IMPROVING THE UNSUPERVISED DISENTANGLED REPRESENTATION LEARNING WITH VAE ENSEMBLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Variational Autoencoder (VAE) based frameworks have achieved the state-of-the-art performance on the unsupervised disentangled representation learning. A recent theoretical analysis shows that such success is mainly due to the VAE implementation choices that encourage a PCA-like behavior locally on data samples. Despite this implied model identifiability, the VAE based disentanglement frameworks still face the trade-off between the local orthogonality and data reconstruction. As a result, models with the same architecture and hyperparameter setting can sometimes learn entangled representations. To address this challenge, we propose a simple yet effective VAE ensemble framework consisting of multiple VAEs. It is based on the assumption that entangled representations are unique in their own ways, and the disentangled representations are "alike" (similar up to a signed permutation transformation). In the proposed VAE ensemble, each model not only maintains its original objective, but also encodes to and decodes from other models through pair-wise linear transformations between the latent representations. We show both theoretically and experimentally, the VAE ensemble objective encourages the linear transformations connecting the VAEs to be trivial transformations, aligning the latent representations of different models to be "alike". We compare our approach with the state-of-the-art unsupervised disentangled representation learning approaches and show the improved performance.

1 INTRODUCTION

Disentangled representation learning aims to capture the semantically meaningful compositional representation of data (Higgins et al., 2018; Mathieu et al., 2018), and is shown to improve the efficiency and generalization of supervised learning (Locatello et al., 2019), reinforcement learning (Watters et al., 2019), and reasoning tasks (van Steenkiste et al., 2019). The current state-of-the-art unsupervised disentangled representation learning deploy the Variational Autoencoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014). The main challenge is to reduce the trade-off between learning a disentangled representation and reconstructing input data. Most of the recent works extend the original VAE objective with carefully designed augmented objective to address this trade-off (Higgins et al., 2017; Burgess et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2017). A recent study in (Locatello et al., 2018) compared these methods and showed that their performance is sensitive to initialization and hyperparameter setting of the augmented objective function.

Recently, Duan *et al.* (Duan et al., 2019) developed an unsupervised model selection method named Unsupervised Disentanglement Ranking (UDR) to address the challenge of hyperparameter search and model selection. UDR leverages the finding in (Rolinek et al., 2019) that the implementation choices of VAE encourage a local PCA-like behavior locally on data samples. As a result, disentangled representations by VAEs are "alike" as they are similar up to signed permutation transformations. On the contrary, the entangled representations by VAEs are "unique" as they are similar at least up to non-degenerate rotation matrices. UDR uses multiple models trained with different initializations and hyperparameter settings, and builds a similarity matrix measuring the pair-wise similarity between the latent variables from different models. A higher score is given to the model that can match its representations to many others models. The results show close match between UDR and commonly used supervised metrics, as well as the performance of downstream tasks using the latent representations.

Inspired by the findings from these studies, we propose a simple yet effective VAE ensemble framework to improve the disentangled representation by VAE. The proposed VAE ensemble consists of multiple VAEs. The latent variables in every pair of these models are connected through linear layers to force the latent representations in the ensemble to be similar up to a linear transformation. We show that the VAE ensemble objective encourages these pair-wise linear transformations to converge to trivial transformations, making latent representations of different VAEs in the ensemble to be "alike", thus disentangled. In this paper, we make the following contributions: (1) We introduce a simple yet effective VAE ensemble framework to improve the disentangled representation learning using the original VAE. (2) We show in theoretical analysis that the linear transformations connecting the latent representations of the individual models in the ensemble tend to converge to trivial transformations thus encourage disentangled representation, and verify this result with experiments. (3) We evaluate our approach using the original VAE model, and show the improved state-of-the-art performance across different datasets.

2 Related Work

Variatioanl Autoencoder is a deep directed probabilistic graphical model consisting of an encoder and a decoder (Kingma & Welling, 2013; Rezende et al., 2014). The encoder $q_\phi(z|x)$ maps the input data $x \in \mathbb{R}^n$ to a probabilistic distribution as the latent representation $z \in \mathbb{R}^d$, and the decoder $q_\theta(x|z)$ maps the latent representation to the data space noted as $q_\theta(x|z)$, where ϕ and θ represent model parameters. The VAE objective is to maximize the marginalized log-likelihood of data. Direct optimization of this objective is not tractable and it is approximated by the evidence lower bound (ELBO) as:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z|x)}[\log q_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \parallel p(z)), \tag{1}$$

In practice, the first term is estimated by reconstruction error. The second term is the Kullback-Leibler divergence between the posterior $q_\phi(z|x)$ and the prior $p(z)$ commonly chosen as an isotropic unit Gaussian $p(z) \sim \mathcal{N}(0, \mathbf{I})$.

Disentangled representation by VAE has achieved the state-of-the-art performance (Higgins et al., 2017; Burgess et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2017), despite the fact that the VAE objective only models the marginal distribution of the data instead of the desired joint distribution over data and latent variables. The reason for this success is the implementation choices of the VAE framework (Rolinek et al., 2019). In practice, the latent variables in VAE often work in "polarized" modes. The "passive" mode is defined by $\mu_j^2(x) \ll 1$ and $\sigma_j^2(x) \approx 1$, while the "active" mode is defined by $\sigma_j^2(x) \ll 1$. The "passive" latent variables closely approximate the prior and have little effect on the decoder. The "active" latent variables, on the other hand, are closely related to both the per sample KL loss and the decoder output. The "polarized regime" enables a reformulated VAE objective showing that VAEs optimize a trade-off between data reconstruction and orthogonality of the linear approximation of decoder Jacobian locally around a data sample. This PCA-like behavior near data points encourages an identifiable disentangled latent space by VAE. Furthermore, it was suggested that finding an appropriate "polarized regime" is dependent on the initialization and the hyperparameter tuning of the state-of-the-art approaches. In this study, we show that the proposed VAE ensemble aligns the "polarized regime" of individual VAE models towards the disentangled representation.

Model selection In practice, we often observe neural networks achieve similar performance with different internal representations when trained with the same hyperparameters (Raghu et al., 2017; Wang et al., 2018; Morcos et al., 2018). For the unsupervised disentanlged representation, as discussed in (Locatello et al., 2018; Duan et al., 2019), we often observe high variance in the performance from the model trained with the same architecture and hyperparameter setting. This poses a challenge for choosing the model in practice. Duan et al. (2019) proposed Unsupervised Disentanglement Ranking (UDR) to address this challenge. The extensive empirical evaluations on UDR using both the supervised metric measurement and the performance of downstream tasks validates its effectiveness. They also confirm that disentangled representations are "alike" and entangled representations are unique in their own ways. The proposed VAE ensemble leverages this finding.

Identifiable VAE Built on the recent breakthroughs in nonlinear Independent Component Analysis (ICA) literature (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019), Khemakhem *et al.* show that the identification of the true joint distribution over observed and latent variables is

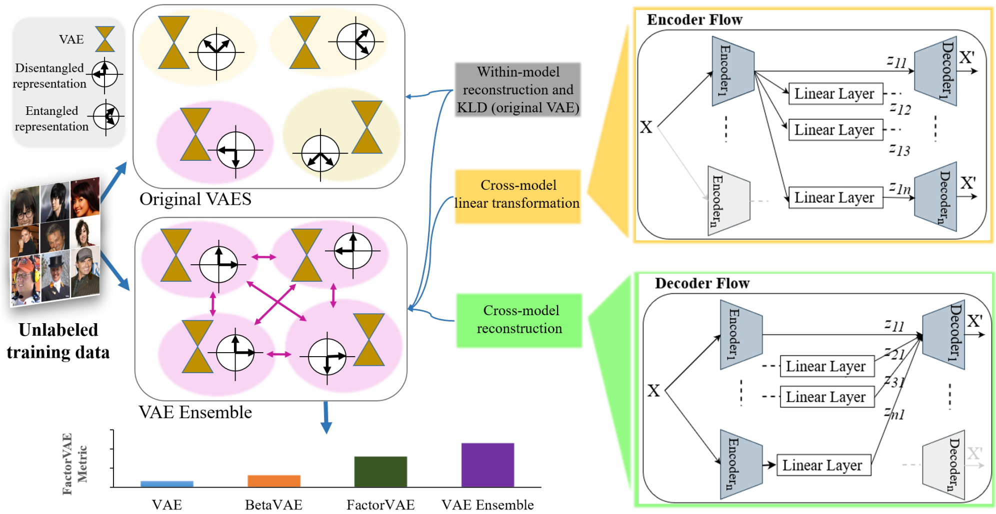


Figure 1: The proposed VAE ensemble consists of multiple original VAE models. The encoders of the VAEs in the ensemble generate input encoding that can be linearly transformed among each other. The decoders of the VAEs in the ensemble reconstruct the input data from both their corresponding encoder and the linearly transformed encodings from other encoders. The x and y axis of the circles on the left hand side of the plot represent two generative factors as an example. The aligned arrows with x and y axis show a model with disentangled representation and unaligned ones show a model with entangled representation.

possible up to very simple transformations (Khemakhem et al., 2019). They proposed identifiable VAE (iVAE) that requires a factorized prior distribution over the latent variables conditioned on an additionally observed variable, such as a class label or almost any other observation. We believe the proposed VAE ensemble is related to such framework where the latent representation from one VAE model can be regarded as the auxiliary observations for another.

Ensemble learning The idea of ensemble learning is to combine multiple learning models (potentially weak learners) to improve the task performance or robustness over a single model (Schapire, 1990). It achieves the improved performance by averaging the bias, reducing the variance thus preventing the over-fitting (Drucker et al., 1994; Breiman, 1996). Early works in neural networks have used the ensemble learning to achieve top performance in the related competition (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014). In this work, we apply the ensemble learning to enforce the alignment among the latent representations of different models. This results in latent representations that are similar among each other in the ensemble up to a trivial transformation.

