GUIDED VARIATIONAL AUTOENCODER FOR DISENTANGLEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose an algorithm, guided variational autoencoder (Guided-VAE), that is able to learn a controllable generative model by performing latent representation disentanglement learning. The learning objective is achieved by providing signal to the latent encoding/embedding in VAE without changing its main backbone architecture, hence retaining the desirable properties of the VAE. We design an unsupervised and a supervised strategy in Guided-VAE and observe enhanced modeling and controlling capability over the vanilla VAE. In the unsupervised strategy, we guide the VAE learning by introducing a lightweight decoder that learns latent geometric transformation and principal components; in the supervised strategy, we use an adversarial excitation and inhibition mechanism to encourage the disentanglement of the latent variables. Guided-VAE enjoys its transparency and simplicity for the general representation learning task, as well as disentanglement learning. On a number of experiments for representation learning, improved synthesis/sampling, better disentanglement for classification, and reduced classification errors in meta learning have been observed.

1 INTRODUCTION

The resurgence of autoencoders (AE) (LeCun, 1987; Bourlard & Kamp, 1988; Hinton & Zemel, 1994) is an important component in the rapid development of modern deep learning (Goodfellow et al., 2016). Autoencoders have been widely adopted for modeling signals and images (Poultney et al., 2007; Vincent et al., 2010). Its statistical counterpart, the variational autoencoder (VAE) (Kingma & Welling, 2014), has led to a recent wave of development in generative modeling due to its two-in-one capability, both representation and statistical learning in a single framework. Another exploding direction in generative modeling includes generative adversarial networks (GAN) Goodfellow et al. (2014), but GANs focus on the generation process and are not aimed at representation learning (without an encoder at least in its vanilla version).

Compared with classical dimensionality reduction methods like principal component analysis (PCA) (Hotelling, 1933; Jolliffe, 2011) and Laplacian eigenmaps (Belkin & Niyogi, 2003), VAEs have demonstrated their unprecedented power in modeling high dimensional data of real-world complexity. However, there is still a large room to improve for VAEs to achieve a high quality reconstruction/synthesis. Additionally, it is desirable to make the VAE representation learning more transparent, interpretable, and controllable.

In this paper, we attempt to learn a transparent representation by introducing guidance to the latent variables in a VAE. We design two strategies for our Guided-VAE, an unsupervised version (Fig. 1.a) and a supervised version (Fig. 1.b). The main motivation behind Guided-VAE is to encourage the latent representation to be semantically interpretable, while maintaining the integrity of the basic VAE architecture. Guided-VAE is learned in a multi-task learning fashion. The objective is achieved by taking advantage of the modeling flexibility and the large solution space of the VAE under a lightweight target. Thus the two tasks, learning a good VAE and making the latent variables controllable, become companions rather than conflicts.

In **unsupervised Guided-VAE**, in addition to the standard VAE backbone, we also explicitly force the latent variables to go through a lightweight encoder that learns a deformable PCA. As seen in Fig. 1.a, two decoders exist, both trying to reconstruct the input data $\mathbf{x}$: Dec_{main}. The main decoder, denoted as Dec_{main}, functions regularly as in the standard VAE (Kingma & Welling, 2014); the secondary decoder, denoted as Dec$_{sub}$, explicitly learns a geometric deformation together with a linear sub-space. In **supervised Guided-VAE**, we introduce a subtask for the VAE by forcing one latent variable to be discriminative (minimizing the classification error) while making the rest of the latent variable to be adversarially discriminative (maximizing the minimal classification error). This subtask is achieved using an adversarial excitation and inhibition formulation. Similar to the unsupervised Guided-VAE, the training process is carried out in an end-to-end multi-task learning manner. The result is a regular generative model that keeps the original VAE properties intact, while having the specified latent variable semantically meaningful and capable of controlling/synthesizing a specific attribute. We apply Guided-VAE to the data modeling and few-shot learning problems and show favorable results on the MNIST, CelebA, and Omniglot datasets.

The contributions of our work can be summarized as follows:

- We propose a new generative model disentanglement learning method by introducing latent variable guidance to variational autoencoders (VAE). Both unsupervised and supervised versions of Guided-VAE have been developed.
- In unsupervised Guided-VAE, we introduce deformable PCA as a subtask to guide the general VAE learning process, making the latent variables interpretable and controllable.
- In supervised Guided-VAE, we use an adversarial excitation and inhibition mechanism to encourage the disentanglement, informativeness, and controllability of the latent variables.

Guided-VAE is able to keep the attractive properties of the VAE and it is easy to implement. It can be trained in an end-to-end fashion. It significantly improves the controllability of the vanilla VAE and is applicable to a range of problems for generative modeling and representation learning.

2 Related Work

Related work can be discussed along several directions.

