[5] β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

Mohamed Mohamed
McGill University
Montreal, QC, Canada
mohamed.mohamed5@mail.mcgill.ca

Abstract

Humans excel at recognizing and distinguishing objects based on attributes such as size, texture, colour, and features, often from limited data. To enable artificial intelligence (AI) to reason similarly, it must learn to uncover the underlying generative factors of data without supervision [2]. This capability is essential for building AI systems that can generalize across tasks and domains. A critical step toward this goal is learning disentangled representations, where individual latent dimensions correspond to distinct generative factors and are invariant to changes in others. While the Variational Autoencoder (VAE) framework [7] introduced the ability to regulate latent spaces for generative modeling, it struggled to achieve disentanglement. The β-VAE framework addressed this limitation by introducing the β hyperparameter, which encourages disentanglement by controlling the trade-off between reconstruction fidelity and independence of latent factors. Increasing β reduces the overlap between generative factors, promoting more interpretable latent spaces. In this paper, we revisit the theoretical contributions of the β-VAE, providing an in-depth analysis of disentanglement through the lens of information theory, as outlined in [3]. Furthermore, we solidify the experimental evaluation of β-VAE to demonstrate its disentanglement properties in various scenarios. We also critique the existing disentanglement metrics, arguing that they inadequately capture feature independence, and propose alternatives to address these shortcomings. Our code reproduction is available at https://github.com/lesupermomo/B-VAE.

1. Introduction

Previous attempts at learning disentangled representations often relied on prior knowledge about the number or nature of the generative factors underlying the data [9]. However, such assumptions are impractical in real-world scenarios, where learners are often exposed to complex and diverse datasets without any prior information about the generative factors of the data. Moreover, in many cases, little to no supervision is available to guide the discovery of these factors. While unsupervised methods have demonstrated potential for disentangled factor learning, they have faced challenges in scaling effectively to more complex data distributions.

Disentangled representations offer several advantages in generalization and transfer learning. A disentangled representation ensures that knowledge about one factor can generalize to new configurations of other factors. This property has been identified as a key component for improving the capabilities of AI systems in areas where humans excel but current AI struggles. Examples include knowledge transfer, where learned representations are reused to accelerate learning across multiple tasks; zero-shot inference, where AI reasons about novel data by recombining learned factors; and novelty detection, where unfamiliar configurations can be recognized as distinct [10].

Before the β-VAE framework [5], most attempts at learning disentangled representations required assumptions about the number or type of generative factors [6, 8, 13–15]. These methods were limited by their reliance on such a priori knowledge, making them unsuitable for applications where generative factors are unknown or cannot be explicitly defined. In contrast, β-VAE introduces an unsupervised approach that facilitates the discovery of disentangled factors without these restrictive assumptions, paving the way for broader applicability in real-world scenarios.

Although InfoGAN [4] introduced a scalable approach to disentangled representation learning, it faced significant limitations. These included training instability and reduced sample diversity. Furthermore, InfoGAN relied on some prior knowledge of the data, as its performance heavily depended on the choice of the prior distribution and the number of regularized noise latents, making it less practical for many real-world scenarios.

The β-VAE framework [5] proposed a novel unsupervised method for learning disentangled representations of independent generative factors in visual data. By extend-

ing the original VAE framework, it introduced a single hyperparameter, β, which modulates the learning constraints applied to the model. These constraints serve two key purposes: limiting the capacity of the latent information channel and encouraging the learning of statistically independent latent factors. When $\beta = 1$, the framework is equivalent to the original VAE [7]. However, increasing β beyond 1 drives the model to prioritize disentanglement, provided that the data contains independent underlying factors of variation.

The β-VAE paper demonstrated state-of-the-art performance in disentangling generative factors both qualitatively and quantitatively. It outperformed leading unsupervised approaches like InfoGAN [4] and semi-supervised methods like Deep Convolutional Inverse Graphics Network (DC-IGN) [8] across multiple benchmarks. Notably, it achieved superior results on datasets such as CelebA [11], Chairs [1], and 3D Faces [12], as evidenced by qualitative evaluations. Their results showcased the framework's ability to generalize and disentangle latent representations in complex data distributions.

Despite its significant contributions, the work of [5] does not fully address why the factorized representations learned by β-VAE tend to align with human intuitions about the data's generative factors, especially when compared to the standard VAE [7]. This lack of theoretical clarity leaves open questions about the underlying mechanisms driving the disentanglement observed in β-VAE. Additionally, β-VAE suffers from limitations such as reduced reconstruction fidelity compared to the standard VAE. This trade-off arises due to its modified training objective, which prioritizes disentanglement at the cost of reconstruction quality by imposing stricter constraints on the latent information bottleneck.

The work of [3] provides valuable insights into the disentangling behavior of β-VAE. By analyzing the relationship between the information bottleneck and latent disentanglement, the authors proposed practical improvements to the original framework. Their extension relaxes the information bottleneck during training, allowing the model to balance reconstruction accuracy and disentanglement more effectively. This modification enables the β-VAE to achieve more robust disentangling performance while significantly improving reconstruction fidelity.