3 The VAE Ensemble Framework

As illustrated in Figure 1, the proposed VAE ensemble consists of n original VAE models with the same architecture but different initializations. It also consists of $n \times (n-1)$ linear layers connecting the latent representations of every two VAE models. Each model in the ensemble maintains its original VAE objective as Eq. 1. In addition, l_2 loss is used to force mapping between latent representations via pair-wise linear layers (cross-model linear transformation). The decoder of each VAE model generates the input reconstruction from not only their corresponding encoder (within-model reconstruction), but also the linearly transformed encodings from other encoders (cross-model reconstruction). Overall, the VAE ensemble is trained with the following objective:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = &\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_{q_{\phi ij}(z_{ij}|x)}[\log q_{\theta j}(x|z_{ij})] - \sum_{i=1}^{n}\mathrm{KL}(q_{\phi ii}(z_{ii}|x) \parallel p(z_{ii})) \\ &- \gamma\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_{q_{\phi ij}(z_{ij}|x)}\left\|z_{jj} - z_{ij}\right\|^2,\end{aligned} \tag{2}$$

where n is the number of models in the ensemble; $\boldsymbol{\phi} := (\phi_{ij})$ is the encoders parameters where ϕ_{ij} represents the encoder of VAE$_i$ and its associated linear layer mapping the latent representation from VAE$_i$ to VAE$_j$ (notice that ϕ_{ii} represents the encoder parameters of VAE$_i$ only and its associated linear transformation can be be regarded as an identity transformation); $\boldsymbol{\theta} := (\theta_i)$ represents the decoder parameters of VAE$_i$; and z_{jj} represents the latent representation of VAE$_j$ while z_{ij} represents the linearly transformed latent representation from VAE$_i$ to VAE$_j$. γ is a hyperparameter to balance the effect of the estimation error between the latent representaitons. $p(z_{ii}) \sim \mathcal{N}(0, I)$ is assume to be the prior as defined in the original VAE objective.

Comparing to the original VAE objective in Eq. 1, the objective of each individual VAE model in the ensemble, $\mathcal{L'}_{VAE}$, can be written as:

$$\mathcal{L'}_{VAE}(\theta, \phi) = \mathcal{L}_{VAE}(\theta, \phi) + \underbrace{\sum_{j=1}^{n-1} \Big\{ \mathbb{E}_{q_{\phi j}(z_j|x)}[\log q_\theta(x|z_j)] - \gamma \mathbb{E}_{q_{\phi j}(z_j|x)} \left\| z_j - z \right\|^2 \Big\}}_{Ensemble\ Regularization}, \tag{3}$$

where n is the number of models in the ensemble, ϕ_j stands for the parameters of the encoder and its linear transformation layers from other VAEs, and z_j represents the linear transformed latent representation of the encoding from other encoders.

In this form, the VAE ensemble regularizes each VAE model with additional terms on the encoder as $\gamma \mathbb{E}_{q_{\phi j}(z_j|x)} \left\| z_j - z \right\|^2$, and on the decoder as $\sum_{j=1}^{n} \mathbb{E}_{q_{\phi j}(z_j|x)}[\log q_\theta(x|z_j)]$. These regularizations directly constrain the latent representations among different VAE models in the ensemble to be similar. In particular, for a given input data, $\left\| z_j - z \right\|^2$ encourages the encoders to generate similar encodings up to the linear transformations; and $\sum_{j=1}^{n} \mathbb{E}_{q_{\phi j}(z_j|x)}[\log q_\theta(x|z_j)]$ emphasizes the similar effect on the data reconstruction from the latent variables such that the decoders can reconstruct the input data with both the original encoding z and the linearly transformed encoding z_j. As we shall discuss in the next section, together these regularizations encourage similar latent representation up a trivial transformation by different models in the ensemble.

We also introduce the hyperparameter γ to balance the trade-off between these two regularizations: higher value forces closer mapping between the encoders and reduce the cross-model reconstruction error of the decoders; lower value relaxes the mapping between the encoders and increases the cross-model reconstruction error of the decoders. As we show in Section 5, both components are important and balancing the trade-off between them is important as the ensemble size increases.

Computational complexity It is a common practice to train a number of seeds per hyperparameter setting for the current state-of-the-art unsupervised disentanglement VAE models (Locatello et al., 2018; Duan et al., 2019). Comparing to training n original VAEs, the proposed VAE ensemble requires additional $n \times (n-1)$ linear layers. While this addition does not increase the size of the model much, the estimation of the linear transformations loss and the cross-model reconstruction losses grow with $n \times (n-1)$, which may be computationally expensive especially when n is large. That being said, the results in Section 5 show that the VAE ensemble achieves more stable results comparing to the current state-of-the-art models. Also, its computation is highly parallelisable.

4 Theoretical Justification

In this section, we present the theoretical analysis on why the proposed VAE ensemble can improve the disentangled representation. We start with analysing the l_2 objective in Eq. 2 of the pair-wise liner transformations in the VAE ensemble, and show that: (1) the pair-wise linear transformations encourage similar "polarized" regime (see Sec. 2) among the VAEs in the ensemble; (2) the linear transformations are close to the orthogonal transformations. Based on these two properties, we then discuss how the cross-model reconstructions by the VAE ensemble encourage learning a disentangled representation over its entangled counterpart.

4.1 The Effect of Linear Transformation between Latent Representations

Let VAE$_i$ and VAE$_j$ be two different VAE models in the ensemble, and M_{ji} be the linear transformation that maps the latent representation of a given input x by VAE$_j$ to the one by VAE$_i$, as

$\mathbf{z_j}(x) \sim \mathcal{N}(\boldsymbol{\mu}_j(x), diag(\boldsymbol{\sigma}_j(x)^2))$ to $\mathbf{z_i}(x) \sim \mathcal{N}(\boldsymbol{\mu}_i(x), diag(\boldsymbol{\sigma}_i(x)^2))$. In the following we remove the input notation from the VAE latent representations for simplicity (i.e. $\mathbf{z_j}(x)$ is simplified as $\mathbf{z_j}$), while keeping in mind that the analysis is based on the local latent representation of a given input x.

For VAE$_j$, the l_2 term of the VAE ensemble loss in Eq. 2 aims to find M_{ji} and $\mathbf{z_j}$ that minimize $\mathbb{E}\left\|\mathbf{z_i} - M_{ji}\mathbf{z_j}\right\|^2$, where the expectation is over the stochasticity of VAE$_j$. We can write $\mathbf{z_i}$ and $\mathbf{z_j}$ as $\mathbf{z_i} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i$ and $\mathbf{z_j} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_j$, where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, diag(\boldsymbol{\sigma}_i^2))$ and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, diag(\boldsymbol{\sigma}_j^2))$. Hence using bias-variance decomposition, the l_2 term can be written as:

$$
\begin{aligned}
&\min_{M_{ji}, \mathbf{z_j} \sim \mathcal{N}(\boldsymbol{\mu}_j, diag(\boldsymbol{\sigma}_j^2))} \mathbb{E}\left\|\mathbf{z_i} - M_{ji}\mathbf{z_j}\right\|^2 \\
&= \min_{M_{ji}, \mathbf{z_j} \sim \mathcal{N}(\boldsymbol{\mu}_j, diag(\boldsymbol{\sigma}_j^2))} \left\{ \left\|\boldsymbol{\mu}_i - M_{ji}\boldsymbol{\mu}_j\right\|^2 + \mathop{\mathbb{E}}_{\mathbf{z_j}} \left\|M_{ji}\boldsymbol{\mu}_j - M_{ji}\mathbf{z_j}\right\|^2 + \mathop{\mathbb{E}}_{\mathbf{z_i}} \left\|\boldsymbol{\mu}_i - \mathbf{z_i}\right\|^2 \right\} \\
&= \min_{M_{ji}, \mathbf{z_j} \sim \mathcal{N}(\boldsymbol{\mu}_j, diag(\boldsymbol{\sigma}_j^2))} \left\{ \left\|\boldsymbol{\mu}_i - M_{ji}\boldsymbol{\mu}_j\right\|^2 + \mathop{\mathbb{E}}_{\mathbf{z_j}} \left\|M_{ji}\boldsymbol{\mu}_j - M_{ji}\mathbf{z_j}\right\|^2 \right\} + C_1 \\
&= \min_{M_{ji}, \mathbf{z_j} \sim \mathcal{N}(\boldsymbol{\mu}_j, diag(\boldsymbol{\sigma}_j^2))} \left\{ \left\|\boldsymbol{\mu}_i - M_{ji}\boldsymbol{\mu}_j\right\|^2 + \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_j} \left\|M_{ji}\boldsymbol{\epsilon}_j\right\|^2 \right\} + C_1
\end{aligned}
\tag{4}
$$

where the constant C_1 arises from the fact $\mathop{\mathbb{E}}_{\mathbf{z_i}} \left\|\boldsymbol{\mu}_i - \mathbf{z_i}\right\|^2$ does not depend on M_{ji} and $\mathbf{z_j}$. Eq. 4 consists of a deterministic component of $\left\|\boldsymbol{\mu}_i - M_{ji}\boldsymbol{\mu}_j\right\|^2$ and a stochastic component of $\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_j} \left\|M_{ji}\boldsymbol{\epsilon}_j\right\|^2$. The deterministic component can be minimized by adjusting the parameters in VAE$_j$ such that its mean encoding $\boldsymbol{\mu}_j$ is optimized for any given M_{ji}. This simplifies our analysis to focus on the stochastic component. Between M_{ji} and $\boldsymbol{\epsilon}_j$ in this stochastic component, we separately optimize one while having the other fixed.