Generative model families such as generative adversarial networks (GAN) (Goodfellow et al., 2014; Arjovsky et al., 2017) and variational autoencoder (VAE) (Kingma & Welling, 2014) have received a tremendous amount of attention lately. Although GAN produces higher quality synthesis than VAE, GAN is missing the encoder part and hence is not directly suited for representation learning. Here, we focus on disentanglement learning by making VAE more controllable and transparent.

Disentanglement learning (Mathieu et al., 2016; Szabó et al., 2018; Hu et al., 2018; Achille & Soatto, 2018; Gonzalez-Garcia et al., 2018; Jha et al., 2018) recently becomes a popular topic in representation learning. Adversarial training has been adopted in approaches such as (Mathieu et al., 2016; Szabó et al., 2018). Various methods (Peng et al., 2017; Kim & Mnih, 2018; Lin et al., 2019) have imposed constraints/regularizations/supervisions to the latent variables but these existing approaches often involve an architectural change to the VAE backbone and the additional components in these approaches are not provided as secondary decoder for guiding the main encoder. A closely related work is the β-VAE (Higgins et al., 2017) approach in which a balancing term β is introduced to control the capacity and the independence prior. β-TCVAE (Chen et al., 2018) further extends β-VAE by introducing a total correlation term.

From a different angle, principal component analysis (PCA) family (Hotelling, 1933; Jolliffe, 2011; Candès et al., 2011) can also be viewed as representation learning. Connections between robust PCA (Candès et al., 2011) and VAE (Kingma & Welling, 2014) have been observed (Dai et al., 2018). Although being a widely adopted method, PCA nevertheless has limited modeling capability due to its linear subspace assumption. To alleviate the strong requirement for the input data being pre-aligned, RASL (Peng et al., 2012) deals with unaligned data by estimating a hidden transformation to each input. Here, we take the advantage of the transparency of PCA and the modeling power of VAE by developing a sub-encoder (see Fig. 1.a), deformable PCA, that guides the VAE training process in an integrated end-to-end manner. After training, the sub-encoder can be removed by keeping the main VAE backbone only.

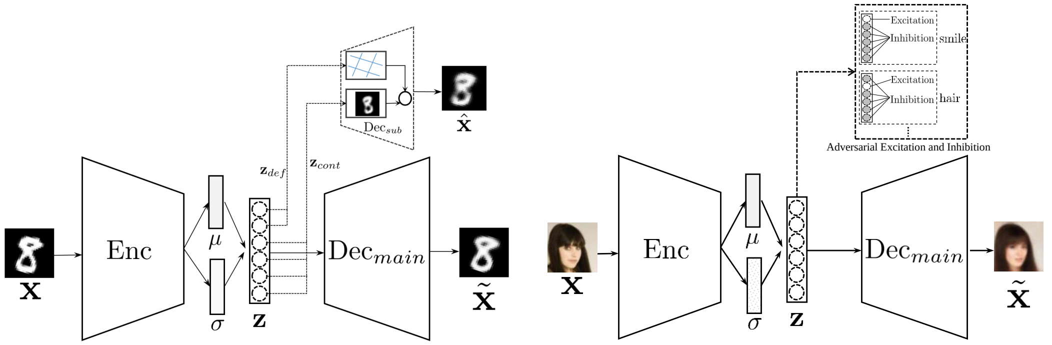
To achieve disentanglement learning in supervised Guided-VAE, we encourage one latent variable to directly correspond to an attribute while making the rest of the variables uncorrelated. This is analogous to the excitation-inhibition mechanism (Yizhar et al., 2011) or the explaining-away

(Wellman & Henrion, 1993) phenomena. Existing approaches (Liu et al., 2018; Lin et al., 2019) impose supervision as a conditional model for an image translation task, whereas our supervised Guided-VAE model targets the generic generative modeling task by using an adversarial excitation and inhibition formulation. This is achieved by minimizing the discriminative loss for the desired latent variable while maximizing the minimal classification error for the rest of the variables. Our formulation has connection to the domain-adversarial neural networks (DANN) (Ganin et al., 2016) but the two methods differ in purpose and classification formulation. Supervised Guided-VAE is also related to the adversarial autoencoder approach Makhzani et al. (2016) but the two methods differ in objective, formulation, network structure, and task domain. In (Ilse et al., 2019), the domain invariant variational autoencoders method (DIVA) differs from ours by enforcing disjoint sectors to explain certain attributes.

Our model also has connections to the deeply-supervised nets (DSN) (Lee et al., 2015) where intermediate supervision is added to a standard CNN classifier. There are also approaches (Engel et al., 2018; Bojanowski et al., 2018) in which latent variables constraints are added but they have different formulations and objectives than Guided-VAE. Recent efforts in fairness disentanglement learning (Creager et al., 2019; Song et al., 2018) also bear some similarity but there is still with a large difference in formulation.