In addition to its methodological contributions, [5] introduced a novel metric for quantifying disentanglement. This measure demonstrated that β-VAE substantially outperforms prior approaches, including ICA, PCA, the original VAE [7], DC-IGN [8], and InfoGAN [4], on multiple benchmarks. These results solidify β-VAE's position as a leading framework for unsupervised disentangled representation learning.

In this project, we implement the VAE and β-VAE models from scratch and reproduce their results on the Faces [12], Chairs [1], and dSprites [5] datasets. Our primary focus is on replicating their experiments involving the manipulation of latent factors and evaluating the disentanglement properties of each model.

We demonstrate that while the β-VAE model promotes disentanglement, it faces significant challenges with reconstruction quality as the β parameter increases. To address this issue, we explore the objective proposed in [3], which achieves a compact latent space with robust disentanglement while maintaining reconstruction fidelity comparable to the baseline VAE model.

Additionally, we critique the disentanglement metric introduced in [5], identifying its limitations. Although the metric captures disentanglement properties to some extent, it falls short in assessing true independence and poses reproducibility challenges. Specifically, the metric assumes prior knowledge of data generative factors and requires access to a dataset generator—conditions that are often impractical in real-world applications.

To overcome these shortcomings, we explore alternative metrics introduced in subsequent research, showcasing their ability to more effectively evaluate disentanglement. By comparing these metrics, we provide a comprehensive analysis of disentanglement properties across models and metrics, further advancing our understanding of this critical aspect of representation learning.

2. Theory

2.1. Variational Autoencoder (VAE)

Consider a dataset x consisting of samples drawn from a distribution parameterized by ground-truth generative factors z. The Variational Autoencoder (VAE) [7] aims to model the marginal likelihood of the data through the generative process by optimizing the following objective:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big],$$

where ϕ and θ parameterize the encoder and decoder distributions, respectively. This expression can be re-written using the evidence lower bound (ELBO) decomposition as:

$$\log p_\theta(x) = D_{\mathrm{KL}}\big(q_\phi(z|x) \,\|\, p(z)\big) + \mathcal{L}(\theta, \phi; x, z),$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler (KL) divergence between the approximate posterior $q_\phi(z|x)$ and the prior $p(z)$. Since the KL divergence is always non-negative ($D_{\mathrm{KL}} \geq 0$), maximizing the lower bound $\mathcal{L}(\theta, \phi; x, z)$ directly aligns with optimizing the VAE objective:

$$\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x, z)$$

$\mathcal{L}(\theta, \phi; x, z) = \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] - D_{\mathrm{KL}}\big(q_\phi(z|x) \,\|\, p(z)\big)$.

To make optimization of this objective practical, several assumptions are typically made: 1. Gaussian Prior and Posterior: The prior $p(z)$ is usually modeled as an isotropic Gaussian, $\mathcal{N}(0, 1)$, and the approximate posterior $q_\phi(z|x)$ is parameterized as a Gaussian with diagonal covariance, $\mathcal{N}(\mu, \sigma^2)$. 2. Reparameterization Trick: The latent variables z are expressed as a differentiable transformation of a noise variable $\epsilon \sim \mathcal{N}(0, 1)$. Specifically, each latent variable z_i is reparameterized as:

$$z_i = \mu_i + \sigma_i \cdot \epsilon,$$

where μ_i and σ_i are outputs of the encoder network.

This reparameterization enables efficient backpropagation by allowing gradients to flow through the sampling process, making it feasible to optimize the ELBO directly with respect to the parameters ϕ and θ.

2.2. β-VAE

The β-VAE is an extension of the Variational Autoencoder (VAE) framework [7] that introduces a tunable hyperparameter β into the original VAE objective. The modified objective function is expressed as:

$$\mathcal{L}(\theta, \phi; x, z, \beta) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta\, D_{\mathrm{KL}}\left(q_\phi(z|x) \,\|\, p(z)\right),$$

where β controls the trade-off between reconstruction quality and the degree of disentanglement in the latent space.

When $\beta = 1$, the β-VAE objective reduces to the standard VAE objective. For values of $\beta > 1$, the increased weight on the KL divergence term imposes a stronger constraint for the approximate posterior $q_\phi(z|x)$ to align with the factorized unit Gaussian prior $p(z)$. This constraint effectively reduces the latent space capacity z, encouraging the model to learn more disentangled and independent latent representations.

A key benefit of increasing β is the promotion of disentanglement, where the learned latent variables represent independent factors of variation in the data. However, this comes with a trade-off: as β increases, the model's ability to reconstruct the input data x with high fidelity often deteriorates. The model faces pressure to create disentangled representations, which can lead to a loss of high-frequency details in the data. This reconstruction-disentanglement trade-off is a known limitation of the β-VAE framework, as discussed in [3].

The optimal β value is dataset and task-dependent, and finding the right balance between disentanglement and reconstruction quality is crucial. While we cannot directly optimize for the optimal value of β, we can estimate it using a disentanglement metric or through visual inspection heuristics. Additionally, optimizing β also influences the choice of other model parameters, contributing to the best disentanglement for a given task.