We start with fixed M_{ji} and optimizing for $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, diag(\boldsymbol{\sigma}_j^2))$. Notice that $\boldsymbol{\sigma}_j^2$ is associated with VAE$_j$ objective. In (Rolinek et al., 2019), the VAE objective is reformulated into the deterministic reconstruction, the stochastic reconstruction and the KL loss. The last two components define the stochastic loss of VAE. It is formulated as:

$$
\min_{V, \boldsymbol{\sigma}_j^2} \quad \sum_X \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, diag(\boldsymbol{\sigma}_j^2))} \left\|D\boldsymbol{\epsilon}_j\right\|^2 \quad s.t. \quad \sum_X L_{KL} = C_1, \tag{5}
$$

where X represents the dataset, D represents the local linear approximation of the Jacobian of the decoder with singular value decomposition as $D = U\Sigma V^T$. Furthermore, the KL loss $L_{KL} = \frac{1}{2}\sum_{k=1}^{d}(\mu_{jk}^2 + \sigma_{jk}^2 - \log\sigma_{jk}^2 - 1)$ can be simplified as $L_{\approx KL} = \frac{1}{2}\sum_{k \in \text{“}active\text{”}}(\mu_{jk}^2 - \log\sigma_{jk}^2 - 1)$ based on the “polarized” regime of VAE. (Rolinek et al., 2019) shows that $\boldsymbol{\sigma}_j^2$ act as the precision control allowed for each latent variable where more influential ones receive more precision. Combining the stochastic loss of the linear transformation in Eq. 4 and the stochastic loss of the original VAE in Eq. 5, the overall stochastic loss on $\boldsymbol{\sigma}_j^2$ can be formulated as:

$$
\min_{\boldsymbol{\sigma}_j^2} \quad \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, diag(\boldsymbol{\sigma}_j^2))} [\left\|M_{ji}\boldsymbol{\epsilon}_j\right\|^2 + \left\|D\boldsymbol{\epsilon}_j\right\|^2] \quad s.t. \quad \sum_k -\log\sigma_{jk}^2 = C_2, \tag{6}
$$

where σ_{jk}^2 is the kth element of $\boldsymbol{\sigma}_j^2$. Here we further simplify L_{KL} with the $L_{\approx KL}$ up to additive constants C_2 when $\boldsymbol{\mu}_j$ is fixed. In addition to the precision control of $\boldsymbol{\sigma}_j^2$ on VAE$_j$, this objective also aims to find an optimal distribution of $\boldsymbol{\sigma}_j^2$ that aligns the “polarized regime” among different VAEs. To see why, let c_k be the kth column of M_{ji}, we then have $\mathbb{E}\left\|M_{ji}\boldsymbol{\epsilon}_j\right\|^2 = \sum_k \left\|c_k\right\|^2 \sigma_{jk}^2$. The Arithmetic-Mean–Geometric-Mean (AM/GM) inequality suggests that $\sum_k \left\|c_k\right\|^2 \sigma_{jk}^2 \geq n \left(\prod_k \left\|c_k\right\|^2 \sigma_{jk}^2\right)^{1/n} = n \left(\prod_k \left\|c_k\right\|^2\right)^{1/n} \exp(-C)$, where the equality is achieved when $\left\|c_m\right\|^2 \sigma_{jm}^2 = \left\|c_n\right\|^2 \sigma_{jn}^2$ for any $m \neq n$. This suggests that latent variables with high $\left\|c_k\right\|^2$ mapping from $\mathbf{z_j}$ to $\mathbf{z_i}$ will have smaller variance. Hence, these latent variables in $\mathbf{z_j}$ are encouraged to stay in the “active” mode. On contrary, the latent variables that do not share similar generative factors between $\mathbf{z_j}$ and $\mathbf{z_i}$ will be assigned larger variance, and being pushed towards the “passive” mode.

Now we fix the optimal distribution of $\boldsymbol{\sigma}_j^2$, and optimize for M_{ji}. Since $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, diag(\boldsymbol{\sigma}_j^2))$, this objective can be understood as optimally rotating the latent space of VAE$_j$ such that the stochastic component in Eq. 4 is minimized. Specifically, we have the following objective:

$$\min_{M_{ji}} \|M_{ji}\boldsymbol{\epsilon}_j\|^2 = \min_{R} \left\|M_{ji}R^T\boldsymbol{\epsilon}_j\right\|^2 \tag{7}$$

where R is an orthogonal transformation. Let c'_k be the kth column of $M_{ji}R$. Similar as before, the AM/GM inequality suggests $\left\|M_{ji}R^T\boldsymbol{\epsilon}_j\right\|^2 = \sum_k \|c'_k\|^2 \sigma_{jk}^2 \geq \prod_k \|c'_k\|^2 \exp(-C_3)$. Hadamard's inequality suggests that $\prod_k \|c'_k\|^2 \geq |\det(M_{ji}R)|$, and the equality is satisfied when c'_ks are pair-wise orthogonal. This can be understood from the geometric perspective where $\prod_k \|c'_k\|^2$ gives an upper bound on Volume$(\{M_{ji}R^T x : x \in [0,1]^d\})$. As a result, the optimization of M_{ji} will lead to an orthogonal transformation. Together the optimization of Eq. 6 and Eq. 7 encourages the alignment of the "polarized regime" among different models under orthogonal linear transformations. They force different models in the ensemble to capture the same mixture of the generative factors. In the next section, we discuss the effect of the cross-model reconstruction in the VAE ensemble that encourages the disentangled representation over the entangled ones.

4.2 The Effect of cross-model Reconstruction

In an entangled representation, each latent variable captures a mixture of generative factors in its unique way. Since different generative factors typically have different effects on data variations (Duan et al., 2019), the orthogonal transformation from one entangled representation $\mathbf{z_j}$ to another one $\mathbf{z_i}$ introduces different encoding variance. Some of the transformed latent variables in $M_{ji}\mathbf{z_j}$ carry larger variance comparing to the corresponding ones in $\mathbf{z_i}$. This discrepancy leads to larger cross-model reconstruction of VAE$_i$ than the within model reconstruction. This error forces both VAE$_i$ and VAE$_j$ to adjust their representations until the effect on the data reconstruction by individual latent variables matches between $M_{ji}\mathbf{z_j}$ and $\mathbf{z_i}$. The process applies to all models in the ensemble and eventually leads to a one-to-one mapping of latent variables between different models, where M_{ji} becomes a trivial transformation (signed permutation matrix). In particular, if one of the models in the VAE ensemble learns a disentangled representation, other models in the ensemble will converge to it. This is because the orthogonal transformation from an entangled representation to a disentangled representation introduces larger cross-model encoding variance due to the mixture of different generative factors in the former, thus a larger cross-model reconstruction by the disentangled model. On contrary, the orthogonal transformation from a disentangled representation to an entangled representation would not introduce larger cross-model encoding variance than the within model encoding, thus similar cross-model reconstruction as within model reconstruction by the entangled model. Such a gap encourages the entangled representations to align with the disentangled representation. We illustrate the geometric interpretation of such a case in Appendix C.

From these discussions, we conclude that the VAE ensemble encourages different individual models to capture similar generative factors, thus learn representations that are "alike" up to a trivial transformation. In the next section, we verify these analytic results with experiments.

5 Experiments

Our experiments are designed to confirm the discussions in the previous sections. Particularly we ask the following questions: **Q1:** Do the linear transformations in the ensemble converge to trivial transformation? **Q2:** Do the VAEs in the ensemble work in similar "polarized" regime? **Q3:** Does VAE ensemble improve the unsupervised disentangled representation learning, and what is the effect of ensemble size? **Q4:** What are the effects of the cross-model reconstruction loss, the linear transformation loss and the hyperparameter γ in the VAE ensemble objective?

We analyze the inner working of the proposed VAE ensemble using the benchmark *dSprite* dataset (Matthey et al., 2017) with fully known generative processes, and the real-world *CelebA* dataset (Liu et al., 2015) with unknown generative process. Furthermore, for *dSprites* dataset, we compare our results with the original VAE model and the state-of-the-art disentanglement VAE models including β-VAE (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018), TC-VAE (Chen et al., 2018) and DIP-VAE (Kumar et al., 2017). We use two widely used supervised metrics including FactorVAE metric (Kim & Mnih, 2018) and DCI Disentanglement scores (Eastwood & Williams, 2018) as the

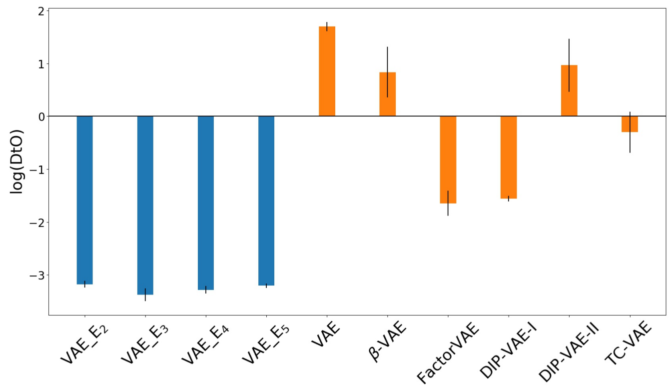
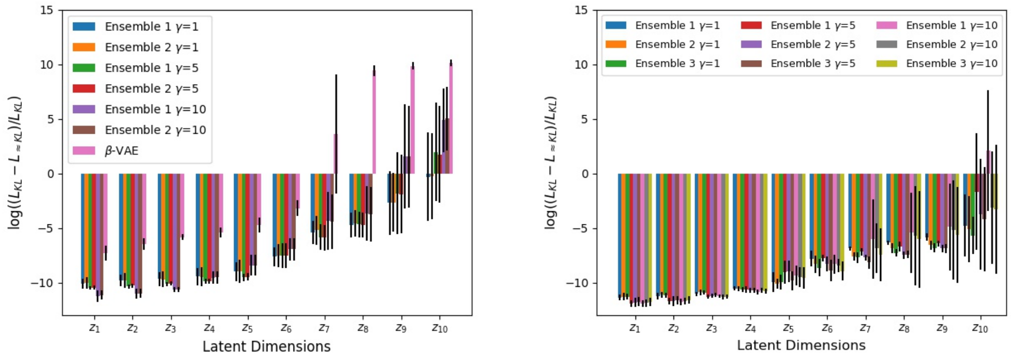


Figure 2: Comparing the DtO of linear transformations in the VAE ensemble (γ=10) with the one between well-trained individual VAEs, as well as the well-trained individual state-of-the-art VAE models. The latent dimension for all models is set to 10 and evaluated on the *dSprite* dataset.

quantitative measurements. They are shown to correlate with other common supervised metrics (Locatello et al., 2018). For example, FactorVAE metric and β-VAE metric (Higgins et al., 2017) capture similar notions, while DCI Disentanglement and Mutual Information Gap (MIG) (Chen et al., 2018) capture similar notions. In addition, DCI Disentanglement is closely related to the unsupervised model selection method UDR (Duan et al., 2019). For *CelebA* dataset, we show the latent traversal visulization as a qualitative measurement in Appendix E. We provide the details of the experiments in Appendix D.