3 Guided-VAE Model

In this section, we present the main formulations of our Guided-VAE models. The unsupervised Guided-VAE version is presented first, followed by introduction of the supervised version.



(a) Unsupervised Guided-VAE

(b) Supervised Guided-VAE

Figure 1: Model architecture for the proposed Guided-VAE algorithms.

3.1 VAE

Following the standard definition in variational autoencoder (VAE) (Kingma & Welling, 2014), a set of input data is denoted as $\mathrm{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ where n denotes the number of total input samples. The latent variables are denoted by vector $\mathbf{z}$. The encoder network includes network and variational parameters $\boldsymbol{\phi}$ that produces variational probability model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The decoder network is parameterized by $\boldsymbol{\theta}$ to reconstruct sample $\tilde{\mathbf{x}} = f_{\boldsymbol{\theta}}(\mathbf{z})$. The log likelihood $\log p(\mathbf{x})$ estimation is achieved by maximizing the Evidence Lower BOund (ELBO) (Kingma & Welling, 2014):

$$ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log(p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}))] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \tag{1}$$

The first term in eq. (1) corresponds to a reconstruction loss $\int q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \times ||\mathbf{x} - f_{\boldsymbol{\theta}}(\mathbf{z})||^2 d\mathbf{z}$ (the first term is the *negative* of reconstruction loss between input $\mathbf{x}$ and reconstruction $\tilde{\mathbf{x}}$) under Gaussian parameterization of the output. The second term in eq. (1) refers to the KL divergence between the variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and the prior distribution $p(\mathbf{z})$. The training process thus tries to find the optimal $(\boldsymbol{\theta}, \boldsymbol{\phi})^*$ such that:

$$(\boldsymbol{\theta},\boldsymbol{\phi})^* = \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{i=1}^n ELBO(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}_i) = \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{i=1}^n \left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}[\log(p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}))] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z}))\right] \tag{2}$$

3.2 UNSUPERVISED GUIDED-VAE

In our unsupervised Guided-VAE, we introduce a deformable PCA as a secondary decoder to guide the VAE training. An illustration can be seen in Fig. 1.a. This secondary decoder is called Dec_{sub}. Without loss of generality, we let $\mathbf{z} = (\mathbf{z}_{def}, \mathbf{z}_{cont})$. $\mathbf{z}_{def}$ decides a deformation/transformation field, e.g. an affine transformation denoted as $\tau(\mathbf{z}_{def})$. $\mathbf{z}_{cont}$ determines the content of a sample image for transformation. The PCA model consists of K basis $B = (\mathbf{b}_1, ..., \mathbf{b}_K)$. We define a deformable PCA loss as:

$$\mathcal{L}_{DPCA}(\boldsymbol{\phi}, B) = \big[\sum_{i=1}^{n} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{def}, \mathbf{z}_{cont}|\mathbf{x}_i)}[||\mathbf{x}_i - \tau(\mathbf{z}_{def}) \circ (\mathbf{z}_{cont} B^T)||^2\big] + \sum_{k,j\neq k} (\mathbf{b}_k^T \mathbf{b}_j)^2, \quad (3)$$

where $\circ$ defines a transformation (affine in our experiments) operator decided by $\tau(\mathbf{z}_{def})$ and $\sum_{k,j\neq k}(\mathbf{b}_k^T \mathbf{b}_j)^2$ is regarded as the orthogonal loss. A normalization term $\sum_k (\mathbf{b}_k^T \mathbf{b}_k - 1)^2$ can be optionally added to force the basis to be unit vectors. We follow the spirit of the PCA optimization and a general formulation for learning PCA, which can be found in (Candès et al., 2011). To keep the simplicity of the method we learn a fixed basis function B and one can also adopt a probabilistic PCA model (Tipping & Bishop, 1999). Thus, learning unsupervised Guided-VAE becomes:

$$(\boldsymbol{\theta}, \boldsymbol{\phi}, B)^* = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}, B} \big\{ \big[\sum_{i=1}^{n} ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_i) \big] - \mathcal{L}_{DPCA}(\boldsymbol{\phi}, B) \big\}. \quad (4)$$

3.3 SUPERVISED GUIDED-VAE

For training data $\text{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, suppose there exists a total of T attributes with ground-truth labels. The t-th attribute, let $\mathbf{z} = (z_t, \mathbf{z}_t^{rst})$ where z_t defines a scalar variable deciding to decide the t-th attribute and $\mathbf{z}_t^{rst}$ represents remaining latent variables. Let $y_t(\mathbf{x}_i)$ be the ground-truth label for the $t-$th attribute of sample $\mathbf{x}_i$; $y_t(\mathbf{x}_i) \in \{-1, +1\}$. For each attribute, we use an adversarial excitation and inhibition method with term:

$$\mathcal{L}_{Excitation}(\boldsymbol{\phi}, t) = -\sum_{i=1}^{n} \mathbb{E}_{q_{\boldsymbol{\phi}}(z_t|\mathbf{x}_i)}[(1 - y_t(\mathbf{x}_i) \times z_t)_+], \quad (5)$$

which is a hinge term. This is an excitation process since we want latent variable z_t to directly correspond to the attribute label. Notice the $-$ sign before the summation since this term will be combined with eq. (1) for maximization.