2.3. Learning the Latent Representation in β-VAE

The work by [3] frames the β-VAE objective as an information bottleneck for the reconstruction task, where the posterior distribution $q(z|x)$ is encouraged to efficiently transmit data information by minimizing the weighted KL divergence and maximizing the data log likelihood. The posterior is constrained to match a unit Gaussian prior, and both the prior and posterior are factorized. Through the reparametrization trick, samples from the posterior are generated by adding Gaussian noise to a deterministic encoder mean. The KL divergence in the objective serves as an upper bound on the amount of information that can be transmitted through latent channels, which increases when the posterior means are dispersed or when posterior variances are reduced. The reconstruction process under this bottleneck encourages data points to be embedded in latent space such that nearby points in data space are also close in latent space. While reducing the KL divergence can be achieved by squeezing the posterior distributions, this increases their overlap and reduces their discriminability, which can negatively affect the log likelihood. Therefore, aligning nearby data points in latent space helps minimize the log likelihood cost, facilitating more efficient reconstruction [3].

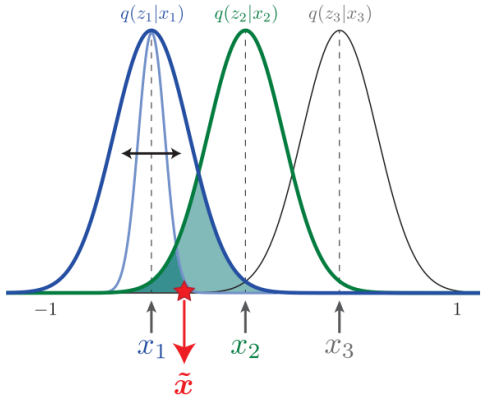

Figure 1. [3] The relationship between posterior overlap, KL divergence, and reconstruction error. A datapoint $\tilde{x}$ sampled from the distribution $q(z_2|x_2)$ is more likely to be confused with a sample from $q(z_1|x_1)$ as the overlap between them increases.

As illustrated in 1 taken from [3], broadening the posterior distributions or bringing their means closer together reduces the KL divergence and increases overlap. However, this overlap introduces confusion between data points from different distributions. Ensuring that neighboring points in data space are also close together in latent space minimizes

the log likelihood cost caused by this confusion.

2.4. β-VAE with controlled capacity increase

In [3], the authors propose a method for controlled capacity increase in β-VAE training to balance the trade-off between latent representation learning and reconstruction quality. Early in training, the capacity of the latent space is kept low to encourage disentangled and meaningful representations. Over time, this capacity is progressively increased, prioritizing reconstruction accuracy in the later stages of training. The capacity constraint is applied through the KL divergence term, D_{KL}, which is encouraged to approximate a controllable target value, C, by penalizing deviations from C using a hyperparameter γ. The training objective combines maximizing the log-likelihood of image reconstructions and minimizing the absolute deviation of D_{KL} from C. This approach allows the model to learn progressively richer representations while maintaining disentanglement during capacity growth.The training objective with controlled capacity increase is given by:

$$L(\theta, \phi; x, z, C) = \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] - \gamma\big|D_{\mathrm{KL}}\big(q_\phi(z|x)\|p(z)\big) - C\big|.$$

To implement this, the capacity C is linearly increased from a low value (e.g., 0.5 nats) to a high value (e.g., 25.0 nats) over the course of training. At low capacities, the model prioritizes only a subset of latent factors, such as the X and Y positional factors, resulting in blurry reconstructions that primarily capture object positions. As C increases, additional factors like scale and other attributes start contributing to the KL divergence, leading to progressively sharper and more detailed reconstructions. This incremental allocation of latent encoding capacity ensures that the model learns to represent new factors of variation without compromising the disentanglement of previously learned ones. By integrating this capacity control principle into β-VAE, the authors demonstrate how to enhance the model's ability to disentangle and reconstruct data effectively.

3. Results

We begin by reproducing the results of [5] on the dSprites dataset. The dSprites dataset, introduced in [5], is a synthetic dataset comprising 737,280 binary 2D shapes (heart, oval, and square) generated as the Cartesian product of shape and four independent generative factors defined in vector graphics: position X (32 values), position Y (32 values), scale (6 values), and rotation (40 values spanning the range $[0, 2\pi]$). The dataset ensures smooth affine object transformations, with consecutive values of each factor selected to minimize pixel-space differences at a resolution of 64×64 pixels. This dataset is particularly suited for disentanglement studies as it eliminates confounding factors, containing only five independent generative factors: shape identity, position X, position Y, scale, and rotation. This provides a ground truth for evaluating and objectively comparing the disentangling performance of different models.

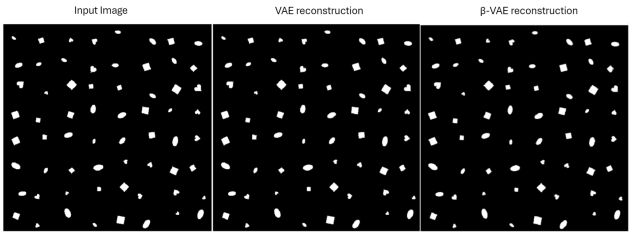

Figure 2. Model Reconstructions. The input image (on the left) consists of 8x8 different images given as input to the VAE model and the β-VAE model with $\beta = 4$. The figure shows that β-VAE reconstructions are not in par with the VAE reconstructions. More specifically, this can be noticed when looking at the small heart shapes.