Q1: We use the *Distance to Orthogonality* (DtO) (Rolinek et al., 2019) to check if the linear transformations in the ensemble converge to a signed permutation matrix during training. DtO is the Frobenius norm of the difference between a matrix M and its closest signed permutation matrix $P(M)$. It is solved with mixed-integer linear programming (MILP) formulation. The details on DtO can be found in Appendix B. In Figure 2, we show the DtO estimation of the linear transformations in the VAE ensemble of different ensemble size for the *dSprite* dataset. We show the mean and standard deviation of DtO across all linear transformations over 10 different runs. Furthermore, we compare these results with a baseline measurement where DtO is estimated for the linear transformations between the mean latent representations of well-trained individual models. Specifically, we use ten well-trained individual models and report the mean and standard deviation of the DtO estimations. As seen in the figure, the VAE ensemble models with different ensemble size all approach to trivial transformations between the individual models, while other VAE models do not have such property. In Fig. 6, we show that during training, the VAE ensemble remains maintains low DtO while the original VAEs do not have such property. A similar result for models trained on the *CelebA* dataset with different latent dimensions is shown in Fig. 7. Further discussion on these results are provided in Appendix E.

Q2: To check if the models in the VAE ensemble work in similar "polarized" regime, we estimate the relative error between L_{KL} and $L_{\approx KL}$ as $\Delta = \frac{L_{KL} - L_{\approx KL}}{L_{KL}}$ for each latent variable. Smaller Δ indicates closer matching between L_{KL} and $L_{\approx KL}$ of a latent variable, thus more "active". Figure 3(a) and 3(b) show $\log(\Delta)$ of the 10 latent variables of individual models in VAE_E$_2$ and VAE_E$_3$ with different γ settings trained on the *dSprite* dataset. The results show that individual models in the ensemble do work in similar "polarized regime'. In Figure 3(a), we also compare the VAE ensemble with the β-VAE where $\beta = 4$. This setting was found previously to be the optimal setting for the *dSprite* data for β-VAE (Higgins et al., 2017). We see that the VAE ensemble encourages more "active" latent variables than β-VAE. When we compare Fig. 3(a) and 3(b), we see that as the ensemble size increases, individual models are forced to have more "active" latent variables by decomposing the generative factors. This can be observed in the latent traversals shown in Appendix E. The *dSprites* dataset contains five ground truth generative factors. The VAE_E$_2$ models can have up to eight "active" latent variables depending on input, and these representations capture a decomposition of the ground truth generative factors.



(a) (b)

Figure 3: The "polarized regime" comparison between models in the VAE ensemble. The latent dimension is set to 10 and the results are over 10 runs of training on the *dSprite* dataset. (a) The "polarized regime" comparison by $\log(\frac{L_{KL}-L_{\approx KL}}{L_{KL}})$ of each latent variable of the two models in VAE_E$_2$ as well as β-VAE model. (b) Similar to (a) but for the three models in VAE_E$_3$.

FactorVAE Metric					
Individual Model		VAE Ensemble (VAE_E)			
VAE	0.635±0.083		γ=1	γ=5	γ=10
β-VAE (β=4)	0.665±0.089	VAE_E$_2$	0.711±0.106	0.736±0.085	0.741±0.086
FactorVAE (γ=40)	0.764±0.075	VAE_E$_3$	0.794±0.030	0.792±0.075	0.821±0.066
DIP-VAE-I (λ_{od}=5)	0.638±0.108	VAE_E$_4$	0.833±0.037	0.790±0.038	0.800±0.078
DIP-VAE-II (λ_{od}=5)	0.676±0.122	VAE_E$_5$	0.828±0.016	0.786±0.051	0.739±0.085
TC-VAE (β=4)	0.808±0.079				
DCI-Disentanglement Metric					
Individual Model		VAE Ensemble (VAE_E)			
VAE	0.143±0.033		γ=1	γ=5	γ=10
β-VAE (β=4)	0.198±0.076	VAE_E$_2$	0.176±0.043	0.243±0.029	0.201±0.037
FactorVAE (γ=40)	0.253±0.072	VAE_E$_3$	0.214±0.064	0.236±0.051	0.311±0.060
DIP-VAE-I (λ_{od}=5)	0.049±0.017	VAE_E$_4$	0.240±0.059	0.223±0.045	0.251±0.038
DIP-VAE-II (λ_{od}=5)	0.106±0.032	VAE_E$_5$	0.242±0.032	0.244±0.039	0.196±0.050
TC-VAE (β=4)	0.303±0.052				

Table 1: Comparison between the proposed VAE ensemble, the original VAE, and the current state-of-the-art disentangled VAE models. We report the mean and standard deviation of the FactorVAE metric and and DCI Disentanglement scores over 10 runs trained on the *dSprite* data.

Q3: In Table 1, we compare the disentangled representation performance between the proposed VAE ensemble and the state-of-the-art models. For the VAE ensemble, we report the performance of the first model in the ensemble. We also report the results for the VAE ensemble with different ensemble size and γ values. As shown in the table, the VAE ensemble significantly improves the performance over the original VAE model. In many settings, the VAE ensemble achieves similar or better performance over the state-of-the-art models. In Table 2, we evaluate the consistency among the models in the ensemble by reporting the standard deviation of the evaluation metrics using different models in the same ensemble. The small values confirm that different models in the ensemble learn similar latent representations. Furthermore, Table 1 shows the joint effect of ensemble size and γ setting. When $\gamma = 1$, the performance of VAE ensemble increases as the ensemble size increases, indicated by the higher mean and smaller variance of both the FactorVAE and DCI Disentanglement metrics. This behavior is consistent with the characteristic of ensemble learning where the increase in performance becomes smaller as the size of ensemble increases. However, as γ increases, having larger ensemble size can reduce the performance. We believe this

	γ	VAE_E$_2$	VAE_E$_3$	VAE_E$_4$	VAE_E$_5$
FactorVAE Metric	(γ=1)	0.0019	0.0090	0.0048	0.0081
	(γ=5)	0.0058	0.0064	0.0089	0.0163
	(γ=10)	0.0060	0.0046	0.0147	0.0139
DCI-Disent Metric	(γ=1)	0.0024	0.0026	0.0028	0.0036
	(γ=5)	0.0037	0.0054	0.0049	0.0040
	(γ=10)	0.0013	0.0024	0.0041	0.0042

Table 2: The comparison between individual models in the same ensemble. We report the average of the standard deviation of the metrics by individual models in the ensemble across 10 runs.

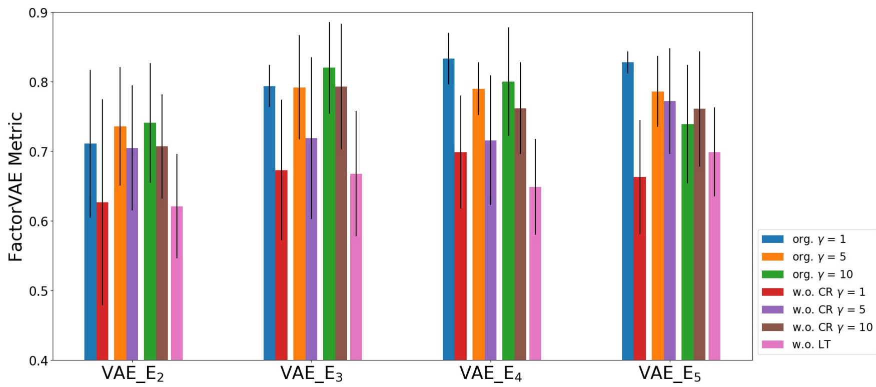


Figure 4: Ablation study to understand the effect of cross-model reconstruction and linear transformation in the VAE ensemble objective using the FactorVAE metric. (*w.o. CR* - without cross-model reconstruction loss; *w.o. LT* - without linear transformation loss; *org* - original VAE ensemble loss)

is due to the increased difficulty of balancing between the cross-model and within model objectives of VAE ensemble for larger ensembles. The reduced alignment of the latent representations among different models can also be seen in Table 2 where difference in the performance among individual models in the ensemble increases as ensemble size increases.

Q4: We conduct the ablation study to further understand the effect of the linear transformation loss and the cross-model reconstruction loss in the VAE ensemble objective. As shown in Fig. 4, removing either component leads to a lower FactorVAE metric for the VAE ensemble. Without the linear transformation loss, the performance of VAE ensemble decreases significantly across different ensemble sizes. Without the cross-model reconstruction loss, the performance of VAE ensemble also decreases but the gap becomes smaller as γ increases. This matches the discussion in Section 3 that higher γ forces closer mapping between the encoders and reduce the cross-model reconstruction error of the decoders. However, this also reduces the effect of cross-model reconstruction as discussed in Section 4.2. A similar result is also found for the DCI Disentanglement metric as shown in Fig. 8 in Appendix E. Overall, the results from the ablation study confirms the importance of both the linear transformation loss and the cross-model reconstruction loss in the VAE ensemble objective.

6 CONCLUSION

In this study, we propose a simple yet effective VAE ensemble framework consisting of multiple original VAEs to learn disentangled representation. The individual models in the ensemble are connected through linear layers that regularize both encoders and decoders to align the latent representations to be similar up to a linear transformation. We show in theory that the regularization by the VAE ensemble forces the linear transformations to be trivial transformations and show improved performance on the unsupervised disentangled representation learning. The theoretical discussion in Section 4 is based on the original VAE objective, and our experiments also focus on the ensemble with original VAE. We believe such framework can be extended to other disentangled VAE models, or even a mixture of different VAE models, as long as the regularization by the ensemble does not conflict with the augmented objective of these models.

REFERENCES

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β -VAE. (NIPS), 2017.

Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. (ICLR):1–13, 2018.

Harris Drucker, Corinna Cortes, Lawrence D Jackel, Yann LeCun, and Vladimir Vapnik. Boosting and other ensemble methods. *Neural Computation*, 6(6):1289–1301, 1994.

Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Chris Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations*, 2019.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. *ICLR*, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-Vae: Learning Basic Visual Concepts With a Constrained Variational Framework. *ICLR 2017*, (July):1–13, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv*, pp. 1–29, 2018.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pp. 3765–3773, 2016.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868, 2019.

AJ Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. Proceedings of Machine Learning Research, 2017.

Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809*, 2019.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *Proceedings of Machine Learning Research*, 80:2649–2658, 10–15 Jul 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *CoRR*, abs/1711.00848, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and L G Dec. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv*, pp. 1–33, 2018.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pp. 14584–14597, 2019.

Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. *arXiv preprint arXiv:1812.02833*, 2018.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pp. 5727–5736, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pp. 6076–6085, 2017.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2019.

Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pp. 14222–14235, 2019.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. In *Advances in Neural Information Processing Systems*, pp. 9584–9593, 2018.

Nicholas Watters, Loic Matthey, Matko Bosnjak, Christopher P Burgess, and Alexander Lerchner. Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *arXiv preprint arXiv:1905.09275*, 2019.

A Technical Lemmas

In this section, we give the lemmas used in the theoretical discussion in Section 4.

Lemma 1. *(Jensen's inequality) If $g(x)$ is a convex transformation on x, then this convex transformation of a mean $g[\mathbb{E}(x)]$ is less than or equal to the mean of the convex transformed value $\mathbb{E}[g(x)]$; it is a simple corollary that the opposite is true of concave transformations.*

Lemma 2. *(AM-GM inequality) As an extension of Jensen's inequality, given a list of non-negative real numbers $x_1, x_2, \ldots, x_n$, the arithmetic mean of this list $\frac{1}{n}\sum_{i=1}^{n} x_i$ is greater than or equal to the geometric mean of the same list $(\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$; and further, the equality holds if and only if $x_1 = x_2 = \cdots = x_n$.*

Lemma 3. *(Hadamard's inequality). if M is the matrix having columns c_i, then $|\det(M)| \leq \prod_{i=1}^{n} \|c_i\|$; and the equality in Hadamard's inequality is achieved if and only if the vectors are orthogonal.*

B Distance to Orthogonality (DtO)

In this section, we introduce the detail of *Distance to Orthogonality* (DtO) that is used in our experiment to check if the linear transformations in the VAE ensemble approach trivial transformations. This measurement is also used in (Rolinek et al., 2019) for a similar purpose. DtO is the Frobenius norm of the difference between a square matrix M and its closest signed permutation matrix $P(M)$. Finding $P(M)$ can be formulated as a mixed-integer linear programming (MILP) problem as following:

$$
\begin{aligned}
\min_P \; & \sum_{i,j} |M_{i,j} - P(M)_{i,j}| \\
s.t. \; & P(M)_{i,j} \in \{-1, 0, 1\}, \quad && \forall (i,j) \\
& \sum_i |P_{i,j}| = 1, && \forall j \\
& \sum_j |P_{i,j}| = 1, && \forall i
\end{aligned}
\tag{8}
$$

By introducing new variables $P^+_{i,j}, P^-_{i,j} \in \{0,1\}$ and $D_{i,j} = |M_{i,j} - P(M)_{i,j}|$, we can reformulate the above optimization problem as:

$$
\begin{aligned}
\min_P \; & \sum_{i,j} D_{i,j} \\
s.t. \; & (P^+_{i,j} - P^-_{i,j}) - M_{i,j} \leq D_{i,j}, \quad && \forall (i,j) \\
& M_{i,j} - (P^+_{i,j} - P^-_{i,j}) \leq D_{i,j}, && \forall (i,j) \\
& \sum_i (P^+_{i,j} + P^-_{i,j}) = 1, && \forall j \\
& \sum_j (P^+_{i,j} + P^-_{i,j}) = 1, && \forall i
\end{aligned}
\tag{9}
$$

Using this optimization formulation, DoT of a given matrix $M \in \mathbb{R}^{n \times n}$ is defined as:

$$
DoT = \frac{1}{n^2} \sum_{i,j} |M_{i,j} - P(M)_{i,j}| \tag{10}
$$

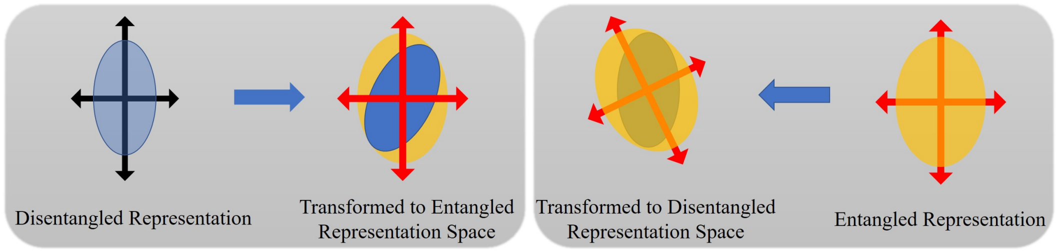


Figure 5: Geometric interpretation of the cross-model reconstruction between a disentangled representation space and an entangled representation space.

C GEOMETRIC INTERPRETATION OF THE EFFECT OF CROSS-MODEL RECONSTRUCTION

Given a disentangled and entangled latent representation space, Fig. 5 illustrates the effect of the cross-model reconstruction by VAE ensemble. The left part shows the orthogonal transformation from a disentangled representation to an entangled space, and the right part shows the transformation in the opposite direction. As shown in the figure, the orthogonal transformation from the disentangled representation to the entangled space does not introduce larger variance than the entangled representation. Hence, we can expect similar cross-model reconstruction and within model reconstruction. However, the transformation from the entangled representation to the disentangled space introduces larger variance (yellow shaded area over blue area on the right) than the disentangled representation. This leads to larger cross-model reconstruction by the disentangled model.

D MODEL ARCHITECTURE AND TRAINING DETAILS

We conducted our experiments, including training and evaluating the current state-of-the-art disentanglement models as well as evaluating the proposed VAE ensembles, using the `disentanglement_lib` [1] open-source library (Locatello et al., 2018).

Table 3 shows the encoder and the decoder architecture of the VAE model used in our experiments. This architecture is the same as the one used in the original β-VAE Higgins et al. (2017).

Encoder	Decoder
Input 64×64 binary/RGB image	Input $\mathbb{R}^d$
4×4 conv, 32 ReLu, stride 2, pad 1	FC d×256, ReLu
4×4 conv, 32 ReLu, stride 2, pad 1	4×4 upconv, 64 ReLu, stride 1
4×4 conv, 64 ReLu, stride 2, pad 1	4×4 conv, 64 ReLu, stride 2, pad 1
4×4 conv, 64 ReLu, stride 2, pad 1	4×4 conv, 32 ReLu, stride 2, pad 1
4×4 conv, 256 ReLu, stride 1	4×4 conv, 32 ReLu, stride 2, pad 1
FC 256 × (2×d)	4×4 conv, nc , stride 2, pad 1

Table 3: Encoder and Decoder architecture, d: dimension of the latent representation; nc: number of input image channel (For *dSprites* dataset $nc = 1$, for *CelebA* dataset $nc = 3$).

Table 4 shows the hyperparameters setting used throughout the experiments. These parameters are fixed for all the experiments.

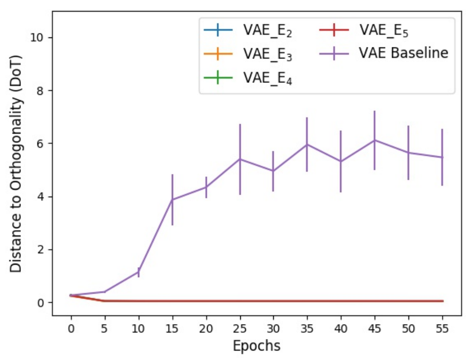
E ADDITIONAL EXPERIMENTAL RESULTS

In this section, we present the additional results including the DtO and "polarized regime" analysis on the models trained on the *CelebA* dataset similar to the ones conducted on *dSprite* dataset in Section 5; the ablation results with DCI-Disentanglement metric and the DtO estimation; and the

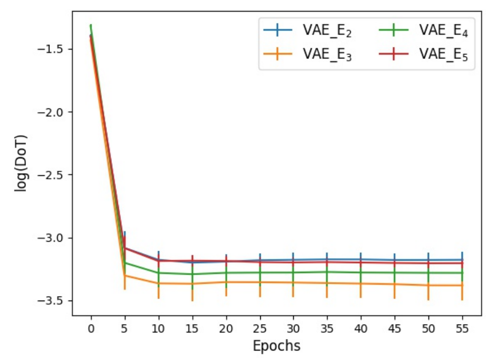
[1] `https://github.com/google-research/disentanglement_lib`

Parameter	value
Batch size	64
Latent dimension	10
Optimizer	Adam
Adam: beta1	0.9
Adam: beta2	0.999
Learning rate	1e-4

Table 4: Hyperparameters setting.



(a)



(b)

Figure 6: Characteristics of the linear transform between latent representations. The latent dimension is set to 10 and the results are over 10 runs of training on the *dSprite* dataset. (a) Comparing the DtO of linear transformations in the VAE ensemble (γ=10) and the one between original VAEs. (b) VAE ensemble (γ=10) with different ensemble size all achieve small DtO of the linear transformations between the models.

latent traversal of the trained model on both *dSprite* and *CelebA* dataset along with further discuss the effect of VAE ensemble on the latent representation.

E.1 Characterization of the Linear Transformation in VAE Ensemble

In Figure 6, we show the DtO estimation of the linear transformations in the ensemble during training for the *dSprite* dataset. We report the mean and standard deviation of DtO across all linear transformations over 10 different runs. Furthermore, we compare these results with a VAE baseline where DtO is estimated for the linear transformations between original VAEs. Specifically, we train ten different VAEs separately and estimate the DtO of the pairwise linear transformation among these models during training. Similarly we report the mean and standard deviation of these DtO estimations. As seen in the figure, the VAE ensemble models with different ensemble size all approach to trivial transformations between the individual models, while the original VAEs do not have such property. A similar result is also found in models trained for *CelebA* dataset. Similar to the results in Figure 6, we observe decreased DtO of the linear transformations in the VAE ensemble during training.