$$\mathcal{L}_{Inhibition}(\boldsymbol{\phi}, t) = \inf_{C_t} \{ \sum_{i=1}^{n} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_t^{rst}|\mathbf{x}_i)}[-\log p_{C_t}(y = y_t(\mathbf{x}_i)|\mathbf{z}_t^{rst})] \}, \quad (6)$$

where $C_t(\mathbf{z}_t^{rst})$ refers to classifier making a prediction for the t-th attribute using the remaining latent variables $\mathbf{z}_t^{rst}$. $-\log p_{C_t}(y = y(\mathbf{x})|\mathbf{z}_t^{rst})$ is a cross-entropy term for minimizing the classification error in eq. (6). This is an inhibition process since we want the remaining variables $\mathbf{z}_t^{rst}$ as independent as possible to the attribute label.

$$(\boldsymbol{\theta}, \boldsymbol{\phi})^* = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \big\{ \big[\sum_{i=1}^{n} ELBO(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_i) \big] + \sum_{t=1}^{T} [\mathcal{L}_{Excitation}(\boldsymbol{\phi}, t) + \mathcal{L}_{Inhibition}(\boldsymbol{\phi}, t)] \big\}. \quad (7)$$

Note that the term $\mathcal{L}_{Inhibition}(\boldsymbol{\phi}, t)$ within eq. (7) for maximization is an adversarial term to make $\mathbf{z}_t^{rst}$ as uninformative to attribute t as possible, by making the best possible classifier C_t to be undiscriminative. The formulation of eq. (7) bears certain similarity to that in domain-adversarial neural networks (Ganin et al., 2016) in which the label classification is minimized with the domain classifier being adversarially maximized. Here, however, we respectively encourage and discourage different parts of the features to make the same type of classification.

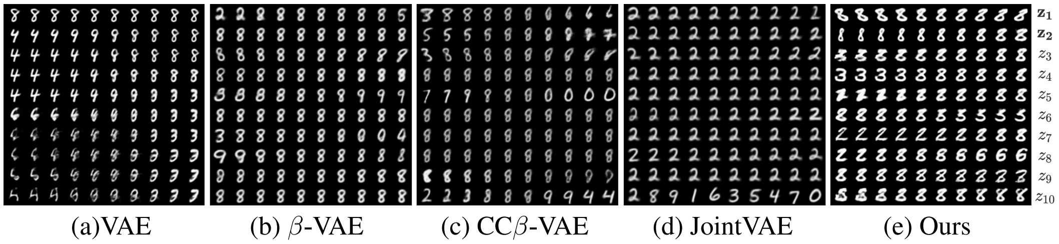
4 EXPERIMENTS

In this section, we first present qualitative results demonstrating our proposed unsupervised Guided-VAE (Figure 1a) capable of disentangling latent embedding in a more favourable way than VAE and previous disentangle methods (Higgins et al., 2017; Dupont, 2018) on MNIST dataset (LeCun et al., 2010). We also show that our learned latent representation can be later used to improve classification performance. Next, we extend this idea to a supervised guidance approach in an adversarial excitation and inhibition fashion, where a discriminative objective for certain image properties is given (Figure 1b) on the CelebA dataset (Yang et al., 2015). Further, we show that our method can be applied to the few-shot classification tasks, which achieves competitive performance on Omniglot dataset proposed by Vinyals et al. (2016).

4.1 Unsupervised Guided-VAE

4.1.1 Qualitative evaluation

We present qualitative results on MNIST dataset by traversing latent variables received affine transformation guiding signal. Here, we applied the Guided-VAE with the bottleneck size of 10 (i.e. the latent variables $\mathbf{z} \in \mathbb{R}^{10}$). The first latent variable z_1 represents the rotation information and the second latent variable z_2 represents the scaling information. The rest of the latent variables $\mathbf{z}_{3:10}$ represent the content information. Thus, the latent variables $\mathbf{z} \in \mathbb{R}^{10}$ are represented by $\mathbf{z} = (\mathbf{z}_{def}, \mathbf{z}_{cont}) = (\mathbf{z}_{1:2}, \mathbf{z}_{3:10})$.



(a)VAE (b) β-VAE (c) CCβ-VAE (d) JointVAE (e) Ours

Figure 2: **Latent Variables Traversal on MNIST:** Comparison of traversal results from vanilla VAE (Kingma & Welling, 2014), β-VAE (Higgins et al., 2017), β-VAE with controlled capacity increase (CCβ-VAE), JointVAE (Dupont, 2018) and our Guided-VAE on the MNIST dataset. z_1 and z_2 in Guided-VAE are controlled.

