} \mathbf{KL}\Big(q_d(Z_d) \parallel p_d(Z_d)\Big) + \alpha I_{\mathrm{q}}\Big(q(Z, N); q(Z)p(N)\Big) \tag{9}$$

As already mentioned, the main differences are the absence of index-code MI and a different reconstruction cost function $f(x_n, z)$. Setting $\alpha = 0$ in Eq. (9) makes the two latent regularizations match but breaks the inequality in Eq. (3). Matching the two reconstruction terms would be possible if we could find a ground cost function c such that $\mathbb{E}_{p(\tilde{x}_n|Z)} c(x_n, \tilde{x}_n) = -\log p(x_n|Z)$. In Section 4.1, we pseudo-match the reconstruction terms by choosing a deterministic decoder and the cross-entropy loss in TCWAEs. However, the cross-entropy is not a valid cost function and is only used for comparison purposes on toy-data sets.

3.3 Estimators

While being grounded and motivated by information theory and earlier works on disentanglement, using the KL as the latent divergence function, as opposed to other sampled-based divergences [32, 26], presents its own challenges. Indeed, the KL terms are intractable, and especially, we need estimators to approximate the entropy terms. We propose to use two estimators, one based on importance weight-sampling [6], the other on adversarial estimation using the denisty-ratio trick [17].

TCWAE-MWS

Chen et al. [6] propose to estimate the intractable terms $\mathbb{E}_q \log q(Z)$ and $\mathbb{E}_{q_d} \log q_d(Z)$ in the KL terms of Eq. (8) with Minibatch-Weighted Sampling (MWS). Considering a batch of observation $\{x_1, \dots x_{N_{\text{batch}}}\}$, they sample the latent codes $z_i \sim q(z|x_i)$ and compute:

$$\mathop{\mathbb{E}}_{q(z)} \log q(z) \approx \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \log \frac{1}{N \times N_{\text{batch}}} \sum_{j=1}^{N_{\text{batch}}} q(z_i|x_j) \tag{10}$$

This estimator, while being easily computed from samples, is a biased estimator of $\mathbb{E}_q \log q(Z)$. Chen et al. [6] also proposed an unbiased version, the Minibatch-Stratified Sampling (MSS). However, they found that it did not result in improved performances, and thus, as Chen et al. [6], we chose to use the simpler MWS estimator. We call the resulting algorithm the TCWAE-MWS. Other sampled-based estimators of the entropy or the KL divergence have been proposed [28, 9]. However, we choose the solution of Chen et al. [6] for 1) its simplicity and 2) the similarities between the TCWAE and β-TCVAE objectives.

TCWAE-GAN

A different approach, similar in spirit to the WAE-GAN originally proposed by Tolstikhin et al. [32], is based on adversarial-training. While Tolstikhin et al. [32] use the adversarial training to approximate the JS divergence, Kim and Mnih [17] use the density-ratio trick and adversarial training to estimate the intractable terms in Eq. (8). The the density-ratio trick [25, 30] estimates the KL divergence as:

$$\mathbf{KL}\Big(q(z) \parallel \prod_{d=1}^{d_\mathcal{Z}} q_d(z_d)\Big) \approx \mathop{\mathbb{E}}_{q(z)} \log \frac{D(z)}{1 - D(z)} \tag{11}$$

where D plays the same role than the discriminator in GANs and ouputs an estimate of the probability that z is sampled from $q(z)$ and not from $\prod_{d=1}^{d_z} q_d(z_d)$. Given that we can easily sample from $q(z)$, we can use Monte-Carlo sampling to estimate the expectation in Eq. (11). The discriminator D is adversarially trained alongside the decoder and encoder networks. We call this adversarial version the TCWAE-GAN.

4 Experiments

We perform a series of quantitative and qualitative experiments, starting with a quantitative comparison of the disentanglement performances of our methods with existing ones on toy data sets before moving to qualitative assessment of our method on more challenging data sets. We finally perform an ablation study on the impact of using different latent regularization functions in WAE. Details of the data sets, the experimental setup as well as the networks architectures are given in Appendix B.

4.1 Quantitative comparison with likelihood-based methods

We train our methods on the dSprites [24] and smallNORB [21] data sets whose ground-truth generative-factors are known and given in Table 2, Appendix B.1, and use three different disentanglement metrics to assess their performances: the Mutual Information Gap (MIG, [6]), the factorVAE metric [17] and the Separated Attribute Predictability score (SAP, [20]). For the implementation of these metrics, we follow Locatello et al. [23]. We compare our methods with β-TCVAE [6], FactorVAE [17] and the original WAE-MMD [32]. As a baseline comparison, we use the cross-entropy as the ground cost function. As mentioned in Section 3.2, while the cross-entropy loss does not make for a valid cost function, it allows for the comparison of the reconstruction error with likelihood-based methods. For the smallNORB data set, we also use the norm in L_1 and the square of the norm in L_2^2 ground cost functions.

We start by tuning γ, which regularizes the dimensionwise-KL, subsequently focusing on the role of the TC term in the disentanglement performances of the TCWAE. Figure 1 shows the heatmaps of the reconstruction error and the disentanglement scores of the TCWAEs on dSprites for $(\beta, \gamma) \in \{1, 5, 10, 50, 100, 150\}^2$. We observe that, while β indeed controls the trade-off between reconstruction and disentanglement, γ affects the best achievable trade-off when tuning β. For large γ ($\gamma \approx 100$ for TCWAE-MWS and $\gamma \approx 5$ for TCWAE-GAN), better disentanglement is obtained without much deterioration in reconstruction, and in particular, better than with $\gamma = 1$ which, for TCWAE-MWS, corresponds to β-TCVAE (with $\alpha = 0$). The same plots for smallNORB are given Figures 7, 8 and 9 in Appendix C. The chosen γ for the different methods are reported in Table 5 in Appendix C.

Tables 1a and 1b report the results, averaged over 5 random runs, for the different methods on dSprites and smallNORB. For each methods, we report the best β taken to be the one achieving an overall best ranking on the four different scores. TCWAEs achieve competitive performances across all the metrics with top scores in several metrics on both data sets. Figure 10 in Appendix C shows the violin plots of the different metrics for each β. We show Figure 2 the latent traversals of the different methods on the smallNORB data set with different reconstruction costs, with samples and reconstructions given Figure 11 in Appendix C. Visually, all WAE-based methods learned to disentangle, each capturing 5 different factors (4 ground-truth generative factors for smallNORB) while arguably generating crispier reconstructions than their VAE counterparts.