In [5], the authors employed three distinct architectures for their β-VAE model, tailored to the datasets used. Based on their reported results, we infer that the largest architecture, comprising 12,749,036 parameters, was used for the dSprites dataset in order to obtain great disentanglement properties due to the increased model capacity. For comparison, the model architecture used for the 3D Chairs dataset had only 542,849 parameters, making the dSprites model approximately 23 times larger in capacity. In our experiments, we employed the architecture they used for the 3D Chairs dataset, which has significantly fewer parameters and requires approximately 500 MB of memory. This difference in model capacity likely explains why our reproduced results fall short of those reported in the original paper. The smaller model's reduced capacity may limit its ability to capture the full complexity of the disentanglement task on the dSprites dataset.

The VAE plots highlight the inherent trade-off between disentanglement and reconstruction quality. Among all the models tested in our experiments, the VAE achieved the best reconstruction loss, approximately 12 in Mean Squared Error (MSE) (see Figure 6 and Table 1) . However, this superior reconstruction accuracy comes at the expense of a fully entangled latent space, as it utilizes all available latent dimensions (see Figure 7). The KL divergence plot in Figure 7 demonstrates that every latent dimension contributes significantly to the KL loss. Ideally, for the dsprites dataset, there are only five true generative factors (as explained earlier), yet the VAE fails to constrain its latent space accordingly, leading to suboptimal disentanglement.

Figure 2 compares the reconstruction performance of the VAE and β-VAE (with $\beta = 4$). While the β-VAE's recon-

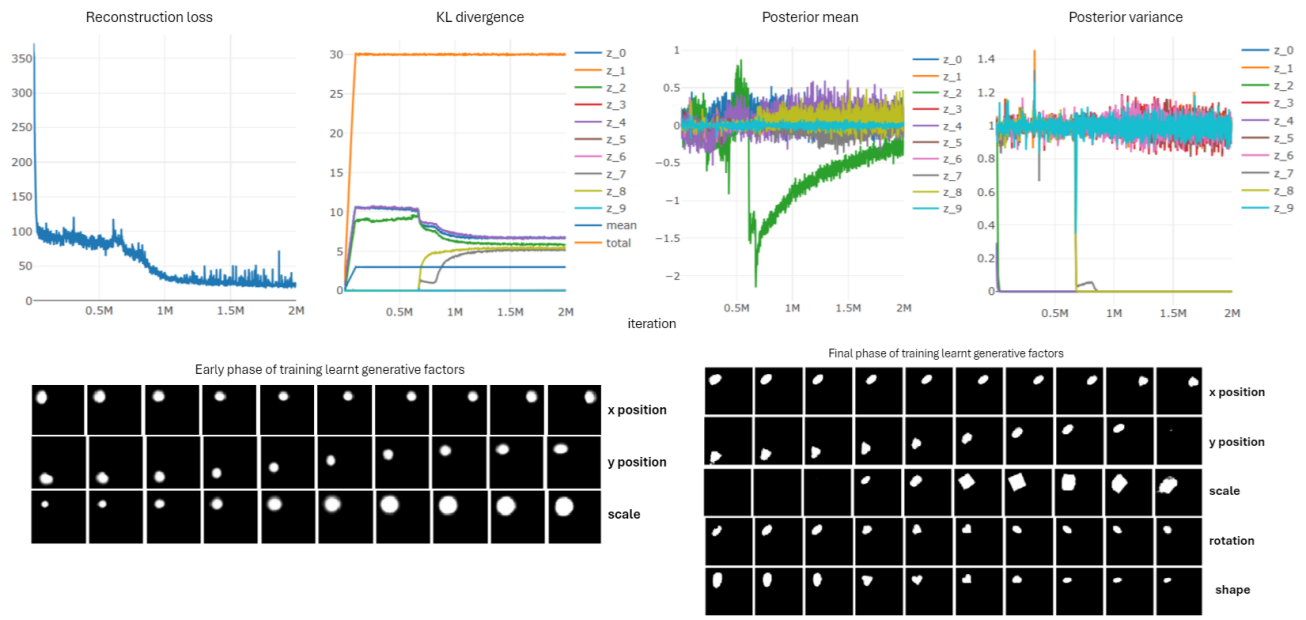

Figure 3. Training results of the β-VAE model from [3] with $\gamma = 100$ and a controlled capacity increase up to $C = 30$, trained over 2 million iterations. The plots display the following: (a) reconstruction loss, (b) KL divergence for each latent dimension, (c) posterior means for each latent dimension, and (d) posterior variances for each latent dimension. These results demonstrate that, early in training, the model successfully learns to disentangle and represent generative factors related to the X and Y positions, as well as scale. Towards the middle of the training process, the model begins to capture the shape and rotation factors, though the latent representations start to become increasingly entangled.