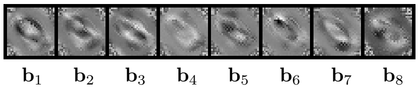


Figure 3: PCA basis learned by the secondary decoder in unsupervised Guided-VAE.

In Figure 2, we show traversal results of all latent variables on MNIST dataset for vanilla VAE (Kingma & Welling, 2014), β-VAE (Higgins et al., 2017), JointVAE (Dupont, 2018) and our guided VAE (β-VAE, JointVAE results are adopted from (Dupont, 2018)). While β-VAE cannot generate meaningful disentangled representations, even with controlled capacity increased, JointVAE is able to disentangle class type from continuous factors. Different from previous methods, our Guided-VAE disentangles geometry properties (z_1 and z_2) like rotation angle and stroke thickness from the rest content information $\mathbf{z}_{3:10}$.

In Figure 3, we visualize the basis $B = (\mathbf{b}_1, ..., \mathbf{b}_8)$ in the PCA part of Dec$_{sub}$. The basis primarily capture the content information.

4.1.2 Quantitative evaluation

For a quantitative evaluation, we first compare the reconstruction error among different models on the MNIST dataset. In this experiment, we set the bottleneck size to 8 in Guided-VAE and use three settings for the deformation/transformation: Rotation, scaling, and both. In Guided-VAE (Rotation) or Guided-VAE (Scaling), we take the first latent variable z_1 to represent the rotation or the scaling information. In Guided-VAE (Rotation and Scaling), we use the first and second latent variables (z_1 and z_2) to represent rotation and scaling respectively. As Table 1 shows, our reconstruction loss is on par with vanilla VAE, whereas the previous disentangling method (β-VAE) has higher loss. Our proposed method is able to achieve added disentanglement while not sacrificing reconstruction capability over vanilla VAE.

In addition, we perform classification tasks on latent embeddings of different models. Specifically, for each data point $(\mathbf{x}, y)$, we use the pre-trained VAE model to obtain the value of latent variable

Table 1: Reconstruction loss on MNIST digits data.

Model	Reconstruction Loss
Vanilla VAE	84.4
β-VAE (Beta=2)	89.2
β-VAE (Beta=4)	100.1
Guided-VAE (Rotation)	86.6
Guided-VAE (Scaling)	85.8
Guided-VAE (Rotation, Scaling)	84.3

$\mathbf{z}$ given input image $\mathbf{x}$. Here $\mathbf{z}$ is a $d_{\mathbf{z}}$-dim vector. We then train a linear classifier $f(\cdot)$ on the embedding-label pairs $\{(\mathbf{z}, y)\}$ in order to predict the class of digits. For the Guided-VAE, we disentangle the latent variables $\mathbf{z}$ into deformation variables $\mathbf{z}_{def}$ and content variables $\mathbf{z}_{cont}$ with same dimensions (i.e. $d_{\mathbf{z}_{def}} = d_{\mathbf{z}_{cont}}$) and use affine transformation as $\tau(\mathbf{z}_{def})$. We compare the classification errors of different models under multiple choices of dimensions of the latent variables in Table 2. It shows that generally higher dimensional latent variables result in lower classification errors. Our Guided-VAE method compares favourably over vanilla VAE and β-VAE.

Table 2: Classification error over different methods.

Model	$d_{\mathbf{z}} = 16$	$d_{\mathbf{z}} = 32$	$d_{\mathbf{z}} = 64$
Vanilla VAE	2.85%	2.63%	2.87%
β-VAE (β=2)	4.70%	5.10%	5.23%
Guided-VAE (Ours)	**2.17%**	**1.51%**	**1.42%**

Moreover, we attempt to validate the effectiveness of disentanglement in Guided-VAE. We follow the same classification tasks above but use different parts of latent variables as input features for the classifier $f(\cdot)$: We may choose the deformation variables $\mathbf{z}_{def}$, the content variables $\mathbf{z}_{cont}$, or the whole latent variables $\mathbf{z}$ as the input feature vector. To reach a fair comparison, we keep the same dimensions for the deformation variables $\mathbf{z}_{def}$ and the content variables $\mathbf{z}_{cont}$. Table 3 shows that the classification errors on $\mathbf{z}_{cont}$ are significantly lower than the ones on $\mathbf{z}_{def}$, which indicates the success of disentanglement since the content variables should determine the class of digits while the deformation variables should be invariant to the class. In addition, when the dimensions of latent variables $\mathbf{z}$ are higher, the classification errors on $\mathbf{z}_{def}$ increase while the ones on $\mathbf{z}_{cont}$ decrease, indicating a better disentanglement between deformation and content.