Finally, we plot the different disentanglement metrics versus the reconstruction error in Figure 3. The TCWAEs behave similarly to β-TCVAE, with a sweet-spot where the best disentanglement is achieved by trading some reconstruction performance when increasing β. As β continues to increase, the disentanglement performances start deteriorating. This is simply because the high reconstruction error prevents to distinguish any disentanglement in the reconstructed points.

4.2 Qualitative analysis: disentanglement on real-world data sets

We train our methods on 3Dchairs [2] and CelebA [22] whose generative factors are not known and qualitatively find that TCWAEs achieve good disentanglement. Figure 4 shows the latent traversals of four different factors learned by the models in the 3Dchairs data set. More specifically, in each sub-

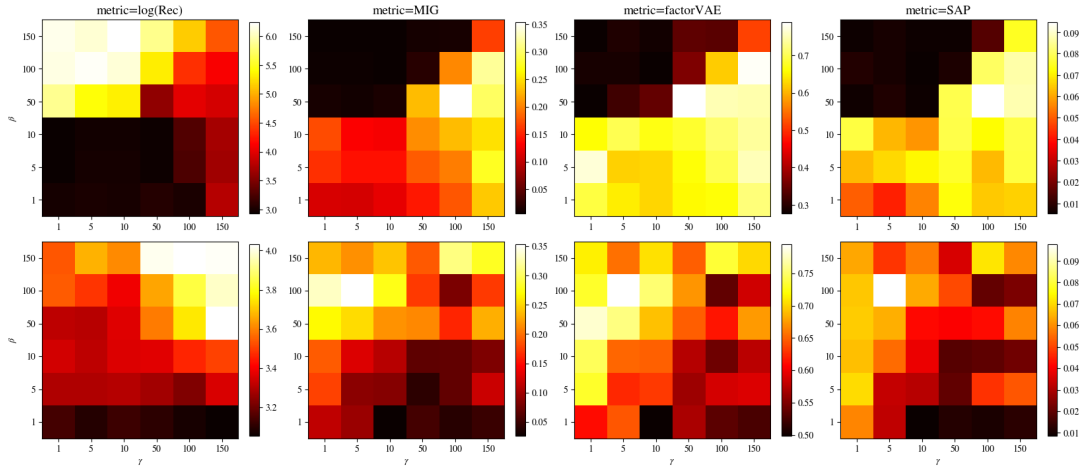


Figure 1: TCWAE-MWS (top) and TCWAE-GAN (bottom) trained on dSprites with the cross-entropy cost function. We use different color-scales for each pair of plots corresponding to the same metric.

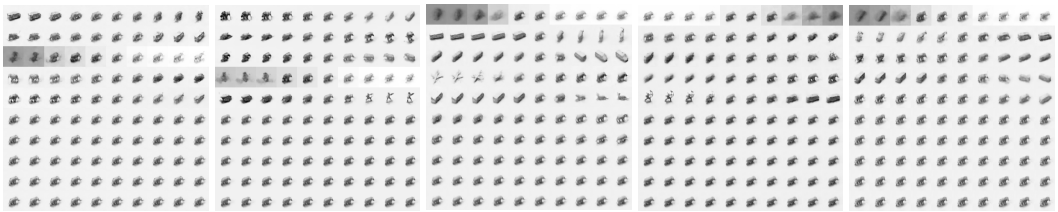
Table 1: Reconstruction and disentanglement scores ($\pm$ one standard deviation) for the different data sets.

Method	Rec	MIG	factorVAE	SAP
WAE ($\lambda = 200$)	$\mathbf{3.01 \pm .03}$	$0.026 \pm .006$	$0.52 \pm .06$	$0.010 \pm .004$
TCWAE MWS ($\beta = 50$, cross-entropy)	$4.16 \pm .31$	$\mathbf{0.36 \pm .04}$	$0.75 \pm .02$	$\mathbf{0.091 \pm .004}$
TCWAE GAN ($\beta = 100$, cross-entropy)	$3.57 \pm .19$	$0.34 \pm .09$	$\mathbf{0.78 \pm .09}$	$0.071 \pm .025$
Chen et al. [6] ($\beta = 8$)	$3.73 \pm .09$	$0.25 \pm .05$	$0.76 \pm .11$	$0.081 \pm .005$
Kim and Mnih [17] ($\gamma = 100$)	$3.47 \pm .13$	$0.26 \pm .09$	$0.70 \pm .05$	$0.070 \pm .030$

(a) dSprites

Method	Rec	MIG	factorVAE	SAP
WAE ($\lambda = 100$)	$7.58 \pm .00$	$0.031 \pm .007$	$0.44 \pm .04$	$0.015 \pm .003$
TCWAE-MWS ($\beta = 10$, cross-entropy)	$7.59 \pm .00$	$0.045 \pm .005$	$0.41 \pm .02$	$\mathbf{0.029 \pm .004}$
TCWAE-GAN ($\beta = 25$, cross-entropy)	$7.58 \pm .00$	$0.034 \pm .002$	$0.46 \pm .01$	$0.019 \pm .002$
Chen et al. [6] ($\beta = 4$)	$7.59 \pm .00$	$0.038 \pm .003$	$0.46 \pm .01$	$0.014 \pm .003$
Kim and Mnih [17] ($\gamma = 10$)	$7.58 \pm .00$	$0.039 \pm .003$	$0.46 \pm .01$	$0.019 \pm .003$
TCWAE-MWS ($\beta = 2$, L_1)	$4.28 \pm .01$	$0.038 \pm .005$	$\mathbf{0.47 \pm .06}$	$0.017 \pm .006$
TCWAE-GAN ($\beta = 4$, L_1)	$4.24 \pm .03$	$0.042 \pm .011$	$0.46 \pm .02$	$0.015 \pm .002$
TCWAE-MWS ($\beta = 4$, L_2^2)	$2.35 \pm .04$	$\mathbf{0.050 \pm .006}$	$0.43 \pm .03$	$0.021 \pm .005$
TCWAE-GAN ($\beta = 4$, L_2^2)	$2.05 \pm .03$	$0.030 \pm .012$	$0.42 \pm .08$	$0.020 \pm .003$

(b) smallNORB



WAE cross-entropy | TCWAE-MWS L_1 | TCWAE-GAN L_2^2 | β-TCVAE | FactorVAE

Figure 2: Latent traversals for each model on smallNORB. The parameters are the same than the ones reported in Table 1. Each row i corresponds to latent z_i and are order by increasing $\mathbf{KL}\Big(1/N_{test} \sum_{testset} q(z_i|x) \parallel p(z_i)\Big)$.