Model	Reconstruction Loss (MSE)
VAE	**11.28**
β-VAE [3]	17.44
β-VAE [5]	23.77

Table 1. Reconstruction loss (MSE) for different models on the dsprites dataset. The VAE achieves the lowest reconstruction loss, followed by the β-VAE [3] model with $\gamma = 100$ and $C_{max} = 30$. The β-VAE [5] model, originally used for the 3DChairs dataset with $\beta = 4$, shows the highest reconstruction loss.

structions are slightly less accurate than those of the VAE, particularly in fine details, such as the small heart shapes, the difference is minimal. This reduction in reconstruction quality is attributable to the β-VAE placing greater emphasis on the KL divergence term in its loss function, which promotes disentanglement at the cost of reconstruction fidelity.

From a detailed analysis of the latent space, as shown in Figure 7 for the VAE where all latent parameters contribute to the KL divergence, and by varying each latent parameter to observe its corresponding generative factor (see Figure 9), it is evident that the VAE model does not learn a disentangled latent space. Instead, its latent space is entangled, with multiple generative factors encoded within each latent parameter. This entanglement likely results from the model's strong emphasis on achieving good reconstruction quality, which it accomplishes through non-interpretable latent representations. This behavior stands in contrast to the results observed in the β-VAE models.

From the KL divergence values of each latent parameter during training (see Figure 8) of the β-VAE, we observe that some latent parameters, such as z_0, z_5, z_6, and z_9, are relatively unused. However, even with the β-VAE, there are still more active latent parameters than the number of generative factors in the dataset. This indicates that the learned latent space is not fully disentangled, which may be attributed to several factors, including the limited capacity of our model or the choice of the prior distribution used during training.

3.1. The role of the prior and model capacity

The prior distribution and the model capacity play a significant role in shaping the latent space. In our experiments, we used a Gaussian prior, and convolutional neural networks following the approach in [5] for the 3DChairs dataset. However, in the original β-VAE paper by [5], a Bernoulli prior was employed instead on the dsprites dataset. This difference in prior distributions likely has a significant impact on the ability of the model to disentangle the latent space, as the Bernoulli prior may encourage

a more discrete and separable encoding of generative factors compared to the continuous nature of a Gaussian prior. Another critical difference lies in the model architectures used in these studies.The original β-VAE paper by [5] employs a fully connected neural network architecture with 23 times more parameters than the CNN used for their 3D chairs dataset. This significantly larger model capacity in [5] likely contributes to its ability to better encode and disentangle generative factors, albeit at the cost of higher computational and memory requirements. The choice of network architecture and prior distribution together plays a pivotal role in determining the balance between reconstruction quality and disentanglement.

3.2. β-VAE with Controlled Capacity Increase Results

The training procedure of the β-VAE with controlled capacity increase, as shown in Figure 3, demonstrates that the model learns the generative factors at different stages of training. The model was trained with $\gamma = 100$ and a maximum capacity $C_{\text{max}} = 30$. The choice of $C_{\text{max}} = 30$ was motivated by the observation from the VAE training process (Figure 7), where the largest KL divergence reached was approximately 30. To ensure that the β-VAE model maintains a similar reconstruction loss, we restricted the KL divergence to not exceed this value during training.

Following the approach in [3], the γ term was set large enough to keep the KL divergence close to the target value C, which was linearly increased from 0 to 30 nats over 100,000 iterations. When we initially used a much larger γ value (1000) as suggested in the paper [3], we observed that the global loss was heavily influenced by this hyperparameter, causing the model to avoid modifying the latent space, and instead focus too heavily on maintaining the learned latent parameters.

In Figure 3, the section on the disentanglement of latent variables illustrates how the β-VAE encodes different generative factors of variation within the latent space. Starting with a seed image, the latent space representation of the image is obtained. In the figure, each row corresponds to a specific latent parameter z_i, while the columns depict the effect of linearly interpolating the value of that latent parameter between -3 and 3, keeping all other latent parameters fixed. This setup allows us to observe how variations in individual latent parameters influence the reconstructed image, thereby highlighting the degree to which the model achieves disentanglement [3, 5].

Throughout the training, the first latent variables to be learned were z_0 and z_4, which encode the generative factors related to the X and Y positions, respectively, along with z_2, which captures the scale factor. These latent parameters remained relatively unchanged for the first 700,000 iterations. Subsequently, the model began to learn the generative factors associated with shape and rotation, represented by the latent variables z_7 and z_8. However, as training progressed, the model started modifying the previously learned latent parameters (z_0, z_4, and z_2), leading to entanglement with the new latent dimensions z_7 and z_8. Despite this entanglement, the reconstruction loss decreased from 81 to 30 (MSE), demonstrating that the model prioritized reconstruction quality while maintaining a stable KL divergence throughout the training process.

3.2.1 Entailments of the Posterior Mean and Posterior Variance

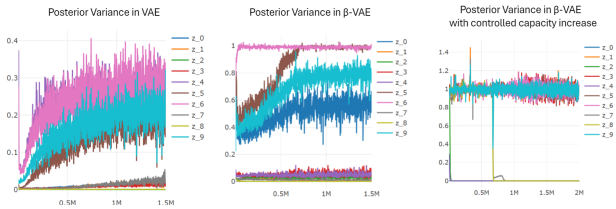

Figure 4. Posterior variance of latent variables for VAE, β-VAE, and β-VAE with controlled capacity increase.