Table 3: Classification error on different latent variables. [↑ means higher is better, ↓ means lower is better]

Model	$d_{\mathbf{z}_{def}}$	$d_{\mathbf{z}_{cont}}$	$d_{\mathbf{z}}$	$\mathbf{z}_{def}$ $Error \uparrow$	$\mathbf{z}_{cont}$ $Error \downarrow$	$\mathbf{z}$ $Error \downarrow$
Guided-VAE	8	8	16	27.1%	3.69%	**2.17%**
	16	16	32	42.07%	1.79%	**1.51%**
	32	32	64	62.94%	1.55%	**1.42%**

4.2 Supervised Guided-VAE

4.2.1 Qualitative evaluation

We first present qualitative results on the CelebA dataset by traversing latent variables of attributes. We select three labeled attributes (emotion, gender and color) in the CelebA dataset as supervised guidance objectives. The bottleneck size is set to 16. We use the first three latent variables z_1, z_2, z_3 to represent the attribute information and the rest $\mathbf{z}_{4:16}$ to represent the content information. During evaluation, we choose $z_t \in \{z_1, z_2, z_3\}$ while keeping the remaining latent variables $\mathbf{z}_t^{rst}$ fixed. Then we obtain a set of images through traversing from the image with t-th attribute to the image without t-th attribute (e.g. smiling to non-smiling) and compare them over methods.

Figure 4 shows the traversal results for β-VAE and our Guided-VAE. β-VAE performs decently for the controlled attribute change, but the individual $\mathbf{z}$ in β-VAE is not fully entangled or disentangled with the attribute. Guided-VAE has a better disentanglement for latent variables and is able to better isolate the attributes w.r.t. the corresponding latent variables.

4.2.2 Quantitative evaluation

In supervised Guided-VAE, we train a classifier to predict the attributes by using the disentangled attribute latent variable z_t or the rest of latent variables $\mathbf{z}_t^{rst}$ as input features. We perform adversarial excitation and inhibition by encouraging the target latent variable to best predict the corresponding t-th attribute and discouraging the rest of the variables for the prediction of that attribute. Figure 5

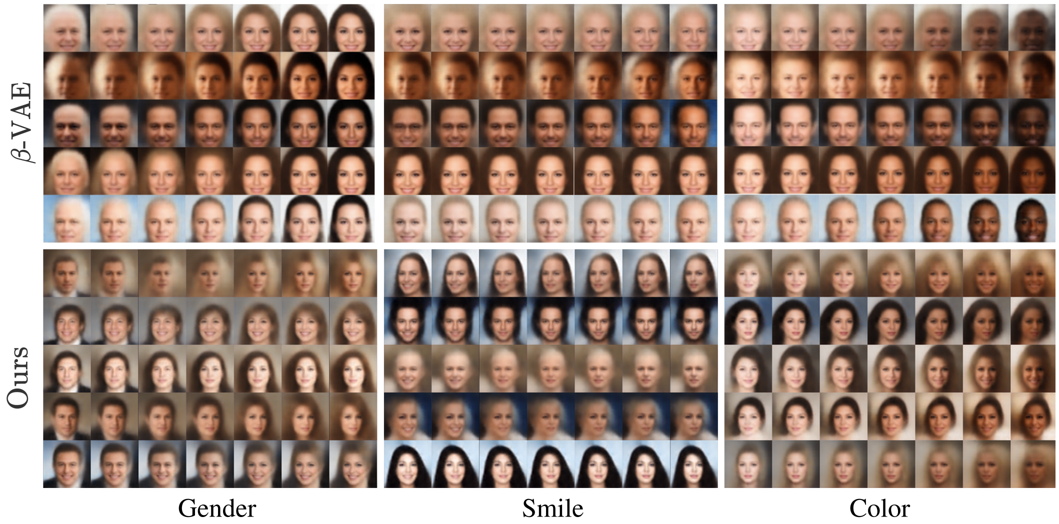


Figure 4: **Traversal of Latent Factors learned on CelebA:** Column 1 shows the traversed images from no-smiling to smiling. Column 2 shows the traversed images from male to female. Column 3 transits the color. The first row is from (Higgins et al., 2017) and we follow its figure generation procedure.

(left) shows that the classification errors on z_t is significantly lower than the ones on $\mathbf{z}_t^{rst}$, which indicates the effectiveness of disentanglement during the training procedure.