We also compare models trained with different latent dimension size. We observe decreased DtO as the latent dimension of the model increases in Figure 7. This is because, as discussed in the main paper, the VAE ensemble encourages more "active" latent variables. Models with higher latent dimension likely to learn a decomposition of generative factors. As a result, the alignment of the latent variable between different models are easier thus the linear transformations between the latent representations is closer to the trivial transformation. On the contrary when there are less latent variables in the model than the generative factors, some of the latent variables will capture more than a single generative factor. As a result, the one-to-one mapping between the latent variables of different models will not lead to a trivial transformation.

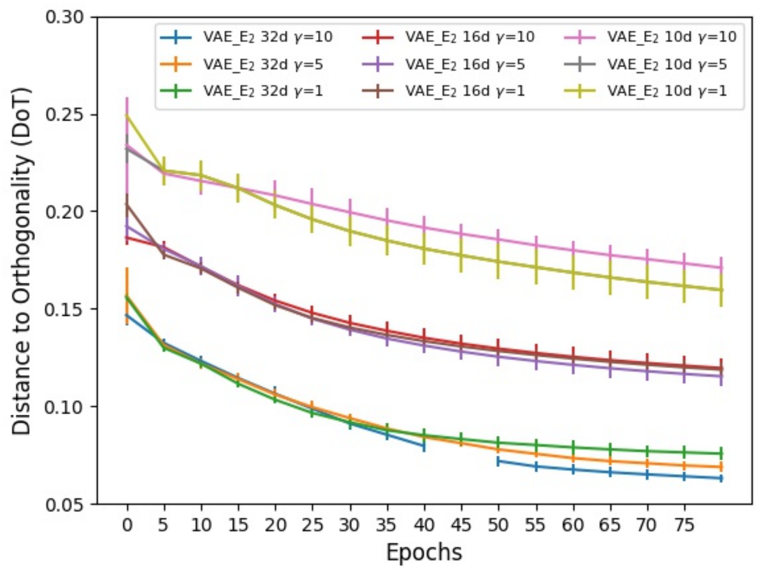


Figure 7: Distance to Orthogonality (DtO) measurement of the linear transforms between latent representations in VAE_E$_2$ during training on the *CelebA* dataset. We also compare models with different latent dimensions of 10, 16 and 32 and the results are averaged over 5 runs. In the figure legend, we use "VAE_E$_i$ nd $\gamma = g$" to represent VAE Ensemble (VAE_E) model with i individual VAE models, n latent dimensions and γ value equal to g.

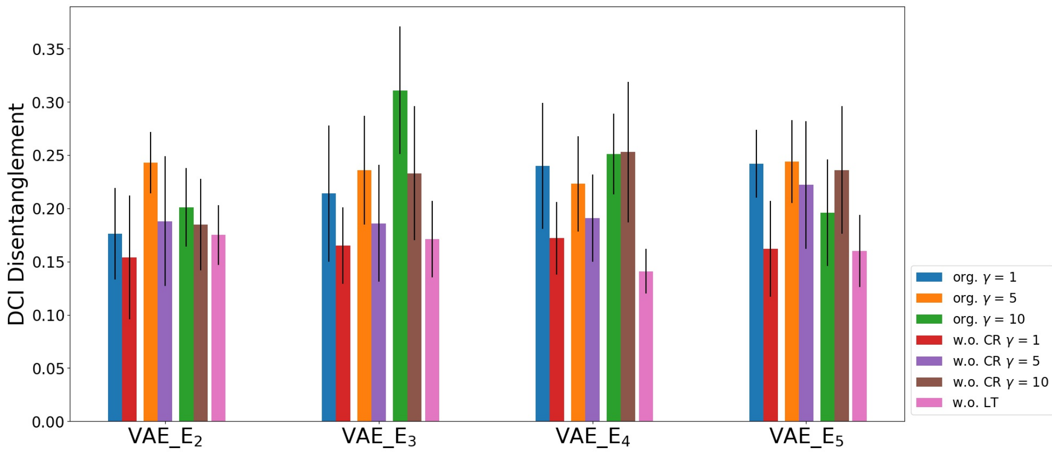
In Figure 9 and Figure 10, we show the "polarized regime" estimation for models in VAE_E$_2$ and VAE_E$_3$ trained for *CelebA* datasets, respectively. Similar to the results in Figure 3, individual models in the VAE ensemble tend to have similar 'polarized regime", and higher γ enforces the "polarized regime" by separating "passive" latent variables from the "active" ones. When we compare between VAE_E$_2$ and VAE_E$_3$, we observe increased "active" latent variables similar to the result on *dSprite* dataset in Section 5. More importantly, as discussed earlier, latent variables in a model with limited latent dimensions need to capture more than a single generative factor, especially for a complicated real-world dataset such as *CelebA*. This makes the linear transformation between the latent representations less trivial. As the latent dimension size grows, such constraint is relaxed and the linear transformations are closer to trivial.

These additional results confirm the conclusion in Section 5: (1) as the ensemble size increases, DtO increases due to the difficulty of aligning the latent representations among different models; (2) as the model latent dimension increases, DtO decreases due to the increased model capacity, and encourages the one-to-one mapping between latent variables in different models; (3) hyperparameter γ does not affect DtO significantly, but plays an important role on separating "active" and "passive" latent variables, especially when the latent dimension is large enough.

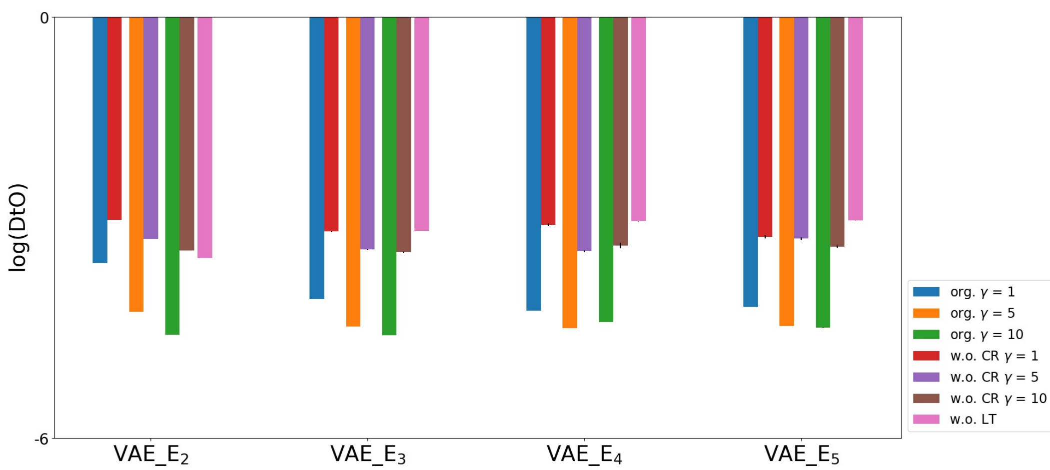
Furthermore, we believe the DtO measurement of the linear transformation in VAE ensemble could be a useful indicator for latent dimension size. As shown in Figure 7 and Figure 9, when the latent dimension is sufficient for a given dataset, the DtO of the linear transformation is small and some latent variables are pushed to "passive" mode.

E.2 Ablation Study

Similar to the ablation result shown in Section 5, here we show the same ablation study using the DCI Disentanglement metric in Fig. 8(a) as well as the DtO measurement 8(b). Similar as the results of the FactorVAE metric in Fig. 4, removing either component leads to a lower DCI Disentanglement metric for the VAE ensemble. Without the linear transformation loss, the performance of VAE ensemble decreases significantly across different ensemble sizes. Fig. 8(b) shows that for VAE



(a) Comparison on the DCI-Disentanglement score



(b) Comparison on the Distance to Orthogonality (DtO)

Figure 8: Ablation study to understand the effect of cross-model reconstruction and linear transformation in the VAE ensemble objective using the DCI Disentanglement metric and DtO. (*w.o. CR* - without cross-model reconstruction loss; *w.o. LT* - without linear transformation loss; *org* - original VAE ensemble loss)

ensemble without the cross-model reconstruction, the linear transformations among models are close to a trivial transformation (signed permutation). This implies the orthonormal transformation of the linear transformations. This result further supports our intuitive justification in Appendix C that the cross-model objective encourages entangled models to align to disentangled models. Indeed, we see that adding the cross-model reconstruction can further reduce the DtO of the linear transformations among the models in the ensemble.

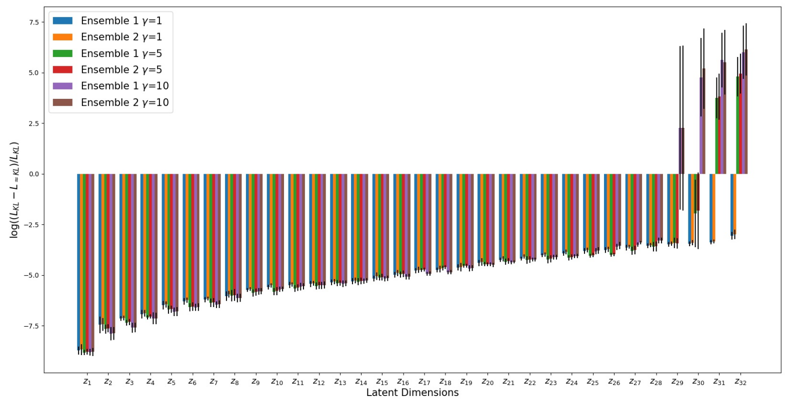
E.3 Latent Traversal

In this section we show the latent traversal of models trained on both *dSprites* and *CelebA* datasets. For a fixed input image, to extract the latent traversal we change the value of a single latent variable z_i in the corresponding encoding, and observe the generated output image to understand the effect of z_i. The range of the value are usually chosen to be from -3 to 3 due to the standard Gaussian prior.