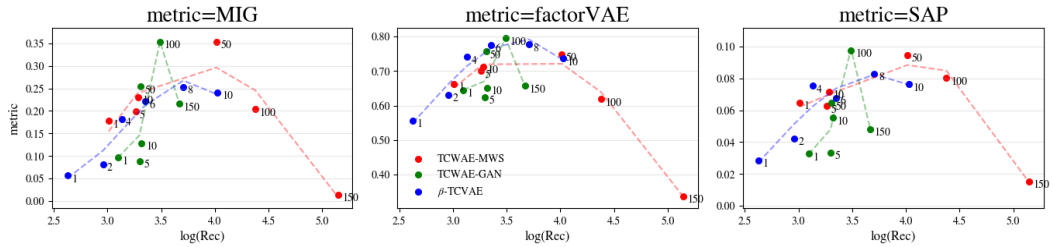


Figure 3: Scores versus reconstruction for TCWAE-MWS , TCWAE-GAN and β-TCVAE on dSprites. Annotations at each point are values of β. Points with low reconstruction error and high scores (top-left corner) represent better models.

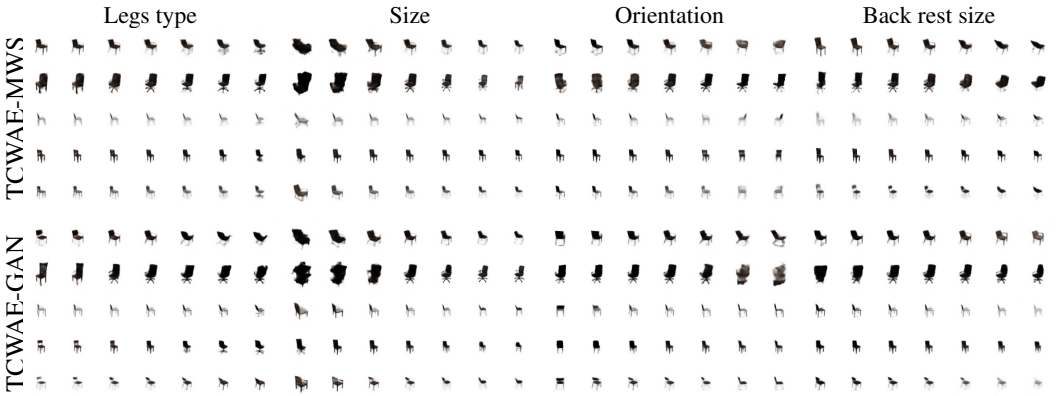


Figure 4: Latent traversals for TCWAE-MWS and TCWAE-GAN. Each line corresponds to one input data point. We vary evenly the encoded latent codes in the interval $[-2, 2]$.

plot, we encode five different observations (rows) and reconstruct the latent traversals (columns) when varying one latent dimension at a time. Figure 12 in Appendix D shows the models reconstructions and samples. Similarly, Figure 5 shows the latent traversals for different attributes discovered in the CelebA data set while the models reconstructions and samples are given Figure 14.

As previously argued, the flexibility in the construction of the reconstruction term of the TCWAE allows for the choice of reconstruction cost functions and the use of deterministic decoders. While comparing the reconstruction error with VAE-based methods is now impossible, we claim, after inspection of the latent traversals, reconstructions and samples, that it improves the reconstruction and sample generation capacities of our methods on more difficult data sets such as CelebA. We

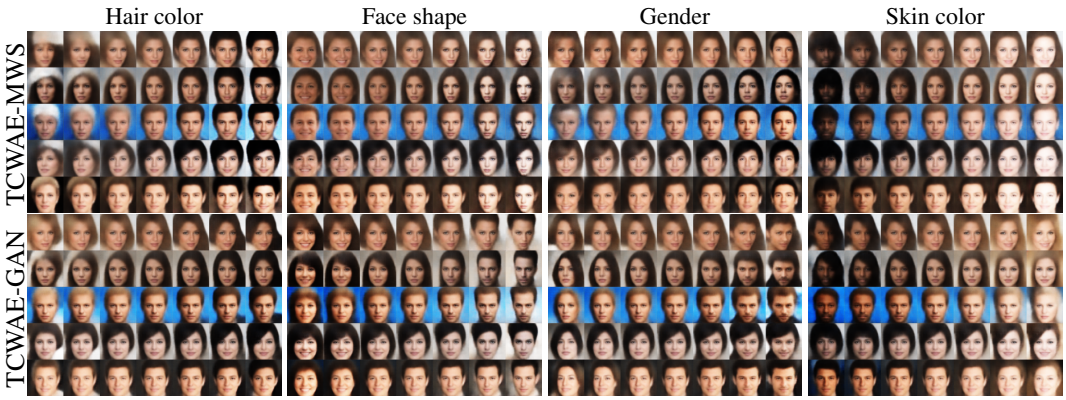
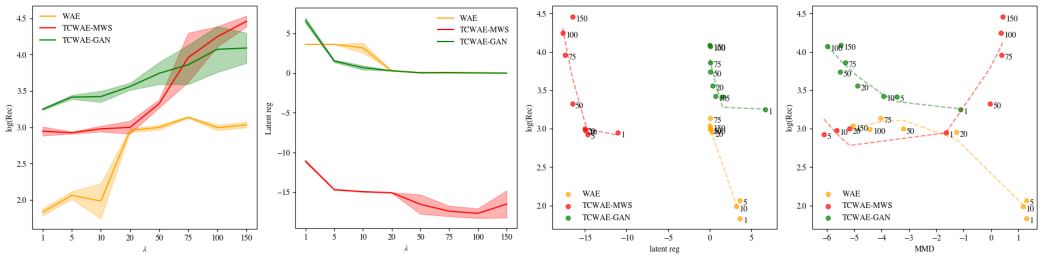


Figure 5: Same as Figure 4 but for CelebA. We vary evenly the encoded latent codes in the interval $[-6, 6]$.