Figure 4 illustrates the posterior variance of latent variables during the training of VAE, β-VAE, and β-VAE with controlled capacity increase. A key observation is that unused latent parameters tend to have a posterior mean of 0 and a posterior variance of 1. This behavior ensures that these parameters minimally contribute to the KL divergence term, effectively "turning off" their influence. The KL term, which is weighted by the hyperparameters γ or β, acts as a regularizer to encourage independence in the latent space, pushing the model to use only a subset of latent parameters that are most critical for reconstruction. This is consistent with findings in [5] and [3], which demonstrate how the KL term enforces sparsity in the latent space by penalizing deviations from the prior distribution.

For the latent variables that are used, their posterior mean is tightly constrained around 0, while the posterior variance approaches 0 as well. This behavior indicates that these latent variables are effectively deterministic. As highlighted by [3], a small posterior variance reduces variability in the reconstruction process, allowing the decoder to rely on precise and stable latent encodings rather than dealing with stochastic noise introduced by larger variances. This deterministic encoding improves reconstruction quality while maintaining disentangled representations.

Unused latent variables, on the other hand, retain a posterior variance close to 1, aligning with the Gaussian prior. This prevents these variables from contributing significantly to the KL divergence term and ensures they do not influence the reconstruction. The distinction between used and

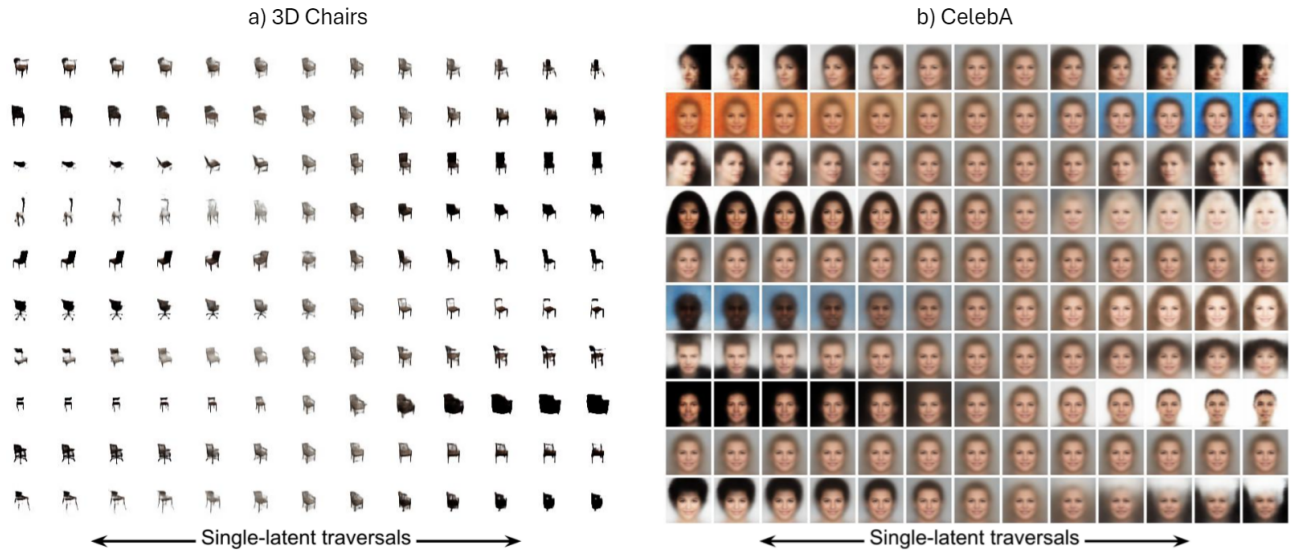

Figure 5. Training results of the β-VAE model with the architecture from [5] with $\beta = 4$. The latent traversals were done similar to Figure 8, where we initialize the latent space with a seed image, then traverse a single latent dimension (in $[-3, 3]$), whilst holding the remaining latent dimensions fixed, and plot the resulting reconstructions.

unused latent variables highlights the ability of the β-VAE framework to prioritize certain generative factors while disregarding others, guided by the capacity of the latent space.

One important consequence of the posterior variance of used latent variables approaching 0 is that it becomes impractical to sample new images directly from the latent space using the prior distribution. Instead, the standard procedure for generating samples in VAEs [7] and β-VAEs [5] involves sampling latent vectors around a known encoding of a dataset example. This ensures that the sampled latent vector falls within a region of the latent space that the decoder can effectively reconstruct, thereby mitigating the challenges posed by the deterministic nature of the used latent variables.

The posterior mean and variance patterns observed during the training of β-VAE models reflect a trade-off between disentanglement and reconstruction. The posterior variance of unused variables aligns with the prior to minimize their impact, while the used variables become deterministic to optimize reconstruction quality. These behaviors are direct outcomes of the regularizing effect of the KL divergence term and its influence on the learned latent space structure.

3.3. Results on Other Datasets

In addition to the results on the dSprites dataset, we replicated the experiments of the β-VAE paper [5] on the CelebA [11] and 3DChairs [1] datasets. The CelebFaces Attributes Dataset (CelebA) is a large-scale facial attributes dataset containing over 200,000 celebrity images, each annotated with 40 attributes. The dataset includes images with significant variations in pose and background clutter. The 3DChairs dataset consists of rendered images of approximately 1,000 three-dimensional chair models. While the true generative factors of these datasets are not explicitly known, the β-VAE model successfully captures several interpretable generative factors.