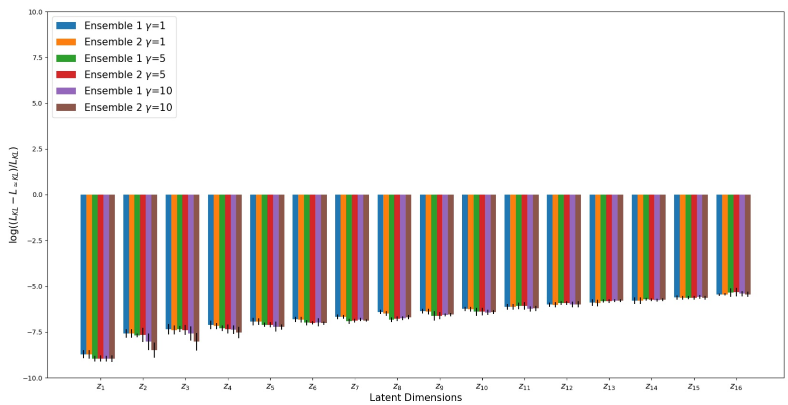
In Figure 11, we show the latent traversal for both VAE_E$_2$ and a single VAE model with 10 latent dimensions trained on *dSprites* dataset. Three images as shown in the last column of each block are used as input. Both models are able to capture certain generative factors of the data including position, shape, rotation and scale. In Section 5, we argue that the representation by VAE ensemble encourages more "active" latent variables, thus can capture a decomposition of the ground truth generative factors. Especially from the "polarized regime" estimation in Figure 3, we observe that some latent variables in the VAE ensemble are in-between "active" and "passive" modes. This suggests that the VAE ensemble model generates input-dependent factors based on the input complexity. In Figure 11, we observe this behavior highlighted with color boxes. The traversal on the second latent variable z_2 shows that an ellipse shape does not lead to an "active" latent variable. However, both heart and square shape lead to an "active" latent variable that changes the output. In contrast, the single VAE model does not have such behavior where the "active" modes are consistent across different input data.

In Figure 12 and Figure 13, we show the latent traversal for both VAE_E$_2$ and a single VAE model with 16 latent dimensions trained on *CelebA* dataset, respectively. In this real-world dataset, the generative factors are unknown. We observe different factors including background, azimuth, gender, hair style being captured by both models. Similar as before, the single VAE model maintains similar "active" mode for all latent variables where similar traversal patterns are observed for both input images. However, VAE_E$_2$ shows semantically consistent but input-dependent "active" mode. This is translated into different traversal effects and more realistic and sharper images by VAE_E$_2$, especially for the first input image that is less common in the dataset. We believe this is important towards a meaningful compositional latent representation learning.

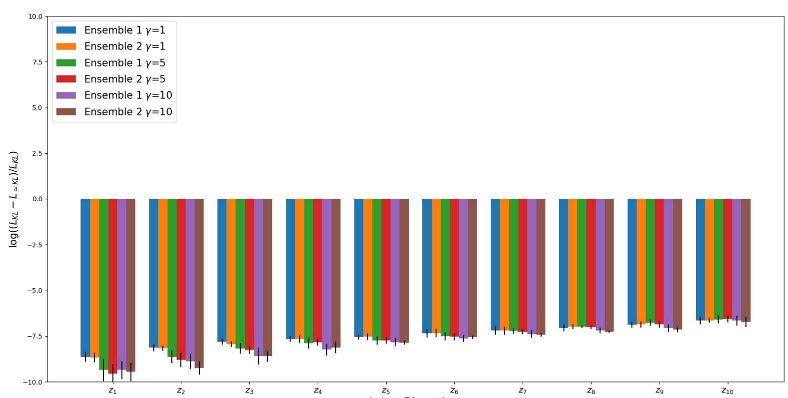
Overall, the latent traversal results in this section confirm the findings on the inner working of the VAE ensemble shown in the previous section as well as the discussion in Section 5.



(a) VAE_E$_2$ with latent dimension of 32, CelebA data.

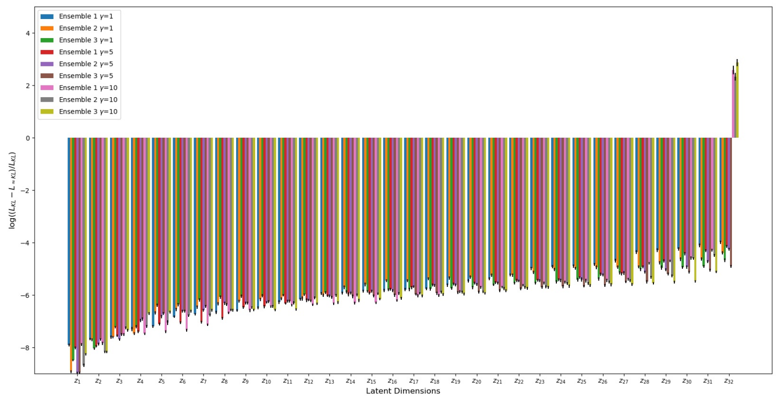


(b) VAE_E$_2$ with latent dimension of 16, CelebA data.

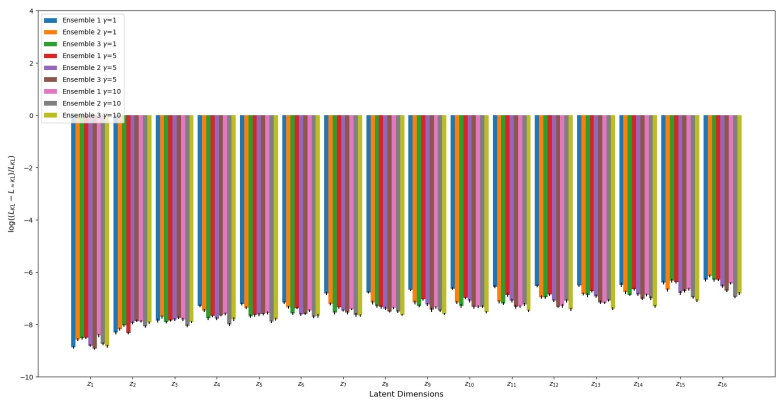


(c) VAE_E$_2$ with latent dimension of 10, CelebA data.

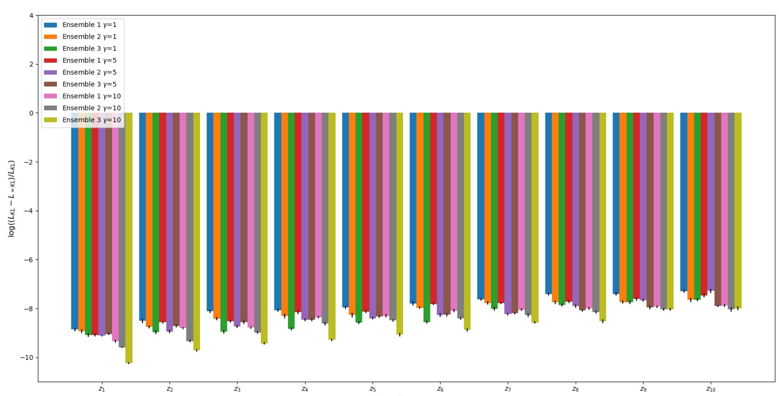
Figure 9: The "polarized regime" comparison between models in VAE_E$_2$. The results are over 5 runs of training on the *CelebA* dataset.



(a) VAE_E$_3$ with latent dimension of 32, CelebA data.

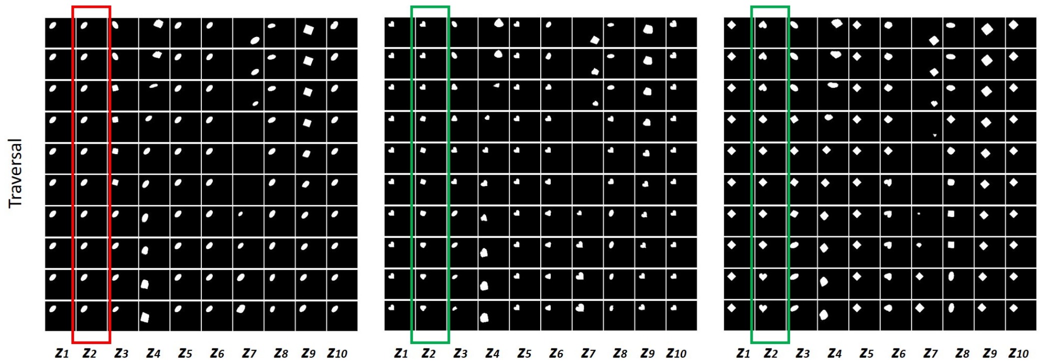


(b) VAE_E$_3$ with latent dimension of 16, CelebA data.

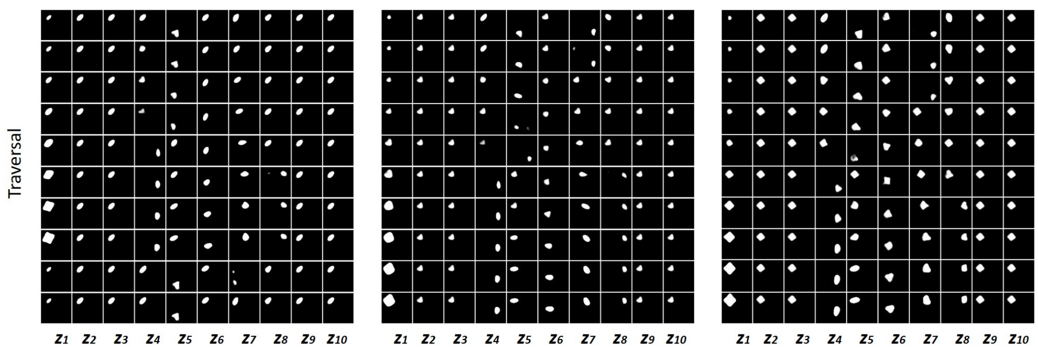


(c) VAE_E$_3$ with latent dimension of 10, CelebA data.