(a) Log(Rec) (b) Latent reg (c) Log(Rec) *vs* latent reg (d) Log(Rec) *vs* MMD

Figure 6: Reconstruction and latent regularization terms as functions of λ for the dSprites data set. (a): log-reconstruction error. (b): latent regularization term (MMD for WAE, KL for TCWAE). (c): log-reconstruction error against latent regularization. (d): log-reconstruction error against the MMD (logarithmic x-axis). Shaded regions show $\pm$ one standard deviation.

acknowledge however than more thoughtful analysis and comparisons, using for example sample-based metrics such as the FID score [12], are needed in order to rigorously assess and compare the reconstruction and sample generation performances of the various methods. We felt that this type of comparison was out of scope of this work given the simple modelling assumptions (small networks and low latent dimensions). Finally, while the choice of reconstruction cost functions for the data at hand has a huge potential to improve the reconstruction performances, it still remains an open and difficult question.

4.3 Latent regularization ablation

Finally, we compare the impact of using different latent regularization functions in the WAE objective. We chose $\beta = \gamma$ in Eq. (8), and compare the resulting TCWAEs with the WAE-MMD [32] for $\beta \in \{1, 5, 10, 25, 50, 75, 100, 150\}$ on the dSprites data set. The reconstruction error and the latent regularizations are shown Figures 6a and 6b. Similar trade-off can be observed between the different methods with high reconstruction error and small latent regularization term for high β and conversely. Note the bias in the MWS estimator in Figure 6b as mentioned Section 3.3. Plotting directly the reconstruction error against the latent regularization in Figure 6c, we observe a domain of β ($\beta \leq 10$) for which the latent regularization term of the TCWAEs decreases significantly without penalizing much the reconstruction error. In the contrast, for the WAE, this trade-off seems log-linear in β. For a more rigorous comparison, we plot the reconstruction error versus the MMD in Figure 6d. While WAE seems to reconstruct slightly better, TCWAEs achieve better latent regularization (in a MMD sense) even if they are not train to minimize the MMD. Surprisingly in the TCWAE-MWS case, as β becomes larger (approximately $\beta \geq 10$), the MMD term starts increasing alongside the reconstruction error. One explanation can be found in the decomposition of the regularization term, with an increase of the dimensionwise-KL compensated by a larger decrease in TC, resulting in looser regularization of the posterior to the prior.

5 Conclusion

Leveraging the surgery of the KL regularization term of the ELBO objective, we design a new disentanglement method based on the WAE objective whose latent divergence function is taken to be the KL divergence between the aggregated posterior and the prior. The WAE framework naturally enables the latent regularization to depend explicitly on the TC of the aggregated posterior, quantity previously associated with disentanglement. Using two different estimators of the KL terms, we show that our methods achieve competitive disentanglement on toy data sets. Moreover, the flexibility in the choice of the reconstruction cost function offered by the WAE framework makes our method more compelling when working with more challenging data sets.

Broader impact

Representation learning plays an important role in machine learning, as it has the potential to impact many area of research, from reinforcement learning, to zero-shot and transfer learning in

downstream tasks. The importance of good representations goes beyond the theoretical world of machine learning as many real-world applications could benefit from improved representations. For example, disentangled representations should allow for more efficient knowledge transfers, resulting in both time and computational resources savings, better robustness would prove very useful in fraud and error detection, while interpretation of the learned representation is already the core of many commercial applications. We identified two situations where the use of representation learning could result in negative outcomes. First, a poor interpretation of the learned representation is, in our opinion, a real risk when using data representations, potentially leading to disastrous consequences if decisions were taken on poor or simply wrong interpretations. Secondly, and potentially with more dramatic consequences, the downstream use of the representations can raise ethical questions in itself as one could use these representations to discriminate based on some features such as gender or ethnicity.

References

[1] A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[2] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *CVPR*, 2014.

[3] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, 2019.

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

[5] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-VAE. *arXiv:804.03599*, 2018.

[6] R. T. K. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in VAEs. In *Advances in Neural Information Processing Systems*, 2018.

[7] K. Do and T. Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv:1908.09961*, 2019.

[8] C. Eastwood and C. K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[9] B. Esmaeili, H. B. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent. Structured disentangled representations. In *AISTATS*, 2018.

[10] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, 2015.

[11] S. Gao, R. Brekelmans, G. Ver Steeg, and A. Galstyan. Auto-encoding total correlation explanation. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[13] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[14] I. Higgins, D. Amos, D. Pfau, S. Racanière, L. Matthey, D. J. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv:1812.02230*, 2018.

[15] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[16] M. D. Hoffman and M. J. Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

[17] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

[18] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[19] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[20] A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

[21] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

[23] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

[24] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[25] X. Nguyen, M. J. Wainwright, and I. J. Michael. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, 2008.

[26] G. Patrini, M. Carioni, P. Forré, S. Bhargav, M. Welling, R. Van Den Berg, T. Genewein, and F. Nielsen. Sinkhorn autoencoders. *arXiv:1810.01118*, 2018.

[27] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[28] P. Rubenstein, O. Bousquet, J. Djolonga, C. Riquelme, and I. Tolstikhin. Practical and consistent estimation of f-divergences. In *Advances in Neural Information Processing Systems*, 2019.

[29] P. K. Rubenstein, B. Schoelkopf, and I. Tolstikhin. Learning disentangled representations with Wasserstein Auto-Encoders. In *ICLR Workshop*, 2018.

[30] M. Sugiyama, T. Suzuki, and T. Kanamori. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. In *Annals of the Institute of Statistical Mathematics*, 2011.

[31] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Annual Allerton Conference on Communication, Control and Computing*, 1999.

[32] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*, 2018.

[33] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

[34] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[35] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, 2019.

[36] G. Ver Steeg and A. Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, 2014.

[37] C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

[38] S. Watanabe. Information theoretical analysis of multivariate correlation. In *IBM Journal of Research and Development*, 1960.

[39] S. Zhao, J. Song, and S. Ermon. InfoVAE: Balancing learning and inference in variational autoencoders. In *AAAI Conference on Artificial Intelligence*, 2019.

A WAE derivation

We recall the Kantorovich formulation of the OT between the true-but-unknown data distribution P_D and the model distribution P_θ, with given cost function c:

$$\mathrm{OT}_c(P_D, P_\theta) = \inf_{\Gamma \in \mathcal{P}(P_D, P_\theta)} \int_{\mathcal{X}\times\mathcal{X}} c(x, \tilde{x})\, \gamma(x, \tilde{x})\, dx\, d\tilde{x} \tag{12}$$

where $\mathcal{P}(P_D, P_\theta)$ is the space of all couplings of P_D and P_θ:

$$\mathcal{P}(P_D, P_\theta) = \left\{ \Gamma \,\middle|\, \int_{\mathcal{X}} \gamma(x, \tilde{x})\, d\tilde{x} = p_D(x), \int_{\mathcal{X}} \gamma(x, \tilde{x})\, dx = p_\theta(\tilde{x}) \right\} \tag{13}$$

Tolstikhin et al. [32] first restrain the space of couplings to the joint distributions of the form:

$$\gamma(x, \tilde{x}) = \int_{\mathcal{Z}} p_\theta(\tilde{x}|z)\, q(z|x)\, p_D(x)\, dz \tag{14}$$

where $q(z|x)$, for $x \in \mathcal{X}$, plays the same role as the variational distribution in variational inference.