As shown in Figure 5b, for the CelebA dataset, the β-VAE model with a latent dimension of 10 identifies the following generative factors:

Table 2. Latent dimensions and generative factors for the CelebA dataset.

Latent Dimension	**Generative Factor**
z_1	Rotation
z_2	Background color
z_3	Rotation of the face
z_4	Hair color and background
z_5	Unused
z_6	Gender and background
z_7	Age and sex
z_8	Background and sex
z_9	Unused
z_{10}	Age

Although these latent variables correspond to meaningful generative factors, they are not entirely disentangled, consistent with observations in [5].

For the 3DChairs dataset, as shown in Figure 5a, the β-VAE model identifies several generative factors:

The remaining latent variables for the 3DChairs dataset

Table 3. Latent dimensions and generative factors for the 3DChairs dataset.

Latent Dimension	Generative Factor
z_2	Back height
z_4	Rotation
z_5	Leg style
z_8	Width

are entangled or non-interpretable, similar to the findings in [5]. These results demonstrate that while the β-VAE model effectively identifies some generative factors, achieving complete disentanglement remains a challenge, especially for datasets with complex or unknown generative factors.

3.4. Disentanglement Metric

The authors in [5] propose the disentanglement metric to quantify the independence and interpretability of latent representations by evaluating their sensitivity to fixed generative factors. The method involves generating images where one generative factor is fixed while others vary. Latent representations inferred from these images are analyzed for variance, with lower variance indicating better disentanglement.

Specifically, two sets of latent vectors are sampled, ensuring the value of the fixed generative factor remains constant across both sets. Images are then simulated and passed through the encoder to produce latent representations. The absolute difference between the two inferred latent vectors is computed for each sample, and the average difference is used as input to a low-capacity linear classifier. This classifier predicts the index of the fixed generative factor, and its accuracy across batches serves as the disentanglement score. The use of a linear classifier ensures that the score reflects the disentanglement properties of the latent space, rather than the classifier's capacity for complex transformations. This approach measures how well each latent dimension aligns with a single generative factor, providing an interpretable disentanglement metric.

3.5. Limitations of the Metric

The disentanglement metric proposed in [5] has several limitations. One primary limitation is the assumption that the generative factors of the dataset are known a priori. This assumption inherently makes the generative factors interpretable by default. Furthermore, applying this metric requires access to a dataset generator capable of producing images with one generative factor fixed while others vary. As such, the metric cannot be used with datasets lacking a generator, which is why we did not employ this metric in this paper.

Another limitation, which is less obvious, is that the metric does not directly measure independence between latent features. Independence is primarily encouraged through the KL divergence term in the β-VAE objective. To illustrate this, consider a scenario where the latent representation redundantly encodes a generative factor, such as having both z_i and z_{i+1} represent the same factor (e.g., scale). In this case, the differences $z_{\text{diff},i}$ and $z_{\text{diff},i+1}$ would have identical values, and a linear classifier could easily exploit this redundancy by performing an OR operation: if either value is small, the classifier predicts a high score. This demonstrates that the metric does not strictly enforce independence but relies on the KL term and the limited capacity of the linear classifier to encourage it. As noted by the authors, the classifier must have low capacity to ensure it does not learn complex nonlinear relationships [5].

4. Conclusion

In this paper, we reproduced and extended the results presented in [3, 5], highlighting the advantages of introducing the β term to enhance the disentanglement of latent factors. Our experiments demonstrated that β-VAE consistently and robustly discovers a greater number of meaningful factors of variation in the data compared to other baseline models, such as the standard VAE. This improved disentanglement results in latent representations that cover a broader range of factor values and exhibit cleaner separation, which is crucial for interpretable and generalizable generative modeling. Additionally, we verified that the objective proposed in [3] achieves comparable levels of disentanglement to those in [5], while requiring models with significantly reduced capacity. This finding underscores the efficiency of the modified objective and its potential for broader adoption in applications constrained by computational resources. We also investigated the role of the prior distribution in shaping the capacity of β-VAE models and analyzed the implications of posterior variances learned during training. Our results revealed practical limitations in both the sampling and training procedures, particularly in achieving low reconstruction loss without sacrificing disentanglement quality. Furthermore, we critically examined the disentanglement metric proposed in [5], identifying key limitations. Specifically, the metric's reliance on pre-defined generative factors and the assumption of access to a dataset generator restrict its applicability to real-world datasets where such information is unavailable. We also highlighted the metric's inability to strictly enforce independence among latent features, noting that disentanglement is influenced more by the β-VAE objective than the metric itself.