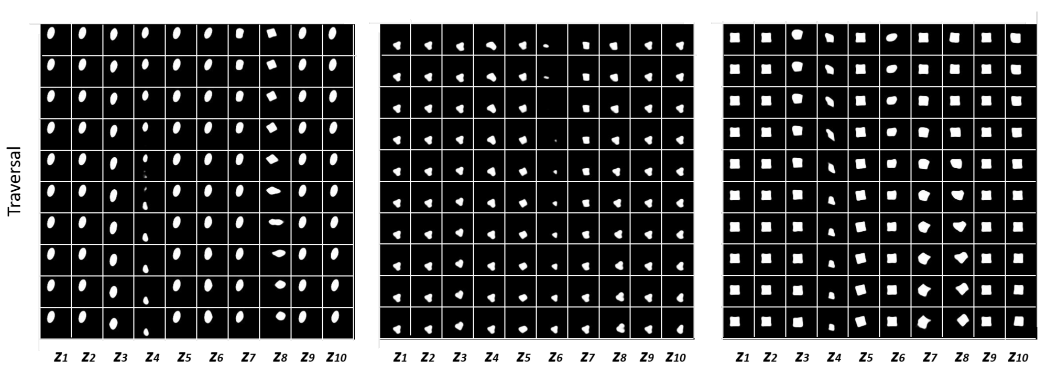
Figure 10: The “polarized regime” comparison between models in the VAE_E$_3$ on the *CelebA* dataset.



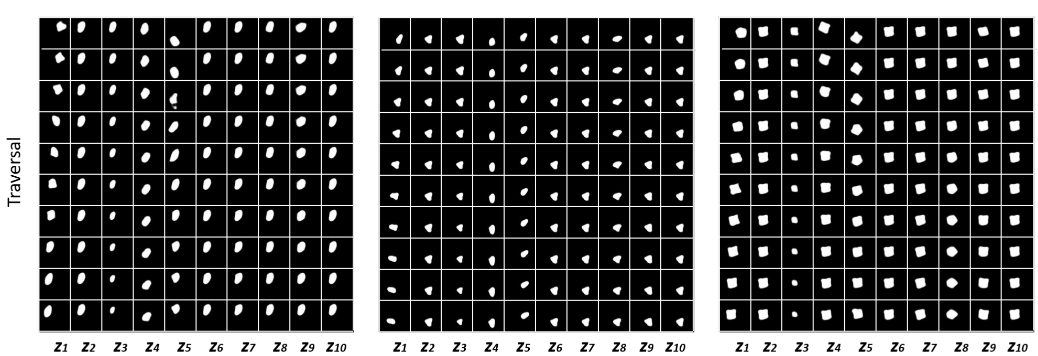
(a) VAE_E$_2$



(b) Single VAE

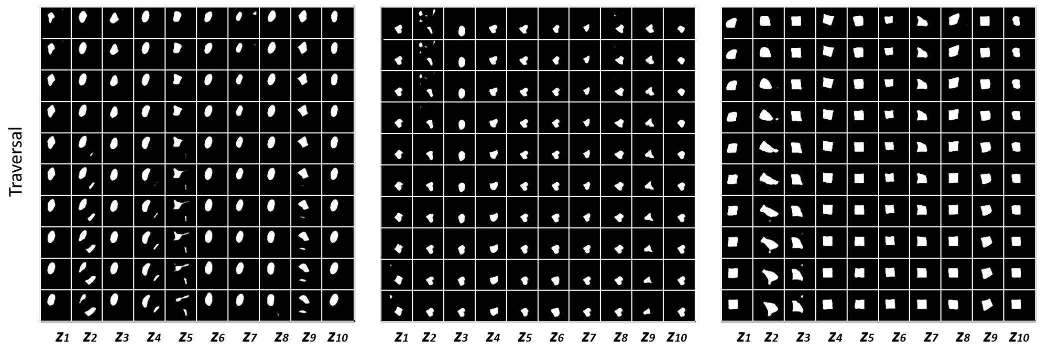


(c) β- VAE

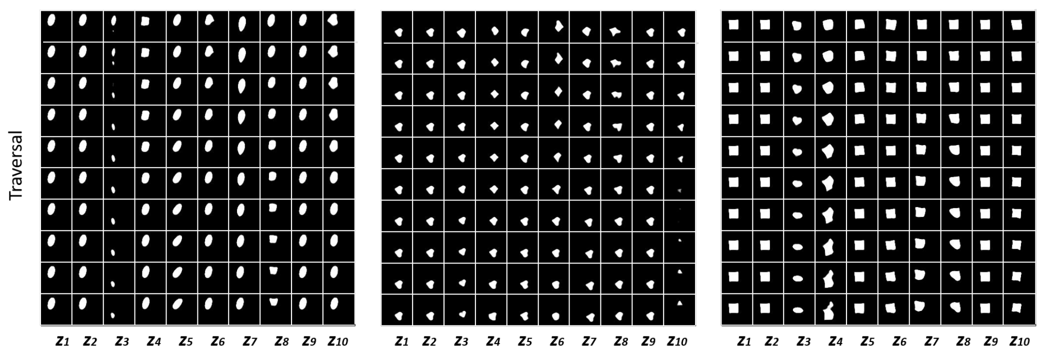


(d) FactorVAE

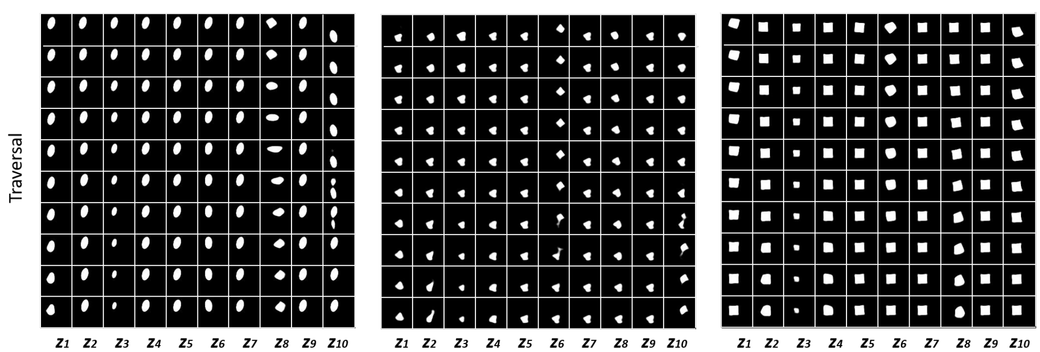
Figure 11: Latent traversal on three different input images using VAE_E$_2$, a single VAE and the state-of-the-art VAE models with 10 dimensional latent representation. The three input images are ellipse, heart and square shapes as shown in the last column.



(e) DIP-VAE-I



(f) DIP-VAE-II



(g) TC-VAE

Figure 11: (cont.) Latent traversal on three different input images using VAE_E$_2$, a single VAE and the state-of-the-art VAE models with 10 dimensional latent representation. The three input images are ellipse, heart and square shapes as shown in the last column.

Figure 12: Latent traversal on two different input images of *CelebA* dataset using VAE_E_2 with latent dimension of 16.

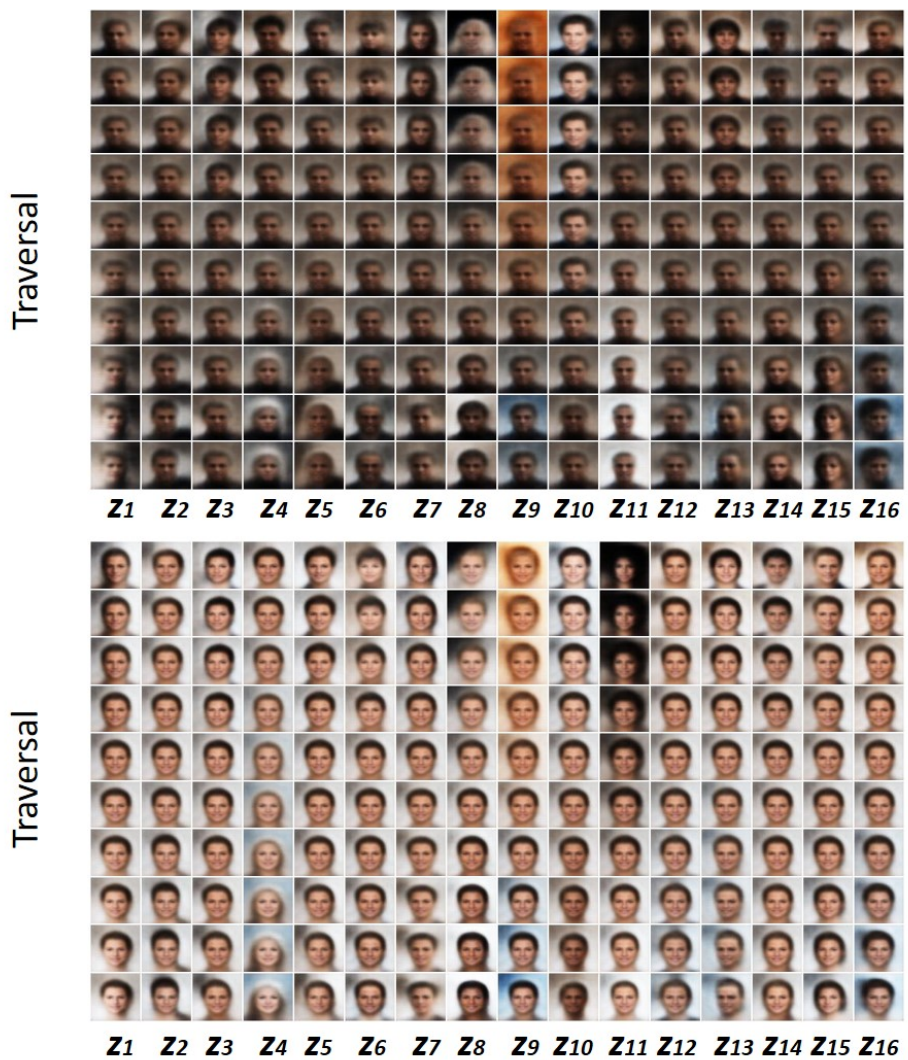


Figure 13: Latent traversal on two different input images of *CelebA* dataset using a single VAE with latent dimension of 16.