While the marginal constraint on x (first constraint in Eq. (13)) in Eq. (14) is satisfied by construction, the second marginal constraint (that over x giving p_θ in in Eq. (13)) is not guaranteed. A sufficient condition is to have for all $z \in \mathcal{Z}$:

$$\int_{\mathcal{X}} q(z|x)\, p_D(x)\, dx = p(z) \tag{15}$$

Secondly, Tolstikhin et al. [32] relax the constraint in Eq. (15) using a soft constraint with a Lagrange multiplier:

$$\widehat{W}_c(P_D, P_\theta) = \inf_{q(Z|X)} \left[\int_{\mathcal{X}\times\mathcal{X}} c(x, \tilde{x})\, \gamma(x, \tilde{x})\, dx\, d\tilde{x} + \lambda\, \mathcal{D}\Big(q(Z) \parallel p(Z)\Big) \right] \tag{16}$$

where $\mathcal{D}$ is any divergence function, λ a relaxation parameter, γ is defined in Eq. (14) and $q(Z)$ is the aggregated posterior as define in Section 2. Finally, they drop the closed-form minimization over the variational distribution $q(z|x)$, to obtain the WAE objective, as defined in Section 3.1:

$$\begin{aligned} W_{\mathcal{D},c}(\theta, \phi) &\triangleq \mathop{\mathbb{E}}_{p_D(X)} \mathop{\mathbb{E}}_{q_\phi(z|x)} \mathop{\mathbb{E}}_{p_\theta(\tilde{x}|z)} c(x, \tilde{x}) + \lambda\, \mathcal{D}\Big(q(Z) \parallel p(Z)\Big) \\ &\approx \mathop{\mathbb{E}}_{p(x_n)} \mathop{\mathbb{E}}_{q_\phi(z|x_n)} \mathop{\mathbb{E}}_{p_\theta(\tilde{x}_n|z)} c(x, \tilde{x}_n) + \lambda\, \mathcal{D}\Big(q(Z) \parallel p(Z)\Big) \end{aligned} \tag{17}$$

B Implementation details

B.1 Experimental setup

We trained and compare our methods on four different data sets, two with known ground-truth generative factors (see Table 2): dSprites [24] with 737,280 binary, $64\times64\times1$ images and smallNORB [21] with 48,600 greyscale, $64 \times 64 \times 1$ images; and two with unknown ground-truth generative factors: 3Dchairs [2] with 86,366 RGB, $64 \times 64 \times 3$ images and CelebA [22] with 202,599 RGB $64 \times 64 \times 3$ images.

Table 2: Ground-truth generative-factors of the dSprites and smallNORB data sets.

data set	Generative factors (number of different values)
dSprites	Shape (3), Orientation (40), Position X (32), Position Y (32)
smallNORB	categories (5), lightings (6), elevations (9), azimuths (18)

For the main experiments in Section 4.2, we use batch size of 64 and train the methods over 600,000 iterations. In Sections 4.1 and 4.3 we use a bigger batch size of 256 in order to reduce the bias of the MWS estimator (Chen et al. [6] however show that there is very little impact on the performance

Table 3: Networks architectures

Encoder	**Decoder**	**Discriminator**
Input: $64 \times 64 \times$ nchan.	Input: $d_{\mathcal{Z}}$	Input: $d_{\mathcal{Z}}$
4×4 conv. 32 ReLU stride 2	256 FC ReLU	1000 FC ReLU
4×4 conv. 32 ReLU stride 2	$4 \times 4 \times 64$ FC ReLU	1000 FC ReLU
4×4 conv. 64 ReLU stride 2	4×4 conv. 64 ReLU stride 2	1000 FC ReLU
4×4 conv. 64 ReLU stride 2	4×4 conv. 32 ReLU stride 2	1000 FC ReLU
256 FC Relu	4×4 conv. 32 ReLU stride 2	1000 FC ReLU
$2 \times d_{\mathcal{Z}}$ FC	4×4 conv. nchan. ReLU stride 2	1000 FC ReLU
		2 FC

Table 4: FactorVAE discriminator setup

Parameter	**Value**
Batch size	64
Learning rate	$1e^{-4}$ (Section 4.1) / $1e^{-5}$ (Section 4.2)
beta 1	0.5
beta 2	0.9
epsilon	1e-08

of the MWS when using smaller batch size). We train the methods over 300,000 iterations in Section 4.3. For all experiments, we use the Adam optimizer [18] with a learning rate of 0.0004, beta1 of 0.9, beta2 of 0.999 and epsilon of 0.0008. For all the data sets except CelebA, we use the same latent dimension $d_{\mathcal{Z}} = 10$. We take $d_{\mathcal{Z}} = 20$ in the case of CelebA. We use Gaussian encoders with diagonal covariance matrix in all the models and deterministic decoder networks when possible (WAE-based methods), Gaussian decoders with diagonal covariance matrix otherwise (VAE-based methods). Details of the architectures of the encoder and decoder networks are given Section B.2. We use a (positive) mixture of Inverse MultiQuadratic (IMQ) kernels and the associated reproductive Hilbert space to compute the MMD when it is needed (WAE and in the ablation study of Section (4.3)).

B.2 Models architectures

The Gaussian encoder networks, $q_\phi(z|x)$ and decoder network, $p_\theta(x|z)$, are parametrized by neural networks as follow:

$$p_\theta(x|z) = \begin{cases} \delta_{\boldsymbol{f_\theta}(\boldsymbol{z})} & \text{if WAE based method,} \\ \mathcal{N}\big(\boldsymbol{\mu_\theta}(z), \boldsymbol{\sigma^2_\theta}(z)\big) & \text{otherwise.} \end{cases}$$
$$q_\phi(z|x) = \mathcal{N}\big(\boldsymbol{\mu_\phi}(x), \boldsymbol{\sigma^2_\phi}(x)\big)$$

where $\boldsymbol{f_\theta}$, $\boldsymbol{\mu_\theta}$, $\boldsymbol{\sigma^2_\theta}$, $\boldsymbol{\mu_\phi}$ and $\boldsymbol{\sigma^2_\phi}$ are the outputs of convolutional neural networks whose architectures are given Table 3.