References

[1] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014. 2, 7

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 1

[3] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018. 1, 2, 3, 4, 5, 6, 8

[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint*, arXiv:1606.03657, 2016. 1, 2

[5] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 4, 5, 6, 7, 8

[6] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 44–51. Springer, 2011. 1

[7] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2, 3, 7

[8] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. *Advances in neural information processing systems*, 28, 2015. 1, 2

[9] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *arXiv preprint*, arXiv:1604.00289, 2016. 1

[10] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 1

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 7

[12] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2009. 2

[13] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International conference on machine learning*, pages 1431–1439. PMLR, 2014. 1

[14] Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013. 1

[15] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. *Advances in neural information processing systems*, 27, 2014. 1

Appendix

Verifying the Role of β in Latent Space Disentanglement

To ensure that the hyperparameter β is not merely acting as a regularizer but actively promoting disentanglement in the latent space, we conducted additional experiments. These experiments were designed to test whether high β values encourage disentanglement. Specifically, we evaluated a scenario where the latent space dimensionality was set to 5, corresponding to the five known generative factors: x position, y position, scale, shape, and rotation.

Surprisingly, our results showed that the VAE model, despite limiting the latent space dimensionality to 5, did not achieve a disentangled latent representation. The resulting GIF, which illustrates the latent traversals, can be found on our GitHub repository at https://github.com/lesupermomo/B-VAE/blob/main/misc/dsprites_VAE_z5_traverse_random.gif. For instance, we observed that the first latent factor encoded both scale and shape, while the second latent factor encoded a mixture of position, shape, and rotation.

In contrast, when using the β-VAE model with a controlled capacity increase, the latent space demonstrated significantly improved disentanglement. This suggests that the β hyperparameter, particularly with appropriately tuned values, plays a critical role in encouraging the separation of generative factors in the latent space.

A. Additional Figures

Here, we provide additional visualizations to complement the main results.

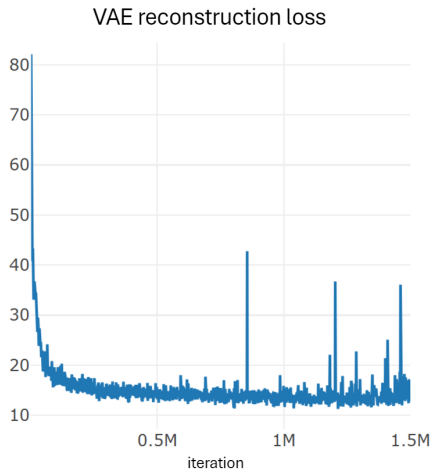

Figure 6. Reconstruction loss progression during VAE training.

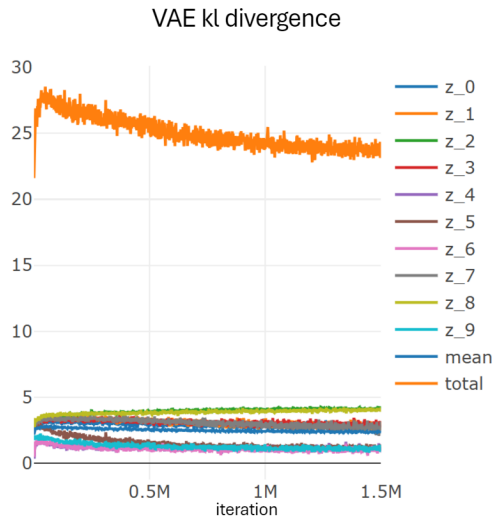

Figure 7. KL divergence progression during VAE training.

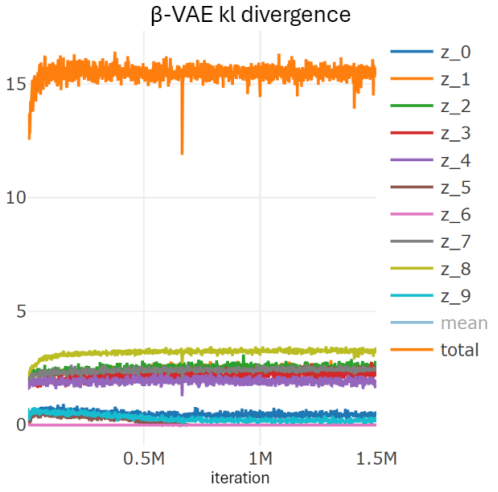

Figure 8. KL divergence progression during β-VAE training.

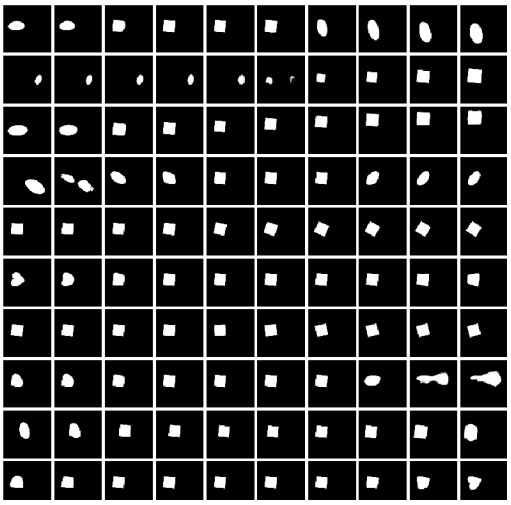

Figure 9. Latent parameters learned by the VAE model. This figure illustrates the variations in the latent space by modifying each of the 10 latent dimensions ($z = 10$) independently. For a randomly selected input image, the encoded latent representation was obtained, and each latent variable was varied from -3 to 3 (columns 1 to 10). The rows represent the visual changes induced by varying each latent dimension, showcasing how the VAE encodes entangled features in its latent space.