Model	**Latent Variable**	**Classification Error**
GENDER	z_1	$3.57\% \downarrow$
	$\mathbf{z}_1^{rst}$	$24.51\% \uparrow$
SMILE	z_2	$8.85\% \downarrow$
	$\mathbf{z}_2^{rst}$	$32.81\% \uparrow$

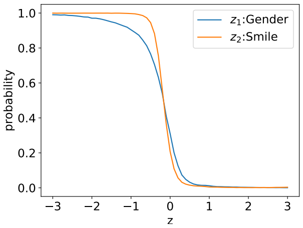


Figure 5: (left) Classification error on CelebA training set. (right) Experts (high-performance external classifiers for attribute classification) prediction for being negatives on the generated images. We traverse z_1 (gender) and z_2 (smile) separately to generate images for the classification test. Each latent z is traversed from -3.0 to 3.0 with 0.1 as the stride length.

Furthermore, we attempt to validate that the generated images from the supervised Guided-VAE can be actually controlled by the disentangled attribute variables. Thus, we pre-train an external binary classifier for t-th attribute on the CelebA training set and then use this classifier to test the generated images from Guided-VAE. Each test includes $10,000$ generated images randomly sampled on all latent variables except for the particular latent variable z_t we decide to control. As Figure 5 (right) shows, we can draw the confidence-z curves of the t-th attribute where $z = z_t \in [-3.0, 3.0]$. For the gender and the smile attributes, it can be seen that the corresponding z_t is able to enable ($z_t < -1$) and disable ($z_t > 1$) the attribute of the generated image. Besides, for all the attributes, the probability monotonically decreases when z_t increases, which shows the controlling ability of the t-th attribute by tuning the corresponding latent variable z_t.

4.3 FEW-SHOT LEARNING

Previously, we have shown that Guided-VAE can generate images and be used as representation to perform classification task. In this section, we will apply the proposed method to few-shot classification problem. Specifically, we use our adversarial excitation and inhibition method in the Neural Statistician (Edwards & Storkey, 2017) by adding a supervised guidance network after the statistic network. The supervised guidance signal is the label of each input. We also apply the Mixup method (Zhang et al., 2018) in the supervised guidance network. However, we couldn't reproduce exact reported results in the Neural Statistician, which is also indicated in Korshunova et al. (2018). For comparison, we mainly consider the Matching Nets (Vinyals et al., 2016) and Bruno (Korshunova et al., 2018). Yet it cannot outperform Matching Nets, our proposed Guided-VAE reaches equivalent performance as Bruno (discriminative), where a discriminative objective is fine-tuned to maximize the likelihood of correct labels.

Table 4: Classification accuracy for a few-shot learning task on the Omniglot dataset.

Omniglot	5-way 1-shot	5-way 5-shot	20-way 1-shot	20-way 5-shot
PIXELS	41.7%	63.2%	26.7%	42.6%
BASELINE CLASSIFIER	80.0%	95.0%	69.5%	89.1%
MATCHING NETS	98.1%	98.9%	93.8%	98.5%
BRUNO	86.3%	95.6%	69.2%	87.7%
BRUNO (DISCRIMINATIVE)	97.1%	99.4%	91.3%	97.8%
BASELINE	97.7%	99.4%	91.4%	96.4%
OURS (DISCRIMINATIVE)	97.8%	99.4%	92.1%	96.6%

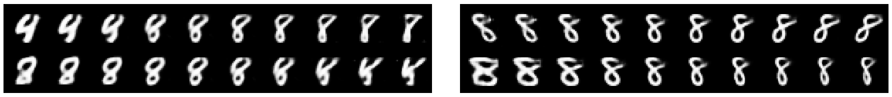
5 Ablation Study

We conduct a series of ablation experiments to validate our proposed Guided-VAE model.

5.1 Geometric Transformations

In this part, we conduct an experiment by excluding the geometry-guided part from the unsupervised Guided-VAE. In this way, the nudging decoder is just a PCA-like decoder but not a deformable PCA.

The setting of this experiment is exactly same as described in the unsupervised Guided-VAE section. The bottleneck size of our model is set to 10 of which the first two latent variables z_1, z_2 represent the rotation and scaling information separately. In the ablation part, we drop off the geometry-guided part so all 10 latent variables are controlled by the PCA-like light decoder.



(a) Unsupervised Guided-VAE without geometry transformation

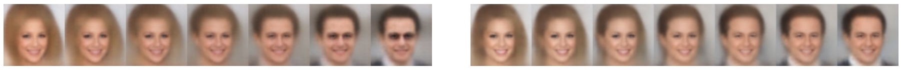
(b) Unsupervised Guided-VAE

Figure 6: **Ablation study on geometry transformation:** The results here are traversed on z_1 and z_2. Compared to (b), (a) presents little about the rotation and scaling information.

5.2 Adversarial Excitation and Inhibition

In this part, we conduct an experiment of using the adversarial excitation method. We design the experiment using the exact same setting described in the supervised Guided-VAE part.

As Figure 7 shows, though the traversal results still show the traversed results on some latent variables. The results from the adversarial excitation method outperforms the results from the discriminative method. While traversing the latent variable controlling the smiling information, the left part (a) also changes in the smiling status but it's controlled by another latent variable.