All the discriminator networks, D, are fully connected networks and share the same architecture given Table 3. The optimisation setup for the discriminator is given Table 4.

C Quantitative experiments

Hyper parameter tuning

Heat-maps for the different metrics on the smallNORB data set are given Figures 7, 8 and 9 while the chosen γ for each models in the experiments of Section 4.1 are given Table 5.

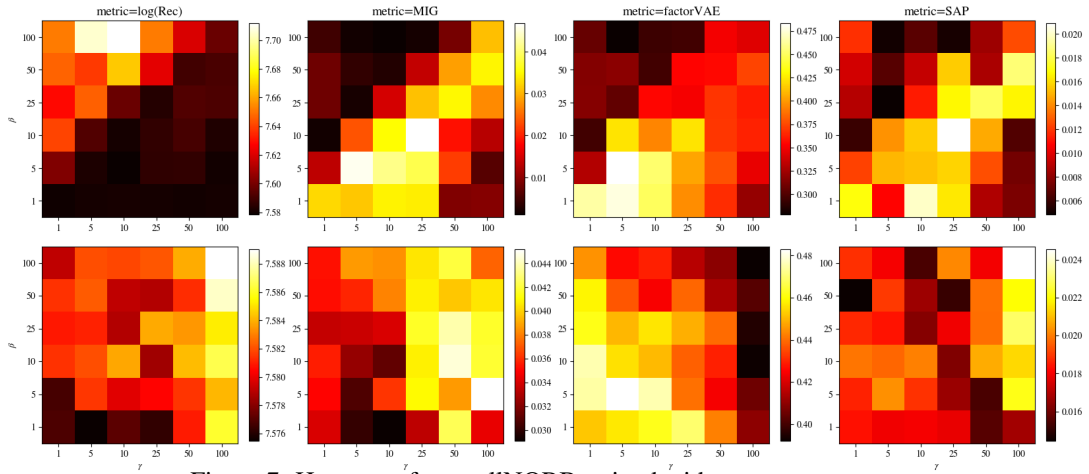


Figure 7: Heatmaps for smallNORB trained with cross-entropy.

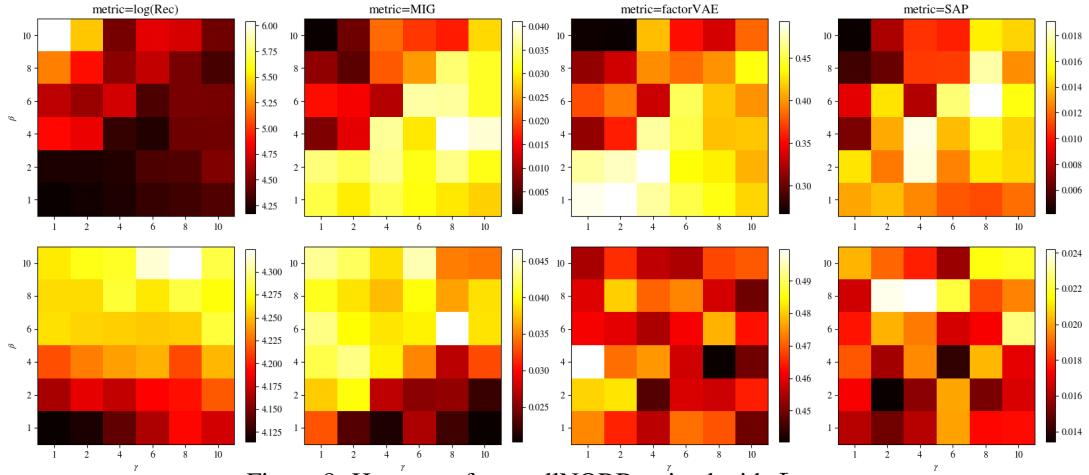


Figure 8: Heatmaps for smallNORB trained with L_1.

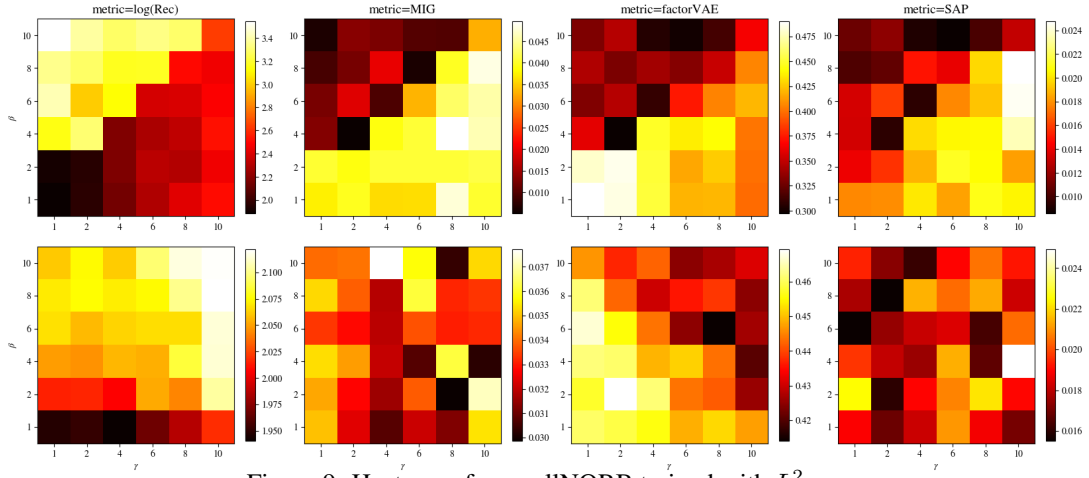
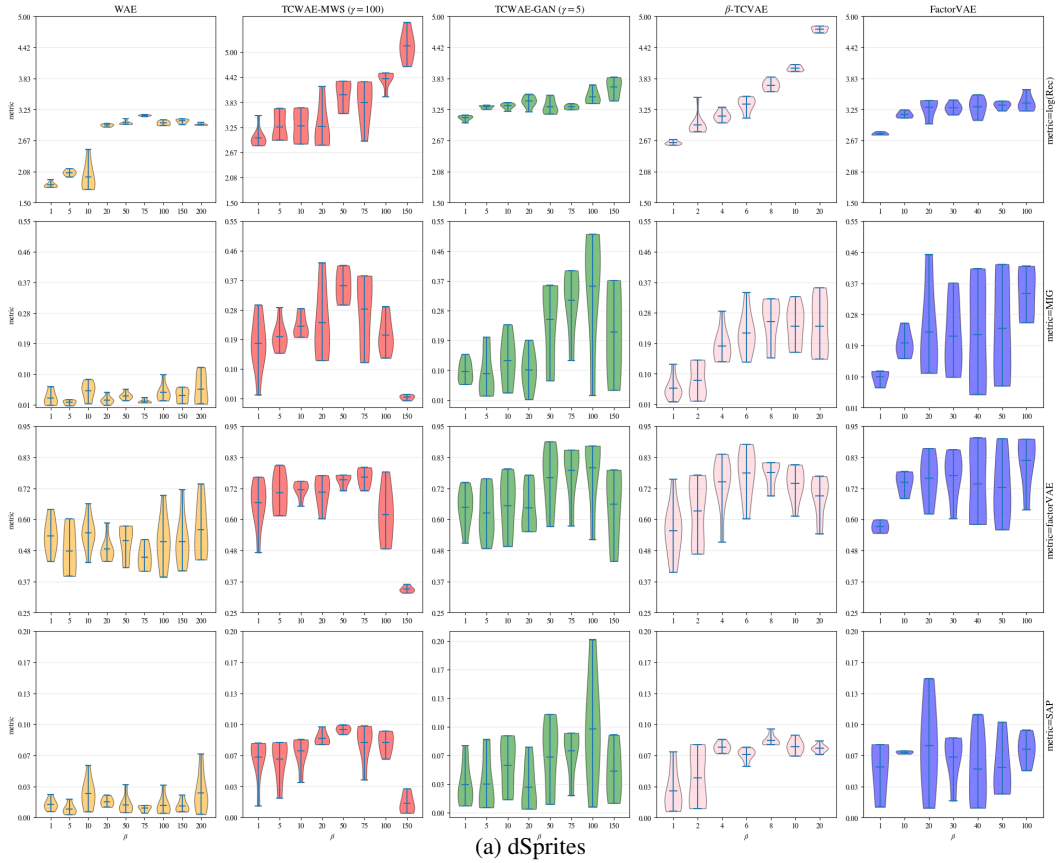


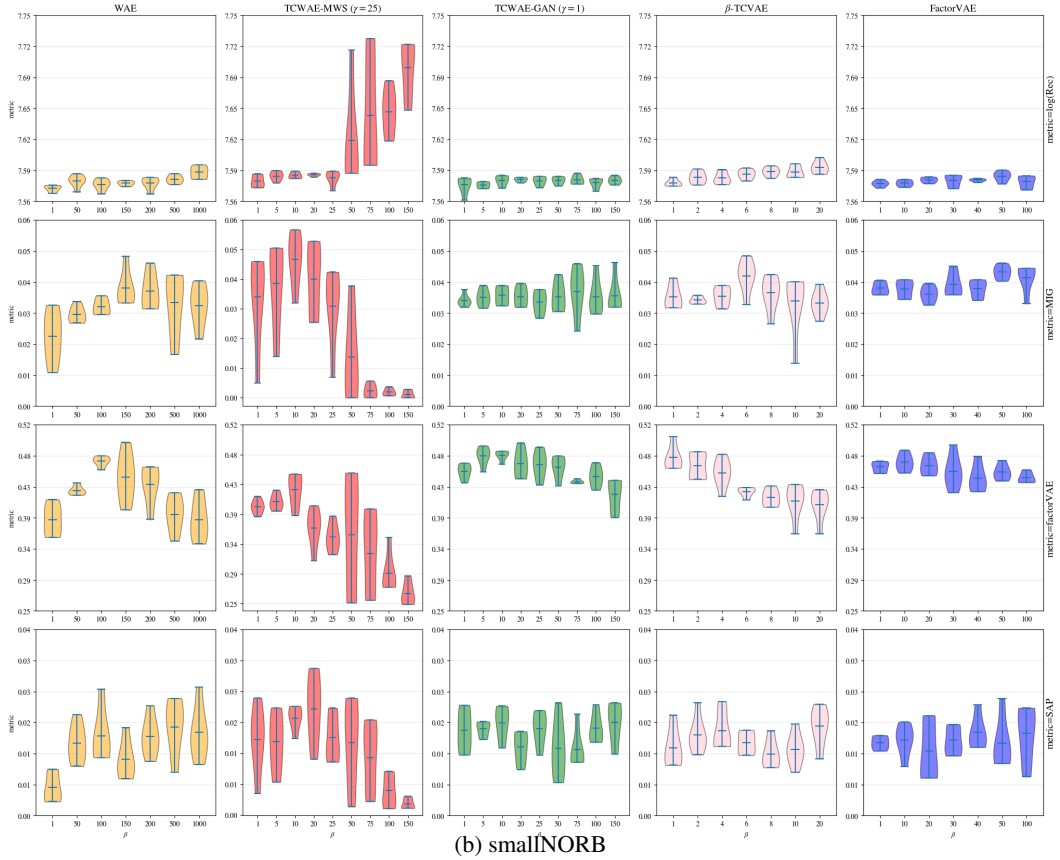
Figure 9: Heatmaps for smallNORB trained with L_2^2.

Disentanglement scores *vs* β

For each methods, we plot the distribution (over five random runs) of the different metrics for different β values Figure 10.