(a) Supervised Guided-VAE without Inhibition

(b) Supervised Guided-VAE

Figure 7: **Ablation study on the adversarial excitation and inhibition method for gender:** The left part shows the traversed images from the supervised Guided-VAE without adversarial inhibition. The right part shows the traversed images from the supervised Guided-VAE using adversarial excitation and inhibition. Both images are traversed on the latent variable that is supposed to control the gender information.

6 Conclusion

In this paper we have presented a new representation learning method, guided variational autoencoder (Guided-VAE), for disentanglement learning. Both versions of Guided-VAE utilize lightweight guidance to the latent variables to achieve better controllability and transparency. Improvements on disentanglement, image traversal, and meta-learning over the competing methods are observed. Guided-VAE maintains the backbone of VAE and can be applied to other generative modeling applications.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *ICML*, 2018.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, 2019.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*, pp. 710–720, 2018.

Harrison Edwards and Amos Storkey. Towards a neural statistician. *ICLR*, 2017.

Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. In *ICLR*, 2018.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, pp. 1287–1298, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT Press, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pp. 3–10, 1994.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psycholog*, 24, 1933.

Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *CVPR*, 2018.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *ICLR Worshop Track*, 2019.

Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *European Conference on Computer Vision*, pp. 829–845. Springer, 2018.

Ian Jolliffe. Principal component analysis. *Springer Berlin Heidelberg*, 2011.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

Iryna Korshunova, Jonas Degrave, Ferenc Huszár, Yarin Gal, Arthur Gretton, and Joni Dambre. Bruno: A deep recurrent model for exchangeable data. In *Advances in Neural Information Processing Systems*, pp. 7190–7198, 2018.

Yann LeCun. *Modeles connexionnistes de lapprentissage*. PhD thesis, PhD thesis, These de Doctorat, Universite Paris 6, 1987.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2:18, 2010.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570, 2015.

Jianxin Lin, Yingce Xia, Sen Liu, Tao Qin, Zhibo Chen, and Jiebo Luo. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. *arXiv:1902.03782*, 2019.

Yen-Cheng Liu, Yu-Ying Yeh, Tzu-Chien Fu, Sheng-De Wang, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Detach and adapt: Learning cross-domain disentangled deep representation. In *CVPR*, 2018.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. In *ICLR Workshop Track*, 2016.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5040–5048, 2016.

Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 1623–1632, 2017.

Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2233–2246, 2012.

Christopher Poultney, Sumit Chopra, Yann LeCun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pp. 1137–1144, 2007.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. In *ICLR Workshop Track*, 2018.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Michael P Wellman and Max Henrion. Explaining'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, 1993.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3676–3684, 2015.

Ofer Yizhar, Lief E Fenno, Matthias Prigge, Franziska Schneider, Thomas J Davidson, Daniel J Oshea, Vikaas S Sohal, Inbal Goshen, Joel Finkelstein, Jeanne T Paz, et al. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, 477(7363):171, 2011.

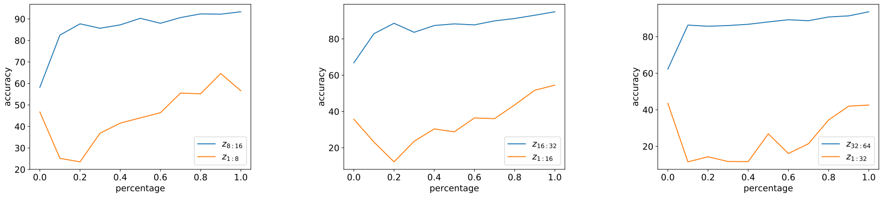
Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2018.

A APPENDIX

A.1 PERCENTAGE OF DATA PARTICIPATING IN THE GUIDED SUB-NETWORK

In this part, we design an experiment to show how the percentage of data participating in the guided sub-network can influence the final prediction. We conduct this ablation study on MNIST using unsupervised Guided-VAE. We change the percentage of data participating in the guided sub-network and then present the classification accuracy using the first half latent variables (represent geometry information) and the second half latent variables (represent content information) separately.

From Figure 8, we observe consistent improvement for the last half latent variables when adding more samples to guide sub-network. This indicates adding more samples can improve disentanglement, which causes that more content information is represented in the second half latent variables. Similarity, the improvement of disentanglement leads the first half latent variables can represent more geometry information, which is indiscriminative for classes. We also observe accuracy improvement when large amount of samples are used to train sub-network. We hypothesize this is because geometry information is still partially affected by classes.



(a) $z_{dim} = 16$

(b) $z_{dim} = 32$

(c) $z_{dim} = 64$

Figure 8: **Study on changing the percentage of the data participating in the guided sub-network:**The three figures present the accuracy from the unsupervised Guided-VAE of which the bottleneck size is 16, 32, 64 separately.