(a) dSprites



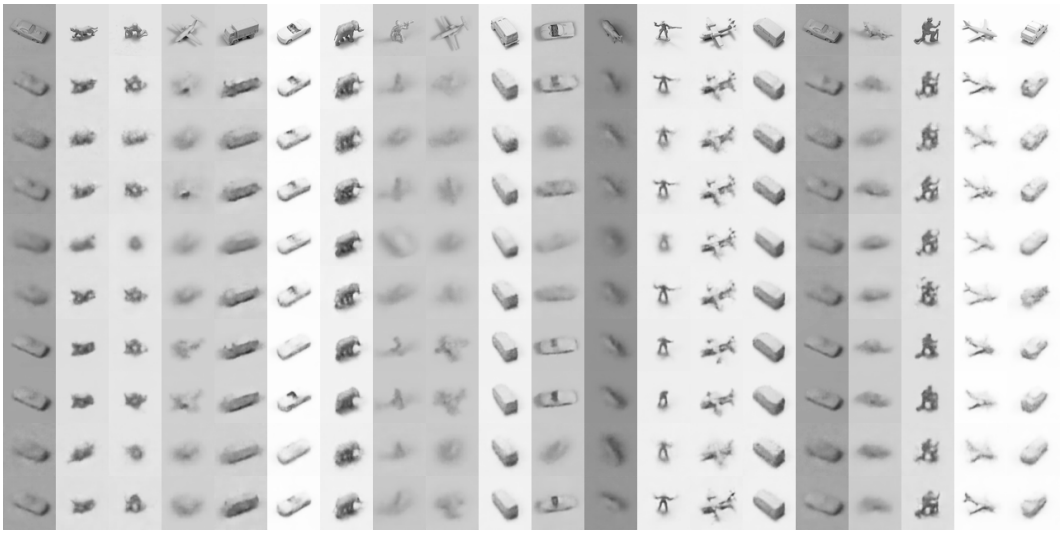
(b) smallNORB

Figure 10: Scores versus γ violin plots for dSprites and smallNORB. All models are trained with the cross entropy reconstruction los.

Table 5: γ values for methods for each data set.

Method	dSprites	smallNORB
TCWAE MWS cross-entropy	100	25
TCWAE MWS L_1	n/a	4
TCWAE MWS L_2^2	n/a	8
TCWAE GAN cross-entropy	5	1
TCWAE GAN L_1	n/a	1
TCWAE GAN L_2^2	n/a	1

Reconstructions, latent Traversals, samples



(a) Reconstructions

(b) Samples

Figure 11: Samples and reconstructions for each model. (a): Reconstructions. Top-row: input data, from second-to-top to bottom row: WAE, TCWAE-MWS (cross-entropy), TCWAE-GAN (cross-entropy), β-TCVAE, FactorVAE, TCWAE-MWS (l_1), TCWAE-GAN (l_1), TCWAE-MWS (l_2^2) and TCWAE-GAN (l_2^2). (b) Samples. Top-line: input data, from top to bottom row: WAE, TCWAE-MWS (cross-entropy), TCWAE-GAN (cross-entropy), β-TCVAE, FactorVAE, TCWAE-MWS (l_1), TCWAE-GAN (l_1), TCWAE-MWS (l_2^2) and TCWAE-GAN (l_2^2). Parameters are the ones reported in Tables 1 and 5

D Qualitative experiments

3Dchairs

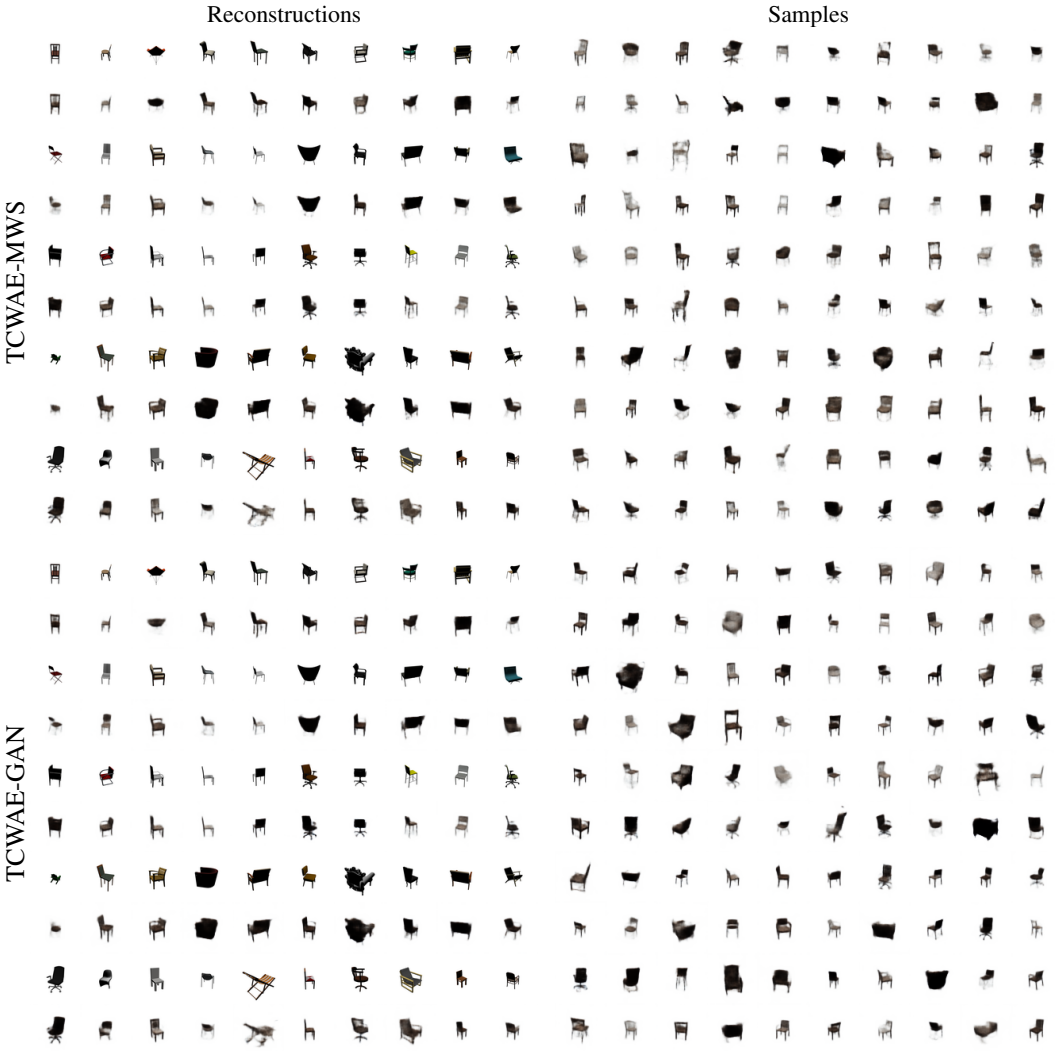


Figure 12: Reconstructions (left quadrants) and samples (right quadrants) for TCWAE-MWS (top quadrants) and TCWAE-GAN (bottom quadrants).

CelebA

Figure 13: Same as Figure 12 for the CelebA data set.

Figure 14: Same as Figure 14 for β-TCVAE (top quadrants) and FactorVAE (bottom quadrants).